UNIVERSITY OF
LEADING
THE WAY
WESTMINSTER⌗

WestminsterResearch

http://www.westminster.ac.uk/research/westminsterresearch

An investigation into BBC Radio 4 comedy appreciation ratings (AIs): 'the cesspool with the velvet lid'

Amanda Benson

Faculty of Media, Arts and Design

An Investigation into BBC Radio 4 Comedy Appreciation Ratings (AIs): 'The Cesspool with the Velvet Lid'.

Amanda Benson

Abstract

BBC Radio 4 comedy is heard by over five million people every week in the UK. Selecting which comedy programmes are to be broadcast on the station is mainly the duty of a single person: Radio 4's Commissioning Editor for Comedy. She faces a delicate task, since her personal choices will affect the listening experience of millions of people. A poor selection will potentially result in a waste of licence payers' money, can have an effect on the perception of the station as a whole, and may even enrage the listeners. The BBC currently generates 'objective' numerical evaluations of audience responses in order to aid the commissioning decision-making process, but are the resultant figures useful? This thesis attempts to answer that question by investigating the suitability of this data in determining the quality of Radio 4 comedy programmes, using analysis of over 650,000 responses.

The key measure of objective audience evaluation of BBC broadcast programmes is the Appreciation Index (AI), a weighted mean value derived from programme appreciation ratings on a 10-point scale from a panel of respondents. This metric is also published as an aggregate station-level score on a quarterly basis as a station performance indicator, used as a 'meta score for channel quality'.

For BBC radio, there is no other recognised objective measure of programme performance that allows episode-level evaluation; the industry standard, RAJAR, does not allow sufficient granularity for audience size information at this level of detail. Thus BBC radio (in comparison to BBC television) has a particular need for the AI to provide a meaningful figure to facilitate programme evaluation.

Strict statistical theory dictates, however, that calculating a mean score using data taken from subjective ratings may not give useful results, particularly if the data spread is not unimodally distributed nor taken from a non-interval level scale. Given evidence that comedy is a divisive genre (and indeed as a constituent genre of Radio 4's broadcasting there is desirability for it to be so), comedy may be particularly poorly represented by AI scores. Indeed, the analysis of the data shows that that the responses are *not* distributed in a fashion that allows the mean to be a useful measure of central tendency. Not only are the aggregate responses not spread in a unimodal distribution, individual respondents have been found to adopt patterns (types) of responses that differ from the topline distribution.

Academic investigations relating to programme appreciation ratings have been relatively scarce, partly due to the limited number of broadcasters that measure this aspect of audience research, as well as the tendency for those that do to refrain from disseminating the data. Where studies have actually been published, researchers have not addressed the issue of the shape of the data's distribution nor the nature of its scale. This work's original contribution to knowledge considers these aspects specifically and does so for radio comedy rather than for the more typically utilised television

Contents

Dedication

To Florence and Louis

Declaration

The work included in this thesis is the author's own. It has not been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning

Notes and Terminology

Style Notes

- Where a programme title is in bold and italic, this specifically denotes a BBC radio comedy show, for example, ***The News Quiz***. In the appendix, there is a list giving the original TX dates for all BBC radio programmes mentioned throughout the document to allow easy reference.[1]

- Some frequently referenced radio comedy shows have been shortened to their colloquial title after their first mention in a chapter. These are:

  - ***I'm Sorry I Haven't A Clue = Clue.***
  - ***Just a Minute = JAM.***
  - ***Andrew Lawrence: How Did We End Up Like This? = Andrew Lawrence.***
  - ***Count Arthur Strong's Radio Show = Count Arthur Strong.***
  - ***The Hitchhiker's Guide To The Galaxy = Hitchhiker's.***
  - ***It's That Man Again = ITMA.***
  - ***I'm Sorry, I'll Read That Again = ISIRTA.***

- Where Radio Comedy is referred to with capitalisation, this denotes the BBC radio production department rather than the genre.

- Where quotations are in italics this means that the quote is a verbatim transcription from an interview conducted as part of this research. As they are verbatim, they may include non-standard grammar.

- Where blue page numbers can be seen, these contain hyperlinks to the page to which they refer. I.e., if you are reading this document in Microsoft word (or similar word processing package), you can click on the page number and it will take you directly to the page to which is being referred. Some hyperlinks will have a connection to specific pages, whereas, if the referral reads, 'see *from* px' this means that you are merely being pointed to the *beginning* of a section that may be relevant.

- In referring to data from the AI survey, where verbatim responses are quoted this is done so in the format below, prefixed by the programme title, the date of TX, the gender and age of the respondent and the respondent's unique identification number, along with appreciation score (when relevant), for example:

  > ***The Secret World***   25-Sep-12   Male 60 (respondent 719314)
  > Appreciation score – 1
  > 'Just rubbish – turned to another radio station'.

- Figures inserted into the body of the thesis which are taken from other sources are shown with a shadow behind them, to visually differentiate them from figures that present data originating from this research.

- Text in lilac boxes is used in the method and results chapters to summarise findings from each section.

Terminology Used

Some broadcast or research specific terminology is used throughout the thesis and is defined within. However, for ease of reference, some of these terms are summarised below:

- AI – appreciation index (for full definition see from p84).

- TX – transmission or broadcast of a programme.

- RX – recording of a programme.

- SoPP – statement of programme policy, i.e., part of BBC broadcasting commitment.

- WoCC – window of creative competition, i.e., BBC airtime that is open to proposals from both BBC departments and independent suppliers.

- RAJAR –Radio Joint Audience Research: a syndicated research body in the UK which measures radio listening.

---

[1] See from p291.

- BARB –Broadcasters' Audience Research Board: a syndicated research body in the UK which measures television viewing.

- GfK – the company that collects appreciation data for the BBC. Originally known as Gesellschaft für Konsumforschung.

- PDG – programme development group – Radio Comedy's regular meeting to discuss show ideas.

- PSB – public service broadcasting.

- IP radio – internet protocol radio is linear audio broadcasting over the internet.

- DTV radio – digital television radio, i.e., radio heard through a television.

- DAB – digital audio broadcast.

- R&M or A&M – the BBC's Radio and Music division, previously known as Audio and Music.

- A, B, C1, C2, D, E – NRS social grades used to denote UK class groups based mainly on the occupation of a household's chief income earner.

- Pulse – the systematic survey that asks a panel of respondents about their opinions of BBC programmes.

- Verbatims – the BBC's Pulse survey includes open ended questions where respondents can write whatever they wish about the programmes they have heard. These are referred to as verbatims or verbatim responses.

- ' and " – programme durations are written with ' denoting minutes and " denoting seconds. For example, 28' 30" would indicate a twenty-eight and a half minute programme.

- Gateway – the BBC's staff intranet portal, which includes access to audience research data.

- Talent – a word often used in broadcasting to denote creative people.

- Programmer - used as an umbrella term for those involved in commissioning and scheduling of radio programmes.

- Dippable - programming that a listener can dip in and out of, without needed to give full attention.

- Ratings – in the industry this word is used in two ways. Firstly, 'ratings' is used to denote audience size, i.e., the number of listeners to a radio station. Secondly, in audience appreciation, a respondent might give a subjective appreciation 'rating' to a programme. Within this thesis, meaning of this word will differ depending on the context.

- Daypart – a term denoting that a day is divided into sections, for example, drive-time or late night.

- Indies – Independent radio production companies.

Chapter 01 – Introduction

Often referred to as the "poor cousin" of television, it might be better likened to its drunken uncle.
It gets away with more: it's frequently sick; it does some outrageous things after 11pm
(and it's particularly amusing over the Christmas holidays).

(Hawes, 2004, p74)

This chapter provides an overview of the thesis; it explains the fundamental issue identified together with the challenge it presents, before outlining the basic argument of this study. The background to the subject area is given, highlighting the importance and originality of the topic. Next is a summary of existing work in this field, which leads in to the specific aims of this thesis in terms of hypothesis and research questions. This is followed by a summary of the limitations. Finally, an outline of the full thesis structure is presented.

1.1 Overview

Hypothesis: Radio 4 comedy programme quality is poorly represented as an unqualified mean of listeners' appreciation scores.

1.1.1 The Fundamental Issue

For BBC broadcasts, a panel of respondents – via the Pulse survey[2] – is asked to evaluate any radio or television show that they might have heard or seen. A summary figure for each episode is calculated from ratings given on a 1-10 scale provided by this panel, presented as a factor of their mean[3] known as the AI (Appreciation Index). However, strict statistical theory[4] implies that presenting a mean score (a type of parametric analysis) as a measure of central tendency from data extracted from a scale that is not interval or ratio (i.e., non-parametric data) can lead to misrepresentation of that data. This calls into question the validity of AI data as a representation of audience estimation of programme quality. No proof is evident that the scale used is any more robust in nature than merely ordinal,[5] meaning that the AI may not be a useful illustration of the feelings of the audience.[6] Radio 4 programming employs AIs as part of the recommissioning process[7] and should these figures prove to be misleading, it could be that decision-making is being influenced by poor information and that Radio 4 audiences are being presented with programmes that are not best suited for them. While the BBC measures appreciation across all of its channels, there has been little published research in the area particularly pertaining either to radio or comedy.[8] Comedy would be of particular interest due both to its divisiveness,[9] its impact on the audiences[10] and its commercial importance.[11]

The BBC has been using AI scores as programme evaluation aids for many decades,[12] and they are broadly unchallenged as a measurement tool, meaning that a challenge of the incumbent measurement procedure is therefore difficult. Furthermore, because the BBC considers the data sensitive, little research has been published in this area. In contrast to television, radio does not benefit from overnight audience size

---

[2] The Pulse is the BBC's systematic audience research survey. See from p85.

[3] The mean multiplied by 10.

[4] See from p141.

[5] See from p137 for an explanation of the various types of numerical scale.

[6] See from p160.

[7] See from p88.

[8] See from p8.

[9] See from p48.

[10] See from p35.

[11] See from p44

[12] See from p70.

ratings,[13] and therefore has particular need for audience appreciation figures. If these figures are incorrect, there is currently no other recognised objective measure for comparison.

1.1.2 Simple Numbers, Complex Meanings

The title of this thesis - 'The Cesspool with the Velvet Lid'- is drawn from a remark by Dixon[14] (Took, 1981, p77) which relates back to a description of BBC radio comedy from the 1940s when the department was prone to corruption despite its glossy, acceptable façade.[15] A parallel can be drawn with the accepted presentation of AIs as a single, simple figure, despite the quagmire of data that they superficially represent.

> Averages take the whole mess of human experience, some up, some down, some here, some there, some almost off the graph, and grind the data into a single number. They flatten the hills and raise hollows to tell you the height of the land – as if it were flat. But it is not flat. Forget the variety behind every average and you risk trouble, like the man who drowned in a river that rose, he heard, on average only to his knee. (Blastland and Dilnot, 2008, p68)

1.1.3 The Challenge

To date there has been no published research evaluating the distribution shape of BBC programme appreciation ratings, or the appropriateness of their presentation as a mean score. Such academic research as has been undertaken[16] has focused on television rather than radio appreciation, and whilst genre differences have been considered, detailed genre analysis has not been published. Furthermore, these studies treat the data as parametrically robust without any verification.

The original contribution to knowledge provided by this study is fundamentally based on the consideration of the distribution of BBC Radio 4 comedy appreciation ratings. The hypothesis is that the ratings may *not* be distributed in a fashion that allows parametric analysis, despite the BBC's continued practice to do so, and academic studies that use this type of examination. This work specifically examines whether appreciation responses are distributed in a unimodal, humpbacked fashion, and by inference assesses whether the measurement scale may be considered to be, at least, interval in nature.[17] If it is not, then aspects of Radio 4's decision-making may be in question. The data this research pertains to has never before been examined in this fashion.

The thesis specifically examines appreciation of speech radio programmes, an area not before covered in published studies dealing with audience appreciation ratings. Compared to television, radio as a medium is relatively overlooked despite its penetration: 'Its profile in the social landscape is small and its influence large' (Hendy, 2000, p3). The genre of radio comedy is examined specifically as it is seen to incite polarised responses,[18] and as such, may be particularly unsuitable for parametric analysis. No previous published studies on audience appreciation have solely focused on a single radio genre.

This research is significant as the data is taken from real-world audience responses rather than from an experimental design and, as such, can be applied directly to BBC Radio 4 programmes. Since the combination of shows covered by this study is typically heard by many millions of listeners every week, an improved understanding of programme performance is of value.

---

[13] See from p81.

[14] Pat Dixon was a BBC Radio producer who worked with The Goons and Tony Hancock.

[15] See from p20.

[16] See from p8

[17] See from p144.

[18] See from p48

Television evaluation has overnight, minute-by-minute viewing figures available to aid performance assessment, but this level of information is not available for radio due to limited time granularity and a diary sweep collection method.[19] Thus, appreciation ratings are of greater significance for audio broadcasters. AIs are used by programmers[20] in their decision-making and if these figures are misleading, there is a risk that they could lead to poor programming choices for the Radio 4 listeners. Findings from this research suggest alternative interpretive techniques that may be developed to help the BBC commissioners, its regulators and academic researchers use this valuable data to assess the performance and quality of BBC output.

1.1.4 This Research

This study ascertains whether BBC Radio 4 comedy programme quality is appropriately reflected in AI scores. It does so by analysing the distribution of the appreciation ratings, particularly in the light of comedy potentially being a polarising genre, and the analysis indeed indicates that the existing method of analysis of appreciation ratings is not strictly statistically valid. An additional output of this study indicates that BBC Pulse respondents have patterns of responses to scaled questions, which indicates that Radio 4 listeners are giving *comparative* responses of programme appreciation rather than absolutes. Furthermore, analysis illustrates that there are a number of independent variables that appear to have an effect upon the appreciation rating over and above the quality of the programme. This research also identifies a substantial source of error in the figures used by the BBC during 2012 and quantifies its effect upon figures published by the BBC. Finally, it is demonstrated that the survey used to evaluate programme appreciation incorporates some redundancy in its questions, resulting in duplication of results.

Certain phenomena identified in this work, such as the distribution shape of the ratings[21] and the use of response patterns by survey participants,[22] are likely to be applicable to all BBC programme audience evaluations (i.e., beyond just that of radio comedy), and might initiate useful changes to the BBC's audience appreciation evaluation process. Moreover, if the response patterns found in the data set are indicative of typical reactions to surveys with numerical scales, there could be implications for the evaluation of any ongoing study panel with scaled questions.

1.1.5 Summary of Key Findings

The primary finding of this research is that, on examination of the distribution of the appreciation scores, AI ratings are not statistically valid (despite continued use by the BBC), indicating that the hypothesis[23] is correct. In the process of the analysis, a number of related factors were observed:[24]

- 10% of the responses appeared to be illegitimate yet were still being included in figures used and published by the BBC.[25]

- The remaining aggregate legitimate scores were bimodally distributed. Thus, AI scores are a poor representation of programme performance.[26]

- Comedy is a divisive genre compared to others on Radio 4 (when considered at programme level).[27]

---

[19] See from p81.

[20] Programmer is used as an umbrella term for those involved in commissioning and scheduling of radio programmes.

[21] See from p176.

[22] See p88

[23] See p1.

[24] See from p176 – chapters 7 to 9.

[25] See from p167.

[26] See from p178.

- Respondents appeared to have patterns of response; 94% of responses analysed could be categorised into one of three types.[28]

- 10-point scale increments do not appear to have a direct linear relationship with appreciation.[29]

- High appreciation ratings correlated extremely highly with 'agreement' to a number of other questions on the survey, indicating redundancy across the survey's questions.[30]

- A number of independent variables, such as the number of series and the demographics of the audience, appear to have an effect upon appreciation ratings regardless of the actual 'quality' of a programme:[31] further indicating the unsuitability of AIs as an indicator of programme performance.

Aside from the indications of poor practice in collecting AI data, the findings of bimodally distributed scores and the identification of alternative patterns in the appreciation responses suggest that:

- The BBC could interpret and use AI data in a significantly more accurate and effective way than it does at present.

- In particular, the understanding of the performance of comedy programming could be significantly improved.

- Public, regulatory and academic understanding of the BBC's most prominent station quality performance indicator could be significantly improved.

- Statistical analysis of this type of data in other forms of research - for example market research – could be significantly improved, particularly in relation to response patterns.

- The number of responses received for a programme may be used to indicate the audience size within genres.

1.2 Background and Context of This Study

1.2.1 The Importance of Radio Comedy

'The rise of television in the post-war period has not brought about, as many people anticipated, the death of radio' (Shingler and Wieringa, 1998, Px). Indeed, Radio 4 is still holding fast, measured in Q4 2013 by RAJAR,[32] it attracted over 11 million listeners weekly and took a 12.5% (one eighth) share of all radio listening in the UK. Comedy is a key part of Radio 4's output commitment and a significant proportion of its listening, heard by 5.5 million Radio 4 listeners every week (ibid). One of the conditions of its service licence demands that, as part of Radio 4's Statement of Programme Policy[33] (BBC Executive, 2010, p35), there must be at least 180 hours of comedy originations[34] broadcast across a financial year. Part of the reason for comedy's presence in the schedule is to fulfil a cultural requirement; radio comedy is seen as having a particular significance to the listeners, particularly in its demonstration of the innovation and creativity of the BBC. The Trust's audience research for Radio 4's service review found that comedy is the 'key' genre in

---

[27] See from p225.

[28] See from p179.

[29] See from p216.

[30] See from p220.

[31] See from p251.

[32] RAJAR stands for Radio Joint Audience Research: a syndicated research body in the UK which measures radio listening.

[33] Known as SoPPs.

[34] Originations means a first broadcast, i.e., repeats are not included.

fulfilling the BBC's creativity requirement (BBC Trust, 2010, p4 – p44). Radio comedy is important in the following respects:

- **Entertainment** – '*I think we gain enormously by having comedy on BBC radio… Part of the BBC's mission is to entertain and comedy is curiously good at that'* (Davie, 2009).
- **Finance** –
  - o **Cost** – Radio 4 comedy costs the licence payer over £5 million per annum – around 10% of Radio 4's programming cost (excluding news programming).
  - o **Income** – 'Comedy is the area of Radio 4 output which is perceived to have the greatest potential commercial value.' (BBC Executive, 2010, p66).
- **Developing Ideas** – Comedy is often considered to be polarising (Williams, 2011, Raphael, 2012a) and therefore difficult to predict in terms of audience appreciation. Radio allows comedy ideas and formats to be tested in a relatively low-cost environment before a larger commitment is made in television. '*It's a unique resource that the BBC has for developing comedy*.' (Mitchell, 2010).
- **Developing talent** – In a survey undertaken for this thesis, 70% of comedians (over 100 surveyed) indicated that they still thought BBC radio to be an important aspect of their career development – higher than social networking at 60%. [35]
- **Attracting new audiences** – Comedy can be a key attraction in helping younger people find Radio 4, for example, *The Hitchhiker's Guide to the Galaxy* in the late 1970s (Hendy, 2007, p186, p193).

1.2.2 Radio 4 Comedy Evaluation

It is the Radio 4 Comedy Commissioning Editor's job to make choices about which new programmes to buy for the network. This might be done by viewing live performances, reading a synopsis or a script and, occasionally, by commissioning a pilot episode. These resources are evaluated subjectively with raw '*gut instinct*' (Raphael, 2012a). However, once a programme has been broadcast, audience research data is available to help objectively evaluate the reaction to a show. Unfortunately, for radio, quantitative data about the size of the audience is limited as the syndicated measurement system, RAJAR, is restricted to quarterly figures at 15-minute slot granularity:[36] unlike the television (BARB)[37] overnight figures which allows greater detail. Consequently, the best audience size measure available is an average across 13 weeks for 15-minute time slots, so it is therefore not possible to know how many people listened to a Radio 4 comedy episode in any specific week.

The BBC also collects data regarding audience appreciation (the AI) for each television and radio episode via its Pulse Survey (BBC Gateway Audience Portal, 2011), in order to supplement the audience size figures. These AIs, along with other Pulse survey responses, are the nearest thing to an objective measure of an individual episode's quality. Thus, AIs tend to be the primary component of non-subjective evaluation for the Radio 4 Comedy Commissioning Editor. 'We always look at them' (Raphael, 2011), '*I'll very often do it week by week to see trends* [as] *we have so little* [else] *to go on'* (Raphael, 2012a).[38]

1.2.3 Limitations of AIs: Distribution

Pulse respondents are asked to evaluate BBC programmes that they have seen or heard, based on a 10-point numerical scale that runs 1-10. For each show, the mean score taken from these responses is presented as a

---

[35] See from p285 for survey results.

[36] See from p81.

[37] BARB stands for the Broadcasters' Audience Research Board – the body that measures television viewing in the UK.

[38] See from p88 for expansion upon use of AIs by programmers.

representation of the overall programme appreciation and known as the AI. Should, for example, two people hear a Radio 4 comedy and one respondent give a score of 6 and the other a score of 8, then the mean would be 7. This mean score is then multiplied by ten to differentiate it from any individual score. In this example, the AI would be presented as 70. While in theory there is no mathematical inaccuracy in this method, in practice there are many limitations which include: the questionable parametric robustness of the ratings, the limited sample sizes, lack of supplementary data about the listening situation, limitations of being comparative rather than absolute scores and potential sources of error.[39] The focus of this research is, however, specifically the distribution of appreciation ratings.

The mean is an 'average' – a measure of the central tendency of a range of figures, which should be applied only to data that is unimodal and hump-backed.[40] If the distribution of that data is otherwise, it can be argued that the range of figures in such cases is 'poorly represented' by a mean (Ehrenberg, 1986, p3-9). A mean summarises the data but conceals the details: 'Whenever you see an average, think: "white rainbow", and imagine the vibrancy it conceals' (Blastland and Dilnot, 2008, p69).

Furthermore, parametric analysis, such as the mean, should only be used with data that is taken from an interval or ratio scale.[41] As the BBC is using the mean as a presentation of their AI scores, there is an implied proposition that it *is* parametrically robust in nature. But, the analysis of the data in chapter 7 suggests that the data *is not* distributed in a fashion that makes presentation as a mean a suitable summary figure. Should this be the case across a wide range of programme appreciation ratings, the implication could be that AIs, used by Commissioning Editors for decision-making processes, are a poor representation of programme appreciation. Furthermore, since AI's are the main quality measure reported by the BBC[42] this would have implications for the regulation of the BBC as well as for academic understanding of BBC performance.

1.2.4 AIs and Radio 4 Comedy Evaluation

The charts below show simple summaries, firstly of the Radio 4 comedy commissioning process and the role of AIs within it, then the process alongside the independent variables that may affect it:

---

[39] See from p115.

[40] See from p156.

[41] See from p144.

[42] See from p88.

Figure 1 – Role of AIs in Radio 4 Commissioning

Figure 1 – Role of AIs in Radio 4 Commissioning
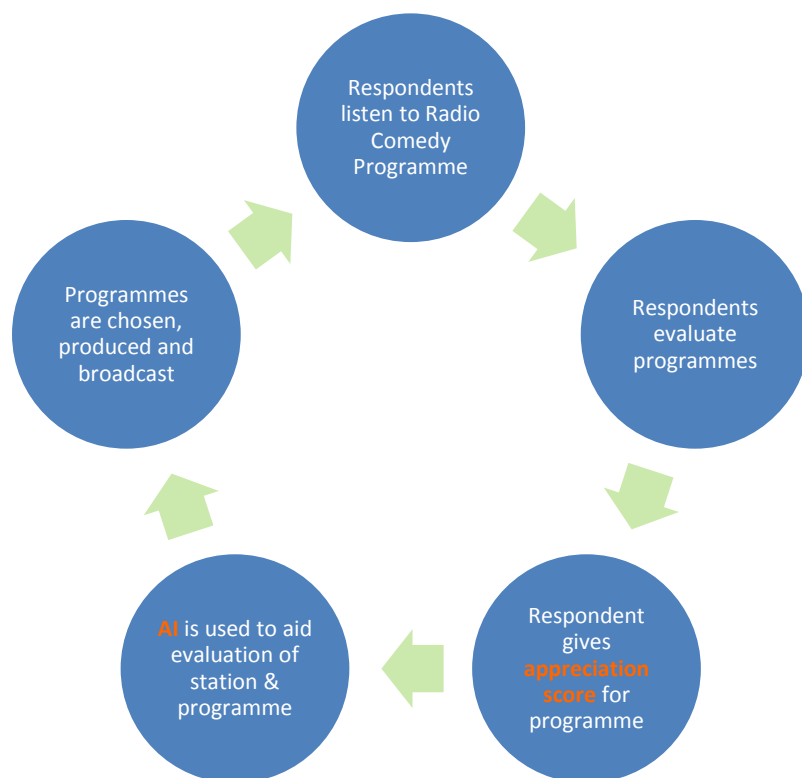
It is the stages between the basic process points that are of most concern to this study:

Figure 2 – Role of AIs in Radio 4 Commissioning – Variables in Addition to Programme Quality

Figure 2 – Role of AIs in Radio 4 Commissioning – Variables in Addition to Programme Quality
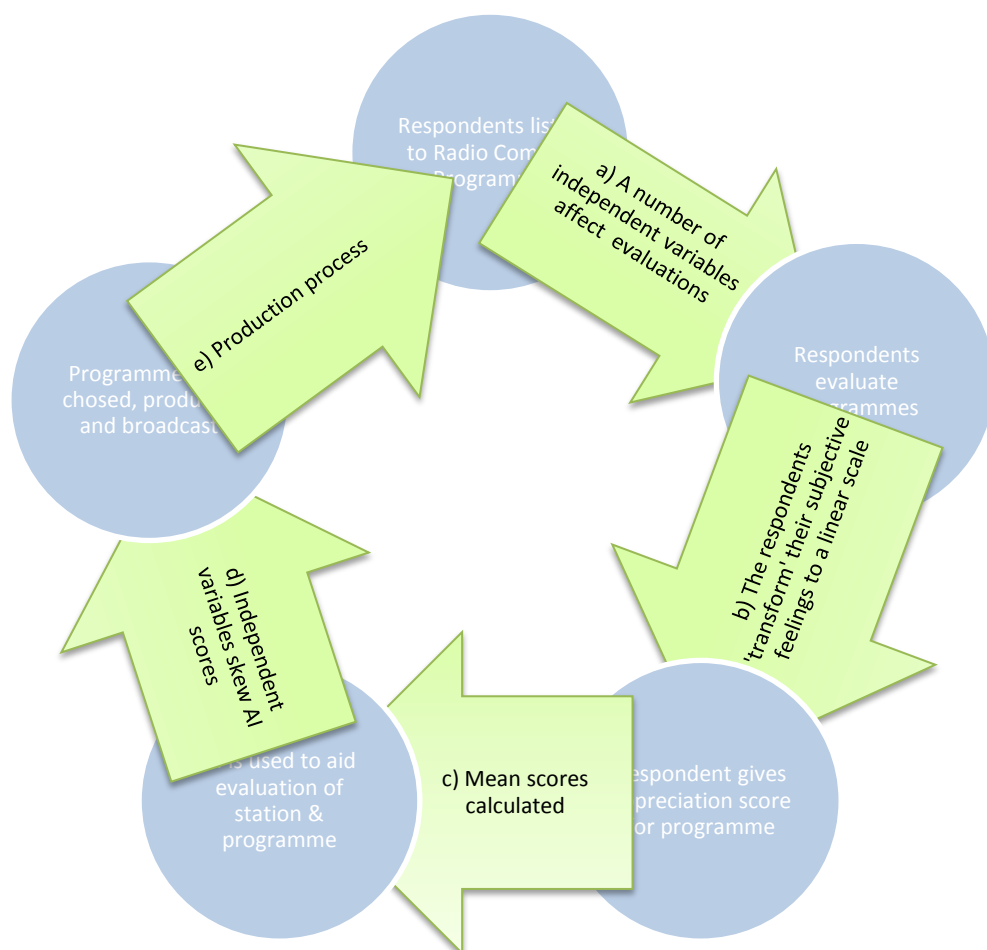
a) Among the many independent variables that can affect respondents' evaluations are the context of the experience, the nature of the programme content and the demographic make-up of the group of respondents. These elements are discussed relating to a variety of sources ranging from complex social science (for example, Kahneman, 2011) to interviews with comedians pertaining to the nature of comedy. Many of these aspects are discussed in chapter 4. A number of these points can be evaluated in point d).

b) Much research on subjective evaluation ratings implicitly treats the numerical ratings scales as at least interval in nature, despite there being a general view that attitudinal variables are no more than ordinal.[43] Discussion in this thesis questions the nature of the process that a respondent uses to transform feelings into a numerical figure. This area falls between statistical theory and the social sciences. Studies looking into online product ratings, such as those undertaken by Hu et al (2009) and Koh et al (2010) are helpful in this respect. Distribution of BBC Radio 4 comedy appreciation scores indicates that the transformation may not be linear,[44] and analysis of the verbatim responses enables this thesis to attempt to offer visibility of the true nature of the relationship.[45] Chapter 5 discusses the nature of response distribution.

c) The calculation of mean scores, while straightforward numerically speaking, is complicated when dealing with non-ordinal data. Text books on basic statistics, such as Ehrenberg, *A Primer in Data Reduction* (1986) and Bryman and Cramer, *Quantitative Data Analysis for Social Scientists* (1994), service this discussion. Research questions 1 and 2 involve this form of analysis.

d) The effect of the independent variables (section a) can be measured at this point and in some cases can be compared to findings from work relating directly to programme appreciation: in particular, Barwise et al, 1979; Barwise and Ehrenberg, 1987; Carrie's 1997 thesis, and BBC (unpublished) findings. Research question 3 deals with these aspects.

e) The vast number of variables between the commissioning of a programme and the point of its being heard by the respondent, i.e., point e) in the above figure, were investigated as part of the research process but are not covered in this thesis.

1.3 The Literature

This thesis examines how radio comedy audience appreciation is measured by the BBC and as such, considers areas that are comparatively neglected.

1.3.1 Radio

While radio is pervasive in society – 91% of UK adults listen at least once a week, and the average listener hears around 21 hours per week (RAJAR Q4 2013) – it has a relatively low profile in academia: 'Radio has not been given anything like the academic or critical attention devoted to film or television' (Shingler and Wieringa, 1998, Pxii). An ironically large number of researchers comment on how little research there is on radio, for example: Crook, 1999, p3 p7; O'Sullivan, 2000, pii; Barnard, 2000, p3; Douglas, 2004, p9; Beck, 2006, p128; Crisell, 2006, pviii; Gray, 2006, p250 and Chignell, 2009, p1. Starkey (2004, p25) argues that this lack of interest stems from radio's early days when contemporary researchers were mainly concerned with media effects, believing that passive audiences were easily manipulated (McQuail, 1997, p58). This meant that the motivation for, rather than the impact of, listening was seldom explored (Barnard, 2000, p85).

[43] See from p143.

[44] See from p177.

[45] See from p216.

While the Royal Television Society was formed in 1927, it was not until 1983 that the equivalent Radio Academy was founded in an attempt to raise the profile of radio in industry and academia (Starkey, 2004, p25). Lacey (2008, p21-22) posits that the introduction of the Radio Studies Network in the late 1990s has actually resulted in the distancing of radio from other media, 'rather than transforming it into a mainstream area of study', reinforcing it as the 'Cinderella of communication studies and a wallflower at the media studies ball'. Shingler and Wieringa (1998, Pix-x) claim that certain aspects of the very nature of radio contribute to researchers underestimating its relative importance: its ubiquity, the fact that it is free to access, and its invisibility. Goodale (2011, p4) argues that sound itself is a neglected area of research compared with the visual, calling it a 'blind spot' resulting in the neglect of 'ear culture'.

Common themes for discussion regarding radio research include the following areas:

- Perception – e.g. Shingler & Wieringa, 1998, Px; Barnard, 2000, p2; Hall et al, 2008, p31.
- Influence – e.g. McQuail, 1997, p99-105; Shingler & Wieringa, 1998, p137; Hendy, 2000, p3; Barnard, 2000, p221; Webster et al, 2006, p83.
- As a democratic medium – e.g. Shingler & Wieringa, 1998, p123-124; Hendy, 2000, p195; Douglas, 2004, p26; Hunter, 2010.
- Uses – e.g. Silvey, 1974, p149; Crisell, 1994 p211; MacFarland, 1997, p43; Dahlgren, 2001 p9353.
- As a secondary medium – e.g. Crook, 1999, p64; Starkey, 2002, p63; Douglas, 2004, p348; Chignell, 2009, p100.
- Engagement – e.g. intimacy (Chignell, 2009, p64), interaction (McLeish, 1999, p288-289), visualisation (Douglas, 2004, p29) and immediacy (Shingler & Wieringa, 1998, p103).
- Situation of listening – e.g. McQuail, 1997, Beck, 1997, p153; Scannell, 1991, p2; p101; Gray, 2006, p252.
- Ephemerality – e.g. McWhinnie, 1959, p12, p21, p42; Scannell, 1991, p1; Crisell, 1994, p9 p13 p58-59 p86; Crook, 1999, p7; Douglas, 2004, p30).
- The language of radio – e.g. McWhinnie, 1959, p78; Hendy, 2000, p152; Barnard, 2000, p176; Starkey, 2004, p176

1.3.2 Radio Comedy

Radio comedy specifically is very poorly covered, its literature 'disappointingly thin' (Chignell, 2009, p17). Those authors who do discuss it, such as Benson and Perry (2006), tend to examine its usefulness to advertisers or commercial broadcasters. Within textbooks about radio, when considering genres, comedy is occasionally highlighted and discussed as a specific area (for example, Crisell, 1994; Barnard, 2000 and Chignell, 2009) but sometimes just as an aspect of light entertainment (for example, Starkey, 2004), or limited to an American perspective (for example, Wertheim, 1979 and Douglas, 2004). As a genre, it is seldom considered with any depth. Some books claiming to cover all radio production include drama but fail to mention comedy in its own right (for example, Evans, 1977 and McLeish, 1999), while others ignore speech formats altogether, focusing only on music stations (for example, McFarland, 1997 and Keith, 2004). The few books that concentrate solely on radio comedy are popular histories of mainly early radio comedy and not academic in nature (for example, Took, 1981; Foster and Furst, 1996 and Coward, 2003). While there are some books specifically about radio drama production techniques (for example, Beck, 1997) and theory (for example, Lewis, 1981 and Crook, 1999), such texts do not exist for radio comedy.

1.3.3 Audience Programme Appreciation

Audience appreciation, as an aspect of media measurement, is seen as difficult to measure (Kent, 1994, p16) and difficult to interpret (Ang, 1991, p145), and is less routinely measured than ratings[46] (McQuail, 1997, p58). As ratings are the primary currency for broadcasters (Webster et al, 2006, p11), appreciation is generally seen merely as a supplementary measure (Gunter and Wober, 1992). It has been argued that, because appreciation offers little meaningful information about audience behaviour, it is possibly of lesser interest to commercial stations (Ang, 1991, p144-145). This is because the number of listeners to a radio station is a sellable commodity but the level of their liking of the programmes is not (at least, not currently). While some research has indicated that 'favourable programme attitudes may beneficially affect advertising reception' (Twyman, 1994, p102), commercial broadcasters tend not to measure this aspect. Data for programme appreciation is therefore rarely available to academic researchers, and published research is therefore scarce. The BBC is one of only a few broadcasters worldwide that has regularly gathered appreciation responses (Gunter and Wober, 1992, p60), but it seldom publishes this data in any detail.

> The nature and confidentiality of audience appreciation over the years has led to its being restricted to internal and largely non-sophisticated usage by the commissioning broadcasters. Few individuals have conducted extensive analyses of these data. (Carrie, 1997, p64)

Systematic research on radio programme appreciation ratings began in the UK with the BBC research department, and listener evaluations were collected systematically from 1941.[47] There was an occasional mention of research results in the annual BBC Handbooks that were available to the general public, but generally any studies done were limited to internal reports. Robert Silvey's 1974 account of his time setting-up and running the BBC's research department is a useful overview of the thought processes behind development of the types of research used by the broadcaster. From the mid-seventies to the mid-nineties, the BBC published an *Annual Review of BBC Audience Research Findings*,[48] and this resource provides insight into some of the changes made to appreciation measurement over the years covered. These books also include snippets of findings from ad hoc research projects relating to programme appreciation such as insight into how age was found to be an independent variable in appreciation ratings (BBC Audience Research Department, 1976, p16), or how the respondents reacted to a specific show such as *The Burkiss Way* (ibid, 1978, p32). However, it was not until 2011 that the Corporation began to regularly publish the aggregate AI scores at channel level (BBC Press Office, 2011).

Academic published work on BBC appreciation is very scarce because access to the data is restricted. Perhaps the only article is Menneer, 1984, his access to the data enabled by his being the BBC's Head of Broadcasting Research at the time. Yet even Menneer's article relates to television rather than radio appreciation. If detailed research has been undertaken at the BBC in recent years, it has been confined to an extremely restricted dissemination within the BBC itself, and very little internal BBC AI research has been made available to aid this research. GfK,[49] who collect the appreciation data for the BBC, do extensive analysis, but it is not published or disseminated widely.[50]

While it may be argued that appreciation measurement is of lesser interest to commercial broadcasters, it has been undertaken on occasion in the UK – albeit for television rather than radio.

---

[46] I.e., RAJAR audience size figures – see from p80.

[47] See from p66 for a brief overview of the development of systematic research at the BBC including examples of reports used.

[48] Known initially as the 'Annual Review of BBC Audience Research Findings', then as the 'Annual Review of BBC Broadcasting Research Findings' from 1980.

[49] GfK – originally known as Gesellschaft für Konsumforschung.

[50] While Gfk's global head of media and entertainment promised to share some analysis to aid this research in its early stages (North, 2010), such as appreciation by day-part and the comparison between scores of 9 and 10, this assurance was not fulfilled.

Researchers were able to utilise data collected by the Independent Broadcasting Authority (IBA) as part of its Audience Reaction Assessment (AURA), which had been in place since the early 1970s, followed by BARB data available from the 1980s. However, this data being commissioned by commercial broadcasters was arguably even more sensitive than BBC information; so perhaps predictably, published data is again limited. Much of the published work focuses on the relationship between audience appreciation and audience size – a relationship that has proved to be complex. Gunter and Wober (1992) summarise the data analysis undertaken in this respect: many of the studies referenced come from conferences rather than published work. Important studies in this area include work by Barwise et al who attempted to quantify appreciation's relationship with audience size (1979) and, later, using US ad hoc data, aimed to evaluate the relationship between appreciation and regularity of viewing (1987). The key text, however, is Carrie's 1997 PhD study which attempts to use existing television appreciation data collected by BARB (p57) to aid the understanding and interpretation of findings 'from data that is already being collected and used at great time and expense to the broadcasting and advertising industries' (p60). Carrie's research furthers the studies by Barwise and Ehrenberg in particular[51] and focuses on a number of areas including patterns of response, audience composition, audience size and repeat viewing. It is telling that despite its comprehensive coverage of the subject, since its completion it has only twice been cited in academic articles, neither of them written in English. Nothing with the scope of Carrie's 1997 research has since been published in academia in the specific area of the detailed analysis of programme appreciation ratings.

1.4 Aim of This Study

Within Radio 4 programming, decision-making based on instinct rather than research is an accepted aspect of the role of the programmers of the network.[52] Where television has overnight figures that give evidence of audience behaviour for each episode, radio's objective measure at this level of granularity is limited to Pulse data; appreciation ratings are the most widely used metric from this source. Appreciation ratings support the decision-making processes that determine which programmes are heard on Radio 4.[53] Comedy is generally considered a highly subjective genre,[54] making it particularly difficult to evaluate objectively.[55] It would thus benefit greatly from improved accuracy of information relating to audience response. The primary aim of this study is to investigate the suitability of BBC AI scores in determining the quality of Radio 4 comedy programmes.

A pilot analysis indicated that some Radio 4 comedies appeared to encourage appreciation responses that were distributed in a non-unimodal fashion.[56] This was an indicator that the data may not be based on an interval or ratio scale and is therefore, not strictly suitable for parametric analysis, such as the mean. This led to the hypothesis that Radio 4 comedy programme quality may be poorly represented as an unqualified mean of appreciation scores. Should the hypothesis be proven to be correct, it could indicate that recommissioning decisions for Radio 4 comedy programmes, largely guided by AI scores, might be based on misleading data. This could then imply that Radio 4 audiences may be hearing the comedy programmes that are not the 'best'.

> H1: Radio 4 comedy programme quality is poorly represented as an unqualified mean of appreciation scores.

---

[51] Ehrenberg was in fact the supervisor of the PhD.

[52] See from p60.

[53] See from p88.

[54] See from p48.

[55] See from p54.

[56] See p158.

The hypothesis drives three research questions. Firstly, are appreciation ratings well represented by a mean score? This study analyses the existing Pulse data to ascertain whether this is the case by looking into the distribution shape of the responses. Secondly, if comedy is indeed a particularly polarising genre,[57] will there be a greater propensity for extreme responses resulting in a reduced likelihood for them to be distributed in a unimodal fashion (exacerbating problems caused by use of inappropriate analysis measure)? Finally, while it has been found that appreciation ratings are not comparable across genres,[58] is it the case that comedy *sub-genres* also need qualification? If so, are there further factors that need to be taken into account? If radio comedy AIs are to be of use in Radio 4 programming processes, a deeper understanding of the factors at play is important and this study seeks to improve that comprehension. The value of this thesis is in aiding the interpretation of information that is already gathered by the BBC.

RQ1: Are appreciation ratings, as collected by the BBC Pulse survey for Radio 4 Comedy, usefully expressed as a mean score?

RQ2: Is the distribution of appreciation scores for Radio 4 comedy particularly polarised, and therefore an extreme case of the problems with using AI scores as a measure of performance?

RQ3: Are there underlying independent variables affecting appreciation ratings that can impact on comparisons of the distribution of appreciation scores across various Radio 4 comedy sub-genres?

1.5 Summary of Limitations of This Study

This research pertains to BBC Radio 4 comedy AIs. There is almost no published data in this area. Therefore, to aid basic understanding, a number of interviews have been conducted to fill in gaps in understanding of the processes involved in collecting and using appreciation ratings.

This analysis specifically investigates Radio 4 comedy in the UK for the calendar year 2012: the most up to date information at the time of the research.[59] The sheer volume of the BBC Pulse survey data dictates that the analysis is not extended beyond a single year. Obtaining the data has proven to be difficult as the information is not in the public domain and did not exist in the raw format required for analysis. If the scope of the data used was extended to further networks and/or for a longer time period, a request for information would have encountered more resistance. Due to these restrictions, the findings in this research are only applicable to Radio 4 comedy for 2012 and may not be generalisable across different media or genres without further validation.

The data used is taken from an ongoing online survey and subject to the limitations of any such data collection route. The Pulse survey is conducted with a sample respondent group which, of course, can only give an indication the population as a whole. When drilling-down in the analysis, sample sizes often precluded more detailed investigation. This becomes most apparent for programmes with smaller audiences such as those in the 23:02 TX slot:[60] the slot that may arguably be of greatest interest to BBC decision makers as it tends to consist of the newer, more experimental shows which demand the greatest need for understanding.

---

[57] See from p48.

[58] See from p114.

[59] See from p162 for the rationale, approach and methods used in this research.

[60] See from p35

The Pulse survey is limited to specific questions, and therefore the analysis can only be conducted across a limited range of variables. Some aspects that would have been of great relevance to radio comedy research, such as group listening or the listener's primary activity,[61] are not measured in the questionnaire. Furthermore, the key aspect of this research relates to question responses on a 1-10 scale so any findings are directly applicable only to this scale type.

In the process of the analysis, this research identifies a source of error within the data, that of illegitimate responses. The elimination of this error necessitated the discarding of a number of responses resulting in a smaller data set than was originally anticipated.

Various components of this study involve an element of subjectivity, which would affect the replicability of the analysis. Allocating sub-genres to each comedy show, identifying illegitimate responses and evaluating response patterns could all be constituents that may have varying interpretations.

Legitimacy of the base data was limited by certain assumptions firstly, that the data supplied was not subject to collection error and, secondly that the intentions of the respondents, for the most part, are without an ulterior agenda and that their responses did in fact attempt to represent their actual feelings toward the programmes.[62]

1.6 Outline of the Dissertation

In chapter 2, this dissertation gives context with a brief history of radio comedy then introduces Radio 4 and explains how comedy, as a programme genre, sits within its output and why it is of importance. This includes a broad overview of the commissioning process and its limitations, and highlights how comedy is a genre that is of particular interest when considering audience response. This section gives context and framing for the reader.

Chapter 3 gives a brief history of the development of systematic research at the BBC then discusses how broadcasters measure radio listening in terms of size and the limitations thereof. This leads into the need for alternative measure of audience programme evaluation.

Appreciation, as the main focus of this thesis, is expanded upon in chapter 4. Firstly, the collection process of the BBC's Pulse survey is explained in detail, including how the AI is calculated as the mean appreciation rating taken from a number of responses given on a numerical scale. It is then shown how it is used as a factor for broadcasters at both an aggregate and a programme level. The discussion then highlights which factors, over and above the actual programme quality, may be influencing the respondents as they come to evaluate a Radio 4 comedy show. Finally, the limitations of AI figures are considered.

As the data used for the analysis is taken from a survey with scaled responses, chapter 5 goes on to present a summary of the different types of scale are typically used in data collection questionnaires. The discussion presents differing theories as to whether subjective evaluations given on a numerical scale may be suited to parametric analysis, as is currently done by the BBC. The hypothesis is introduced, positing that the mean appreciation scores may not be a good representation of the overall appreciation for a Radio 4 comedy programme. This leads to three research questions which seek to ascertain whether: the mean score is indeed a poor representation for appreciation responses, if comedy is particularly prone to poor representation, and whether there are further factors at play.

Chapter 6 explains the method used in analysing the AI data, explaining how illegitimate data was discovered in the data set and how it was stripped out of the figures to give a more accurate result.

---

[61] Radio is often seen as a secondary activity – see from p78.

[62] See from p264 for full description on limitations.

The following three chapters each present analysis of the appreciation data. Chapter 7 shows that Radio 4 comedy appreciation scores are not distributed in a manner that is best represented by a mean score. This section also highlights how groups of respondents were found to show specific response patterns in their ratings. Furthermore, it illustrates how verbatim responses might be used to evaluate true 'distances' between subjective, ordinal scale points. Another key finding is that, while theory indicates that response rate can relate to audience size, this was found not to be evident when comparing different genres, although response rate and appreciation did correlate when considered for a single time slot and sub-genre. This chapter also shows how respondent agreement to a number of additional questions in the Pulse survey, correlate very highly with appreciation.

Chapter 8 shows how comedy can be the most polarising genre in terms of appreciation, but only at programme level, rather than when aggregated.

Chapter 9 goes on to consider sub-genres within comedy and finds observable differences between each type. However, when analysed in terms of demographic and other variables which affect appreciation (identified in chapter 4), these variables appear more significant than programme sub-genre. Some of these findings, such as demographic skews, support earlier academic and BBC research findings.

The 10th chapter of the thesis discusses the findings of the research, their implications and significance. It highlights the study limitations and goes on to suggest areas of further research and to propose changes to existing practices.

The appendix contains information which, while not vital to this research, may be of interest to the reader. It includes:

- Information from a survey of comedians conducted throughout 2012, to ascertain BBC radio comedy's claimed relevance to their career development.

- A brief discussion of terminology applied to numerical scales of 0-10 and 1-10.

- A list of all the variables held on Pulse respondents.

- A humorous illustration of the distrust of the Radio 4 commissioning process held by radio comedy producers (who wish to remain anonymous).

- A list giving the original TX dates for all BBC radio programmes mentioned throughout the document. It is in alphabetical order.

Chapter 02 – Radio 4 Comedy

A lot of rubbish is talked about comedy.
You can't define it and anybody who tries to pontificate is either a liar or the head of a television department.

(Dennis Main Wilson: Nathan, 1971, p119)

This chapter first gives a brief history of radio comedy then goes on to explain its relationship to Radio 4, which is the UK's prime broadcaster of original radio comedy. The size of the audience and the cost to licence fee payers are highlighted. Next, the complications of defining the genre of radio comedy are discussed, as are its suitability and importance as a category of audio broadcasting. Finally, the process of Radio 4's commissioning is explained, particularly in light of the problem of the subjectivity of decision-making.

2.1 A Short History of British Radio Comedy

2.1.1 Introduction

BBC radio comedy has evolved over 90 years, the unique type and success of public service broadcasting in the UK allowing it to persist where most countries have found it too expensive to be commercially viable since television became the primary source of entertainment. Its evolution has been a product of many things: the changing culture, the Second World War, new production techniques, new talent, etc. What we hear on air today is very much a finely balanced consequence of many years of negotiations between talent and management about what is acceptable on air, bound by the limitations of the medium and the zeitgeist of the times. Over such a long time one can argue that change is imperceptible to the listeners as fashions change slowly: 'One of the most intriguing elements of radio comedy is that it grows and develops almost unnoticed and suddenly you find a new mood in comedy – a change of direction, new voices, different jokes' (Took, 1981, p169). On the other hand, to the listeners, the smallest changes can seem huge; the deaths of Kenneth Williams or Humphrey Lyttlelton may have felt like the sudden end of an era for fans of ***Just a Minute*** or ***I'm Sorry I Haven't A Clue*** respectively.

2.1.2 The Early Days

Variety entertainment can be traced, in part, back to the 18th century when the 1737 licensing laws restricted dialogue on stage unless accompanied by music in an attempt by the government to curb the satire boom (Galton and Simpson, 2011a, p105). However, it wasn't until the late 19th century that comic patter, in particular, grew in popularity, developed from back-room pub entertainment with performers such as Dan Leno (Hall, 2006, p3). In the early 20th century this newer style of banter, along with the growth of cinema, began to impact upon the dominance of variety entertainment, particularly with the more sophisticated audiences (Foster and Furst, 1996, p3). But variety was about to be given a boost because it was going to be delivered to its audiences in a whole new way.

From 1895, long before wireless broadcasting had become popular even with enthusiasts, a gadget called the Electrophone allowed people to hear theatre and music hall performances in their own homes. They paid a subscription and were delivered audio over their telephone line, hearing the output on headphones; one of its early adopters was Queen Victoria (Crook, 1999, p15-17). Once wireless broadcasting was introduced, even before the formation of the BBC, listeners at home continued to be offered this kind of

fare. For example, in 1921 performers were paid ten shillings to be in a concert party style of show called *The Co-optimists* (Briggs, 1961, p288).

In 1922 came the formation of the BBC[63] and its initial thrust was purely to broadcast existing popular forms of entertainment rather than to create anything new (Barnard, 2000, p110). Early programmes were likely to reproduce elements typical of what audiences would expect from a music hall, such as the music, comperes, comics and audience reactions (Starkey, 2004, p166), and indeed throughout the 1920s and 1930s, producers continued to view light entertainment purely as a 'means of relaying shows from that habitat or, at best, as a place to recreate it' (Crisell, 2002a, p39).

The BBC broadcast its first variety programme in January 1923 (Briggs, 1965, p76), and from the very beginning the BBC was subject to restrictions from the powerful theatre proprietors (Barnard, 2000, p110), causing 'a turbulent relationship with the live theatrical arts' (Street, 2002, p43-44). The new medium caused much consternation for the owners and agents as there was, at that time, no appreciation for the benefit that broadcasting might bring to their artists (Briggs, 1965, p78), resulting in many of them banning the talent from being on the radio as part of their contracts (Briggs, 1961, p286). From 1923 to June 1925 live broadcasts from music halls were only allowed by exception until a 'strictly limited' agreement was made between the BBC and the president of the Society of West End Theatre Managers (Briggs, 1965, p77). The agreement was not successful as the variety artists refused to sign up for the proposal. The BBC tried to convince the owners and agents that, as radio was just an aural medium, it did not pose a real threat to their business as much as cinema did, a sentiment that did nothing to stem their concerns (ibid, p78). 'Music-hall and theatre owners felt that if their stars were available, as it were free to the listeners, then box office receipts would drop'. They did not imagine at that time that 'it would enhance the size of theatre audiences, who would flock to see a star they had previously only heard' (Took, 1981, p4).

Such was the worry of the potential loss of their powers that theatre managers in 1927 threatened to set up their own broadcasting organisation in order to control the distribution of their artists and their material (Briggs, 1965, p78). This threat drove another attempt for the two factions to come to a compromise and resulted in an agreement between Gerald Cock (Director of Outside Broadcasts for the BBC) and George Black (the Director of the General Theatre Corporation) to broadcast fortnightly shows from The Palladium from October 1928 (Briggs, 1965, p82). Further agreements with those who controlled the theatres and music halls meant that by 1929 Roger Eckersley (BBC Chairman of the Programme Board) could rely on having at least one variety programme each week coming from a London venue (ibid, p83). Unfortunately, the artists remained unhappy with the arrangement as, while they received extra payment if their act was broadcast, there was a concern that the BBC was 'giving their talents away' (ibid, p82). This new paradigm of entertainment meant material that a performer might use for years had become disposable, heard now by a huge audience in one performance.

> Radio simply devoured their material. Two or three routines which in the halls could last a lifetime now vanished into the ether and could never be used again. (Crisell, 2002a, p39)

The BBC had this same problem as they could not broadcast the same material over and over again and most of the performers only had a limited repertoire. This meant that new acts were needed constantly, and in the late 1920s and early 1930s, between 1,500 and 2,000 artists each year were auditioned as only around 1% were considered good enough for broadcast (BBC Year Book, 1932, p209; Briggs, 1965, p84). Many of the big stars of the time, still unsure of radio's potential and isolated from their large audiences, often did not fare as well as lesser known performers who were used to smaller audiences and had less to lose (Briggs, 1961,

---

[63] Then the British Broadcasting Company, it became a Corporation in 1926.

p286). Furthermore, some performers were intimidated by the microphone and others felt restricted by the censorship of the BBC (Street, 2002, p43-44).

While the BBC was for the most part just recreating other media in audio, it did begin to experiment with its own content. The very earliest BBC comic character created specifically for radio was 'Our Lizzie', played by Helena Millais from 21st November 1922 (Briggs, 1961, p286). She talked about 'comedy fragments from life' and had her own catchphrase: 'Ullo, me old ducks! 'Ere I am again with me old string bag' (Foster and Furst, 1996, p4). However, the first radio comic to really capture the imagination of the listeners was John Henry who played a hen-pecked husband (Briggs, 1961, p286). From his first broadcasts in 1923, his wife's call of 'John Henry, come here!' turned him into a national star (Coward, 2003, p15). The first revue that the BBC put on itself, as opposed to simply broadcasting a music hall show, was *The 7:30 Revue* from Manchester in March 1925 (Briggs, 1961, p288), but the first real series, broadcast from London, was *Radio Radiance* (1925) featuring Tommy Handley (Street, 2006, p249).

Ongoing complications with theatre controllers persisting into the 1930s was another driver for the BBC to continue to seek out its own talent, but it was a difficult balance as artists who weren't affiliated to the theatres couldn't be guaranteed a living and still had to rely on income from live performance. Val Gielgud (Drama Producer) wrote to Roger Eckersley in 1931 stating: 'We are not in a position to offer any artiste a complete livelihood' (Briggs, 1961, p286). This situation was not mirrored in the US where successful performers were able to make good livings from radio broadcasting alone as advertisers were able to pay up to $25,000 for an hour's sponsored programme (ibid, p94).

But variety was not a dominant proportion of BBC broadcasts, not just because of the problems securing talent. John Reith[64] was not a fan of 'the rough vulgarities of common comedians' (Took, 1981, p5) so there was bias towards the rather more 'refined, middle-class, and eminently respectable' humour (Took, 1981, p5). The output exemplified by the style of 'sketches from humourists' was rather too gentle to be very funny and 'the novelty of the settings in which these were being placed being the only real artistic development' (Took, 1981, p9). However, whatever the exact nature of the output, there was enough of interest to the listeners to encourage uptake and the number of licences issued grew throughout the decade from 35 thousand in 1922 to over 3 million in 1930 (Took, 1981, p9).

In the 1920s there was no one person in charge of BBC light entertainment, the area merely being a side-line for the Productions Director (Briggs, 1965, p76), with the whole business surrounded by an 'inconsequential, concert-party atmosphere' (Gale Pedrick: BBC Year Book, 1948, p53). It was not until 1930 that the genre was considered to be significant enough for someone to control it exclusively, at which point John Watt from Belfast was brought in to lead the new Revue and Vaudeville department (Briggs, 1965, p89). However, it was in 1933 that humour really began to be a substantial focus for BBC output. The creation of the 'Variety' Department, headed by Eric Maschwitz, began to be systematic in its development of material made specifically for radio (BBC Year Book, 1934, p115, Crisell, 2002a, p40). This was in recognition that there was a demand from the listeners for pure entertainment, 'something less exacting' than the talks and news that educated and informed (BBC Year Book, 1934, p116). The BBC used the word 'variety', in this context, to encompass a wide range of light musical entertainment, spanning vaudeville to operetta (ibid p115). The output of this new department was allotted 17.5 hours per week, including 2 hours each of both revue and vaudeville (ibid).

---

[64] Reith was the BBC's first Director General.

Many of the styles that fell under the Variety banner were not well regarded at the time; they were thought of as '*parochial and homespun*' and described disparagingly by Gielgud as a bunch of 'ukulele players and comedians' (Took, 1981, p11). Part of the issue was that traditional variety relied on visual humour and 'ribald elements', which would not translate well to audio or be allowed on BBC radio respectively (Briggs, 1965, p94), thus resulting in 'formidable shortcomings' (BBC Year Book, 1934) and leaving just 'safe and inoffensive' remnants (Chignell, 2009, p15). Furthermore, satire was a no-go area at this time (Briggs, 1965, p113). Maschwitz felt that freshening up the talent base would lead to new ideas and he did all he could to find new performers. He found that discovering talent ready-made for radio was not easy and thus proposed the creation of an 'act building department' full of 'Variety Apprentices' that could be trained to be radio performers (Briggs, 1965, p106). Unfortunately his idea was not taken up but it did not stop him from investigating other avenues. 1933 saw Maschwitz demonstrating the BBC's independence from the control of the theatres when he acquired one on its behalf – St George's Hall in Langham Place – from where it broadcast regularly, until 1943 when it was damaged in by bombing. Even though they had merely crossed the road, working from this building rather than the stuffier atmosphere of Broadcasting House, gave the Variety department a feeling of freedom (Took, 1981, p17). The relocation to Broadcasting House from its original base in Savoy Hill in 1934 had not been the most popular move. Savoy Hill was better placed for easy reach of the pubs and restaurants of Covent Garden and the theatres and agents based in Soho, while Portland Place was seen merely as, 'a select neighbourhood, away from the show business and the smell of trade… a part of London that most of us associated with women's shopping and the obscurer Embassies' (Gorham, 1948, p40).

Despite Maschwitz's developments, it was only really in 1938 that the BBC truly created a '*radiogenic*' comedy (Crisell, 2002a, p62). ***Band Waggon*** was the first programme to feature a resident comedian (Foster and Furst, 1996, p13), and set a new standard as it portrayed comic action being built around a selection of characters in a specific situation (Barnard, 2000, p112). Ronnie Waldman, later the BBC Head of Light Entertainment, recalled how the programme came about partly from a piece of information from audience research, part toss of a coin and ultimately down to talent availability:

> Waldman recalls that "BBC audience research came up with the information that dance band shows weren't as popular as we thought they were. They were all right but no more. After a lot of thought, John Watt [The Director of Variety] said, "We can't cut out dance bands altogether, but let's have a compere and a comedian and put something more into the show." And Gordon Crier, the producer, said, "There's a young chap called Richard Murdoch who's an extremely polished actor. He'd be a good compere." Well, Harry Pepper and I had just done our annual tour of the seaside summer shows and in our opinion there were only two comedians worth looking at for the job. They both happened to be in Shanklin, Isle of Wight. One was called Tommy Trinder, and the other Arthur Askey. They were equally good, so we tossed a coin, heads for Trinder and tails for Askey, and it came down heads". Thus ***Band Waggon***'s comedian should have been Tommy Trinder. But, he wasn't available, they asked Arthur Askey and fortunately he was. (Took, 1981, p17-p18)

So, BBC radio comedy took a step towards being a bit more exciting and anarchic with this new show. It may have been coincidence, but it just happened to be the same year that John Reith made an emotional departure from his role as Director-General of the BBC (Briggs, 1965, p635).

2.1.3 Wartime

Band Waggon's new style proved popular and in June 1939, over a few pints in the Langham Hotel (BBC Year Book, 1941, p73), there was a discussion about whether this type of success could be replicated. There was also a proposal to include the American-style quick-fire patter, which was popular at the time through

movie artists such as the Marx Brothers (Briggs, 1965, p118). Those in that discussion were writer Ted Kavanagh, radio producer Francis Worsley, and comedian Tommy Handley. As a consequence of that meeting, the programme that they put on air on the 12th of July 1939 was *ITMA – It's That Man Again* (Took, 1981, p22-23). Breaking away from the traditional music hall conventions (Coward, 2003, p19) it was the first BBC radio show to really use sound to the full, in an 'anarchic and irreverent' fashion (Foster and Furst, 1996, p28). Although, its pre-war (just) launch garnered little success (Briggs, 1970, p564), *ITMA* became arguably the most popular radio comedy show in UK history with a regular audience of 20 million domestic listeners and 30 million worldwide (Hendy, 2007, p26). It ran for 310 episodes from 1939 to 1949, only ending due to the death of Handley. It was so popular that his funeral became a national day of mourning (Coward, 2003, p17). *ITMA* filled a valuable role during the war, providing entertainment and solace: attempting to 'jazz the black-out blues' (BBC Year Book, 1941, p72).

> Entertainment at the BBC during the war was really committed to the national cause. So there was a consensus in terms of what it should be doing. What it should be doing was cheering people up, providing a useful sense of distraction at a time of insecurity and anxiety and also mocking the common external enemy. (Graham McCann: *Auntie's War on Smut*, 2008)

For a brief period at the beginning of the War, all theatres and cinemas were ordered to close. As compensation to the British people, the government encouraged the BBC to offer even more comedy and variety in its evening broadcasts (Foster and Furst, 1996, p53). During this wartime period, programmes such as *ITMA* and *Band Waggon* became vital because they were recorded in a studio rather than the theatres, which were no longer available for live relays (Crisell, 2002a, p62). The studio recordings attempted to replicate the conditions of theatres for the audience so they might feel, despite the War, things were carrying on as normal.

The expansion, aiming to divert and cheer up the public and troops (Parsons, 2010, p79), 'took on a new inclusiveness in terms of tone, manner and accent' (Barnard, 2000, p112). The expansion of light-hearted content was welcomed by the audiences. A listener research survey in 1942 found that of all the genres, variety was, by far, the content for which people expressed most enthusiasm, although this was clearly skewed towards the working classes (Briggs, 1970, p572). The increase in variety output, along with providing content for the Forces Programme, meant that the department were now called upon to provide well over 100 'spots' every week in 1940 (BBC Year Book, 1941, p72), compared to around 30 per week pre-war (BBC Year Book, 1945, p50), and reportedly rising as high as 180 in 1941 (BBC Year Book, 1942, p46). But there was an issue of supply; the war was a drain upon the talent available as men were called up to fight, taking many of the youngest, brightest performers, writers and producers (Took, 1981, p33). Some new blood was taken from the London advertising agencies (Barnard, 2000, p112) but this was not sufficient to cope with the increasing output, resulting in the majority of output between 1939-1942 being of 'poor quality' (Took, 1981, p33). Though regional broadcasts had ceased, there was still a modicum of content provided from outside London, such as *King Pins of Comedy* and *Black-outs for the Black-out* from Scotland which helped prop up the output (BBC Year Book, 1941, p73). Unfortunately ENSA – the Entertainments National Service Association based on Drury Lane, concerned with entertainment of the troops and munitions workers – were of little help. They were responsible for providing material for the *ENSA Half Hour* on the BBC, but arguments regarding ownership of content resulted in tension and, on occasion, their scripts were rejected as being 'amateurish' (Briggs, 1970, p312-313). The Variety department was not helped either by the strict 'heavy-handed' (Coward, 2003, p23) rules of the censors, who allowed very little spark to the content, imbuing it with 'apparent dullness' (Took, 1981, p37). Indeed, although there was a proven enthusiasm for variety, research in 1942 showed that there had been a significant slump in the

level of satisfaction from variety, with listeners' expressing dislike of the vulgarity of the programmes (Briggs, 1970, p570). Furthermore, the Board of Governors 'deplored "the low standard of music hall"' (Took, 1981, p39). John Watt, then Director of Variety, was himself critical of his own output. Writing in 1942, he recognised that there was not the available talent to fulfil the required production levels without the affecting the quality:

> We have found that we cannot be funny all the time, as we have tried to be, owing to lack of writers and material. Let us, therefore, try to be funny half the time and do more musical shows, though we invariably have only half the audience. (Briggs, 1970, p572)

Access to talent was not aided by the Variety Department's evacuation to Bristol in 1939 (Crisell, 2002a, p65). They had to move again in 1941 to Bangor, North Wales following a cryptic, coded news item about a camel falling ill at London Zoo (Foster and Furst, 1996, pii). Thus, a round trip of around four hundred miles had to be undertaken by performers, such as Arthur Askey, for them to be able to broadcast (Briggs, 1970, p571).

Luckily, Watt came up with an inspired idea to counter his problem with lack of talent to make BBC radio comedy. Rather than try to unsuccessfully replicate the popular American style, he sent his assistant to buy in American radio directly from the United Services Organization (USO), the US equivalent of ENSA. USO's role was to make radio to entertain the US troops (Took, 1981, p39), and the BBC bought their shows as a sort of syndication (Crisell, 2002a, p65). This gave instant access to extra content and allowed the British listeners to be entertained by the style of comedy that they had come to love via cinema. The new American shows included the fashionable 'slick compering and comedians being allowed to project their personalities through whole programmes rather than contributing "turns" of strictly limited duration' (Briggs, 1965, p109). Excluding *ITMA*, the period between 1942 and 1944 saw all the most popular radio shows on the BBC to be American, including *The Jack Benny Hour* and *The Bob Hope Show* (Barnard, 2000, p112; Took, 1981, p40). 'Thereafter there were to be two distinct strands in wartime Variety, one essentially British Provincial and one American' (Briggs, 1970, p315). Although the American shows were generally more highly thought of, there were also traditional, home-grown shows, in addition to *ITMA,* that remained relatively popular such as *The Pig And Whistle* (ibid). By the end of the War, however, there were complaints that there was too much Americanisation and around this time new UK talent began to surface such as Kenneth Horne and Charlie Chester (Took, 1981, p39).

2.1.4 Post-War

The wartime period had introduced a new sound to radio comedy that continued after The War ended. Programmes such as *ITMA* had allowed 'ordinary voices' to be heard (Barnard, 2000, p112) and, indeed, *Band Waggon* in particular had blurred the barriers of class distinctions as 'Arthur Askey and Richard Murdoch were from two ends of the social spectrum. The listeners loved them both and they loved each other' (Parsons, 2010, p79-80). This less elitist atmosphere persisted as a new wave of entertainers made their way to the BBC after they were demobbed (Took, 1981, p56), bringing with them the barracks humour that they had developed from the concert parties during their service (Barnard, 2000, p112). Many of these ex-servicemen, having been through combat, had an irreverence and complete disrespect for authority and this made a huge difference to the way radio comedy evolved after the war. 'Jovial anarchy' and '"The death of deference" it was called at the time' (Cryer, 2009). Among this group were Frank Muir, Denis Norden, Terry Thomas, Spike Milligan, Peter Sellers, Dick Emery, Cardew Robinson, Tony Hancock, Frankie Howerd, Kenneth Williams, Harry Secombe and Jimmy Edwards (Foster and Furst, 1996, p54).

The BBC committed to give any returning servicemen or women an audition if they wanted it, but BBC variety producers had a number of additional ways in which they could find new talent. Took (1981, p57) argues that there were primarily three more routes via which performers could be 'discovered'. Firstly, the Windmill Theatre was a great place to see unestablished talent. It was a hard training ground for comics where they had to do six shows a day, coming on between the dance numbers and nude women that the majority of the audience were chiefly there to see. Comedians such as Tony Hancock, Harry Secombe, Peter Sellers, Dick Emery and Jimmy Edwards all played the Windmill early in their careers, braving the stony audiences. At the Windmill 'You walk on and off to the sound of your own feet' (Cryer, 2009); Spike Milligan, Benny Hill and Roy Castle were among those who were not considered good enough to even grace its stage (Took, 1981, p57). The second route was through the Nuffield Centre, a services club for 'other ranks' just off the Strand. It was run by Mary Cook, and anyone she thought particularly good she would highlight to the BBC and make sure they were seen by producers. Michael Bentine, Alfred Marks and Frankie Howerd were all recipients of her good references (Took, 1981, p59). The third, rather left-field route to the BBC was merely to hang out at Jimmy Grafton's pub near Victoria in London. Grafton, in addition to running a pub, wrote for comedians including Alfred Marks, Dick Emery and Robert Moreton. Young comics were drawn there and camaraderie centred around the pub. 'After the war there was a very special atmosphere among young comics and actors,' Tony Hancock explained, 'we all seemed to know each other. Anyone who was working helped the others, even paid for their laundry' (Goodwin, 1999, p95). It was a place to meet likeminded people and it was there that, in 1948, Michael Bentine introduced Harry Secombe into the Grafton circle, shortly before Spike Milligan and Peter Sellers too joined in the fun (Took, 1981, p59-p60), all of them having worked separately on the variety circuit (p57). These four found that they had a similar sense of the absurd and it was at Grafton's that *The Goons* performed for the first time; the comedic connections allowed BBC producer, Pat Dixon, to become aware of their work (Foster and Furst, 1996, p146). Although ***The Goon Show*** idea had been systematically rejected by the BBC for 3 years (ibid), when finally broadcast it ultimately epitomised radiogenic comedy, going on 'to achieve extraordinary cult status, and influence a later generation of writers, performers, and producers' (Hendy, 2007, p26).

While newcomers were being discovered, at the other end of the scale, established talent also desperately wanted to get on air. Unlike the early days of radio when the agents and theatre owners feared the impact that radio might have upon artists' incomes (Briggs, 1965, p78), radio was now seen as 'the quickest way to national fame' (Took, 1981, p111). Radio could make huge stars out of comics and with the variety circuit still alive and money to be made (Foster and Furst, 1996, p125), there were powerful parties that had a great interest in ensuring that their people got on air. These parties exerted their influence in nefarious ways, so it was during this period that BBC Variety department obtained a rather sullied reputation. Pat Dixon, the producer responsible for getting *The Goons* onto BBC radio, was horrified by the practices of bribery that were occurring. Thus, when welcoming a new colleague to BBC Variety he described the department as 'the cesspool with the velvet lid' (Took, 1981, p77). Michael Standing, coming in as Director of Variety in 1945, was similarly concerned about the department's commissioning processes:

> The thing that worried me most when I went to Variety was the sense one felt of corruption impinging on the department from the outside. You were surrounded by people who regarded the giving of presents as an absolutely fair and orthodox way of behaving. Song plugging was rife, and the practice of the entertainment business giving expensive presents was widespread… Within a fairly short time of being in the job, I interviewed a band leader about a broadcast, and when he'd gone I had occasion to shift the blotter on my desk and found twenty pounds there. So he didn't do any broadcasting for quite a time after that… You see, in those days salaries in the BBC were shockingly low, and married producers with children had a very difficult time in making ends meet at all. They had to go out and meet artists who lived on a totally different scale, all the important figures in the business, and it was often very difficult indeed for them to resist what seemed to be a perfectly friendly offer, a loan for this, a holiday there – and occasionally things went wrong. (Standing: Took, 1981, p84-85)

Despite any issues the Variety Department may have had, in 1946 the 43 producers managed to fill 12,000 programme slots on BBC radio and issued over 8,400 talent contracts (BBC Year Book, 1947, p49).

During his tenure, Standing made a number of changes to the Variety department in an attempt to improve its working practices. He found that the spirits of the producers were depleted, not just due to pay but also because they were tired of being held responsible for a disproportionate amount of 'flops and inferior work generally' (Briggs, 1979, p710). Standing sought to improve the situation by trying to allocate programmes based primarily on the interests and tastes of the producers. He also introduced the Script Section within the department, recruiting dedicated comedy writers straight out of the services such as Eric Sykes, Bob Monkhouse, Denis Norden, Frank Muir and Ray Galton and Alan Simpson from the TB clinic (Barnard, 2000, p112).

Standing was particularly known for another aspect of BBC comedy broadcasting: the introduction of the infamous *The Green Book – BBC Variety Programmes Policy Guide for Writers and Producers*. With ever more barrack room humour creeping onto the air (Took, 1981, p85), and as the BBC returned to broadcasting live from theatres as they had before the war (Graham McCann: *Auntie's War on Smut*, 2008), there was a need to maintain standards of decency in a way that was clear for all variety producers. Max Miller had been banned from the airwaves for five years for telling the following joke live on air, and now with others fearful for their jobs too (Christie Davies, *Auntie's War on Smut*, 2008), it was the kind of thing that BBC executives wanted to avoid:

> I met this girl on this narrow path along the side of a mountain. I didn't know what to do. I didn't know whether to block her passage or toss myself off. (*Auntie's War on Smut*, 2008)

While the BBC could have accepted that public attitudes were accepting of such bawdiness (they could hear it in the music halls), it was still thought that the BBC should maintain standards; even if it were guilty of being a 'fossilised purveyor of national culture with moral and ethical standards which were wildly different from the way life was lived' (Brian Brivati: *Auntie's War on Smut*, 2008). As such, Standing drew up a series of guidelines for producers in order for there to be clarity in exactly what was and wasn't to be allowed on air.[65] Although there were some *very* specific instructions, such as avoiding jokes about 'Pre-natal influences (for example, "His mother was frightened by a donkey")', it was clear in its introduction that people needed to be practical and use common sense regarding what was suitable:

---

[65] The Green Book guidelines are reproduced in Took, 1981, p86.

This booklet is for the guidance of producers and writers of light entertainment programmes. It seeks to set out the BBC's general policy towards this type of material, to list the principal 'taboos', to indicate traps for the unwary or inexperienced, and to summarise the main guidance so far issued of more than a short-term application. It is however no more than a guide, inevitably incomplete and subject of course to supplementation. It cannot replace the need of each producer to exercise continued vigilance in matters of taste. (*BBC Variety Programmes Policy Guide for Writers and Producers:* Took, 1981, p86-91)

As television did not really take off until the mid to late fifties, this post-war period (1945-59) is seen as a 'golden age' of radio comedy: 'long-running and consistently funny shows reflected the kind of mood and optimism inherent in the nation at this time' (Foster and Furst, 1996, p123). For a considerable period, radio managed to retain substantial audiences (Briggs, 1995, p220), attracting all of the biggest stars and having a staggeringly wide variety of high quality output (Took, 1981, p93). Not only were there many types of shows, but they often had great durability throughout this era of great social change, most shows lasting five years, many ten (Crisell, 2002a, p75). Could it be that, *because* of the social change, they offered a constant that people could cling on to? Also, show longevity was aided on radio as it could maintain its stars without risk of them being lured away by television's big bucks, at least until commercial television launched in 1955. In the early fifties a good radio character performer could earn up to £700 a week; this was staggering when compared to a BBC radio producer earning just £700 per year (Took, 1981, p111). A couple of the most notable shows during this period were **Take It From Here** (1948-1960, 'The first radio show to emerge from the post-war comedy explosion', Nathan, 1971, p19) and **The Goon Show** (1951-1960 – 'one of the most influential post-war programmes', Foster and Furst, 1996, p103). Even **ITMA**'s cohesive appeal continued to be popular; despite the confidence in the new Britain, the post-war austerity was as controlling as it had been during the war, with rationing still in place well after its last episode in 1949 (Briggs, 1979, p711).

Research undertaken in 1946 had shown that audiences preferred continuing comedy shows rather than one-off programmes and there were indeed a number of long-running shows developed from that point that allowed listeners 'who hail with delight the familiar gags, follow with impassioned interest the varying and fantastic adventures of their favourite characters' (BBC Year Book, 1947, p50). Developing series was a key strategy for the department (BBC Year Book, 1948, p79), one that had proven successful in the US for many years (ibid, p55). So while directly after the war the majority of comedic output was from one-off programmes, by the 1950s the bulk were now series running for anything from 6 weeks to 9 months. A benefit of this strategy was an economic advantage of volume production (BBC Year Book, 1951, p142).

The fast paced, American humour still had its influence on continuing domestic comedy shows such as **Meet the Huggetts**, **Life with the Lyons**, and **The Clitheroe Kid**; they all offered comforting, gag-laden views of domestic life. These ran alongside the surreal comedy, silly voices and musical interludes of **The Goon Show**, and the catchphrases and gags of **ITMA** and **Take It From Here**. However, it was in the rejection of these now established radio conventions that a new style emerged. Ray Galton and Alan Simpson wanted to focus purely on character and situation, arguably *inventing* the 'sitcom' via **Hancock's Half Hour** (1954-1959) (Took, 1981, p128). This new format became very successful and inspired further long-running radio shows such as **The Navy Lark** (1959-1975) and **The Men from the Ministry** (1962-1977). However, despite its relatively recent invention, by the end of the 1950s sitcom on radio was seen to be a flagging medium, not because it didn't work on radio, but because more efforts were being put into developing television. In fact, not just sitcoms, not just radio comedy, but radio itself was facing a doubtful future. With the growth in television, driven in part by the Coronation in 1953 and the birth of ITV in 1955, radio audiences were

beginning to decrease rapidly. This created a mood of uncertainty for those involved in radio; would the new medium completely replace it? Clare Lawson Dick[66] wrote of the late 1950s that listeners were draining away, 'like the bath water when the plug is pulled out… None of us knew if, in the end, the bath would empty entirely' (Hendy, 2007, p28). From the time that commercial television pervaded homes and the Suez crisis meant that the sun's rays had permanently set over its Empire, setting about huge social changes (Jonathan Miller: *Auntie's War on Smut*, 2008), in Britain, the 'golden age' of radio comedy was over.

2.1.5 1960s

In the 1950s television was still a medium for the privileged and radio remained the main form of entertainment in the home (Foster and Furst, 1996, p125), but this was not the case in the 1960s as television became ubiquitous and stole audiences and talent from radio. Massive names such as Tony Hancock, Peter Sellers, Frankie Howerd and Kenneth Williams were on radio - where they had made their names - less and less (Foster and Furst, 1996, p235). While Crisell (2002a, p145) claims that there was no decline in quality during the 1960s, just declining audiences, others do regard it as a time of deterioration as money, and with it the best talent, had abandoned radio for television (Coward, 2003, p72). Foster and Furst (1996, p235) claim that this period 'marked a definite decline in quality and quantity of the BBC's radio comedy output', a time of 'low ebb' (Elmes, 2007, p21) for the genre.

Finding replacement stars was not easy either since, as quickly as television was growing, the music halls were dying (Briggs, 1979, p716). Therefore there was no longer that career path to the BBC, nor had there been a recent war to sweep up all barrack room humour (Foster and Furst, 1996, p248). Instead, there came a new breed that were to become the primary drivers of BBC radio comedy for the next fifty years. This new generation of comics were university bred, with their backgrounds in student revues that up to the late 1950s had been based mainly on poking fun at the working classes (Jonathan Miller, Nathan, 1971, p96). Off the back of the popularity of Cook, Moore, Bennett and Miller's *Beyond The Fringe* show, BBC radio producers began to make going to student revues part of their routine, just as producers used to tour concert parties before the war (Took, 1981, p162). It was at a Cambridge Footlights revue in 1963 that Peter Titheridge and Edward Taylor saw Humphrey Barclay, Bill Oddie, and John Cleese performing and, seeing their potential, brought them into the BBC fold. It was these three, along with other Cambridge alumni Graham Chapman, David Hatch, and Tim Brooke-Taylor, who went on to create one of the seminal shows of the 1960s: *I'm Sorry I'll Read That Again – ISIRTA* (1965-1973). While it had its genetics in the surreal lack of logic of *The Goon Show*, the innuendo and puns of *Beyond Our Ken* and the breakneck speed of *ITMA*, its young performers and updated style earned it a '*cult*' status (Briggs, 1995, p805), becoming 'a young person's *Round The Horne*' (Took, 1981, p162). The particular freshness of the programme came from challenging the comedy convention of the 'payoff':

> The next generation of comics – university bred, confident, arrogant – abandoned punch lines almost entirely, on the grounds that they were anachronistic and artificial. The real reason, of course, is that they couldn't be bothered to think of any. (Coward, 2003, p61)

The new guard however, did not completely dominate the airwaves as throughout the 1960s there were still many people who were not enamoured by the 'Oxbridge Mafia' (Coward, 2003, p71), including ex-serviceman and *The Goon Show* producer, Dennis Main Wilson (Nathan, 1971, p119-120):

> What bugs me about the university fellows is that they stand there in front of the cameras and entertain themselves, and if the public don't get it, it proves how clever they are.

---

[66] Later the Controller of Radio 4.

More traditional forms of radio comedy, owing their style to the variety shows of the past, still remained popular – such as **Round The Horne** (1965-1968) (Foster and Furst, 1996, p259). As a hotbed of innuendo and Polari, it still managed regular audiences of 15 million (ibid, p236), taking great pleasure in subverting the BBC's taste and decency rules. Despite the huge social changes taking place during the 1960s, with gambling, abortion, pornography and homosexuality all being decriminalised, there was still a prudery about sexuality that could only be tackled from an obtuse angle. Like *The Goons* before them, the writers considered any censorship to be a challenge and it was all the more funny for it.

> I, ironically, was the script editor who had the job of checking all scripts for taste. I never found anything with which I could take issue. If he wanted to say that he was dangling his grumdingles in the whatever, it meant nothing whatsoever, unless you decided you wanted it to mean something rude. (Edward Taylor: *Auntie's War on Smut*, 2008)

> Took [Barry] said that the most frequent question he was asked by fans was "How on earth did you get away with it?" Years after the show ended, Hugh Green [BBC Director General through the 1960s] is said to have told Took "To tell you the truth, I rather like a good dirty joke" (Coward, 2003, p67).

Another popular show at the time was **The Clitheroe Kid** which, although 'despised by intellectuals' (Coward, 2003, p70), was hugely successful in the ratings. In the mid-sixties, nearly a quarter of the population would be listening to its Sunday lunchtime broadcast. However, it does illustrate the way the tide was turning during this time as, despite its success, over its fourteen year run, its 10.5 million listeners dwindled to just 1 million. This decline was, ironically, in the face of an increase in numbers of radios being bought. The relatively cheap transistor radio meant that in 1967 twice as many radios were being bought versus around a decade earlier and people were able to listen in more places than just the front room with the rest of the family, where they now actually watched television. Listening did, in fact, increase at the end of the sixties, but it was during the daytime and it was through music stations (Radio's 1 to 4 having been created in 1967, replacing the Home and Light services) rather than speech formats (Hendy, 2007, p55).

Declining innovation in radio comedy in the 1960s might arguably be illustrated by some of the more popular shows being mere adaptations of television programmes such as **Steptoe and Son**, **Whack-O**, and **The Likely Lads** (Foster and Furst, 1996, p23). Where a great deal of radio comedy output once came from variety performances – relatively cheap and easy to make – not only was there little opportunity to find talent this way anymore, it was deeply unfashionable now, so there was little call for it as a radio show. This was partly due to the legacy of **Hancock's Half Hour** as it had challenged the validity of a radio show being purely a series of jokes; character and situation were now an integral part of many comedies. Frankie Howerd, speaking in 1970 said, 'I now prefer being funny in a situation, rather than just telling jokes. That after all isn't fashionable any longer' (Briggs, 1995, p215). So, what we might think of as straight *stand-up* comedy today was of little interest to the BBC, or to commercial stations where budgets did not allow for light entertainment when music could offer bigger audiences for less money.

During the 1960s there had been a comedy revolution that bypassed radio to a certain extent, that of satire. The influential *Beyond The Fringe* had in part opened the door for Ned Sherrin and David Frost et al to create *That Was The Week That Was – TW3* – broadcast on BBC television in the early 1960s. Unusually, compared to many other television comedies, it had not had to fight its way onto the screen via radio. It had been directly commissioned by Hugh Carlton Greene who had been brought in as Director-General in 1960 to update the 'Reithian fustiness' and encourage a 'new culture of irreverence' (Dominic Sandbrook: *Auntie's War on Smut*, 2008). It was a huge success:

> The public welcomed the irreverence with open arms. It started with about two and a half million and then went to four, five, six and hit twelve million for a show at 10:30pm at night. 10:30pm then was later than 10:30pm is now, too. It was unheard of. (David Frost: *Auntie's War on Smut*, 2008)

This development left radio behind, and went some way to further the perception of it being a moribund medium. The culture within radio Light Entertainment at the time was a little stagnant. The youngest and brightest producers had left radio to go to television, so there were now none left under 30 years old; most of those remaining had been recruited straight after the war, and were stereotyped as 'ex-bomber pilots', and being 'suspicious of the newfangled satire genre' (Hendy, 2007, p76). It was not until 1965 that Nicholas Parsons managed to persuade the BBC radio team to pilot a satirical show (Parsons, 2010, p82). It was not a programme that was widely supported by the radio comedy department but made its way onto the desk of Greene who gave it the green light (no pun intended). Many taboos had been challenged already by *TW3* but some were still in place in the slightly more old-fashioned medium of radio, and **Listen To This Space** had to fight its corner to be able to broadcast witticisms including names of newspapers or to make jokes about the royal family. Thus, despite the fact that this series petered out due to its writers leaving when they were refused an increase in salary (ibid, p82-90), it did open the door to radio satire which, in 1970, Radio 4 took the opportunity to walk through.

2.1.6 1970s

David Hatch and Simon Brett, supported by Tony Whitby the Controller of Radio 4 (all Oxbridge) developed a weekly topical satire programme which went on air from 1970 and ran for an astonishing 28 years (Took, 1981, p173). **Week Ending** (1970-1998) was developed to be a hard-hitting show, and indeed aimed to court controversy, having to be taken off air during election periods (Coward, 2003, p72). Indeed, in the late 1970s when Hatch was a little worried that it was getting a bit too mellow, he thought that 'a good libel suit was needed' (Hendy, 2007, p188.)

Initially its challenging content proved polarising both internally and with audiences (Took, 1981, p17). Within the BBC it was variously described as having 'comic inventions of a very high order' at the same time as being likened to 'drinking a glass of warm, flat beer' after a good dinner party (Elmes, 2007, p236-237); its lack of studio audience left it sounding rather cold. Listeners too were undecided and the first ever episode garnered an AI of just 38, with 50% of the 200 respondents giving it the lowest scores of C or C- (BBC Archives – Caversham). However, for this episode, 56 of the 200 people (28%) rated it as A or A+, giving an indication that there was a group who thought it had potential.

> Ratings proved stubbornly low for Weekending [*sic*] during its first year. There were also some crucifying reviews in the press. But those who did listen to it over the following eighteen months recorded a measure of appreciation that crept up slowly from a truly terrible RI of 38 to a much more satisfactory 60. By the mid-1970s, when it had become something of a minor cult among younger Radio 4 listeners, and especially among University students, the audience reaction figure had risen to as high as 77. (Hendy, 2007, p77)

Its longevity is in part due to the fashion in which it was made. Firstly, it had a number of different producers over its long life, ensuring that it had freshness about it from new ideas. It became a training ground, to a certain extent, for young, hungry producers including people such as John Lloyd, Douglas Adams, David Tyler, Jimmy Mulville, Armando Iannucci, Jon Magnusson, Geoffrey Perkins and Griff Rhys Jones. Secondly, the writing of the show was done by selecting pieces which could be written by absolutely anybody. This meant that there was no issue with a particular writer becoming burnt out. Paul Mayhew-Archer (told by a fan that he was the second most prolific producer of Week Ending) spoke of how the

process worked (Mayhew-Archer, 2011); picking-up the show after it had already been running for years he found it to be a really smooth operation. Monday or Tuesday there would be a meeting with the commissioned writers to go through the big stories of the week and allocate those stories to the writers. Wednesday was the day when the non-commissioned writers would come in and be briefed. Thursday was script day when they would all come in with their sketches, some of which would be sent away for rewriting. By the end of Thursday there would be a lot of material and Mayhew-Archer would take it away and choose around twenty items for the show so that it would be very pacey. He would work out the running order alternating quickies with longer pieces. Nothing lasted more than 90 seconds. At 9:30 on the Friday morning everyone met and it was all recorded by 1pm and edited in the afternoon by 4 or 5pm for a Friday night TX. Mayhew-Archer explained that as, over the years, there was a variety of producers of the show all having to cherry-pick from a vast array of material, the cleverer writers of *Week Ending* wrote in different styles depending on who the producer was. One writer told Mayhew-Archer that he had realised that he liked items containing music such as advert parodies, so they would offer him that kind of material, whereas another producer, Pete Atkin, liked dense political stuff so that was what they offered him.

Satire didn't have to be cold, dry and cutting. During the 1970s, another popular topical programme was *The News Huddlines* (1975-2001). With the same open door policy on material as *Week Ending*, it took a rather warmer stance as it was an audience show whereas *Week Ending* was not. It resulted in its satire actually being '*very jolly*' (Mayhew-Archer, 2011).

The topical shows of the 1970s meant that radio comedy maintained a hand on cutting-edge talent, but overall it was a rather dismal time for radio Light Entertainment. Shows left over from previous decades such as *The Clitheroe Kid* and *The Navy Lark* were hardly fashionable, and younger, fresher comics had been lured to television, such as the cast of *ISIRTA* going on to create *Monty Python's Flying Circus* and *The Goodies*. Many programmes from this time are now considered poor, or not considered at all: 'Non-topical sketch comedy in the 1970s covered the gamut, from dull to non-existent' (Coward, 2003, p72). It was a time of decreasing production values as BBC bands were cut down and recording venues closed (Foster and Furst, 1996, p265). Genres like comedy and drama were increasingly viewed – partly due to the publication of the 'Broadcasting in the Seventies' report from 1969 – as being more suitable for television (Hendy, 2007, p183). However, there was still hope. It was thought by Clare Lawson Dick, the then Controller of Radio 4, that listening figures could be boosted significantly, particularly with younger audiences, by a noteworthy radio comedy (Hendy, 2007, p186). Therefore, by 1977 The Light Entertainment department came under enormous pressure find the next cult shows in the way that programmes like *The Goon Show* and *Hancock's Half Hour* and *ISIRTA* had in previous decades. In 1978, Con Mahoney, who had worked for the BBC since the 1930s (Hendy, 2007, p76), was replaced as radio's Head of Light Entertainment by the much younger David Hatch, a highly experienced producer who had already worked on *Week Ending*, *ISIRTA* and *I'm Sorry I Haven't A Clue* amongst other things. When it came to developing new shows, there was little opportunity to get content from established writers as television had lured the best with wads of cash and BBC Radio could not compete. To overcome this problem, David Hatch advocated finding new writers who could be secured at lower cost. New talent was already on the doorstep via the open-door policy created by *Week Ending* and *The News Huddlines* and by the end of 1978 Hatch had nine new writers working for his team. Also, a competition for novice writers found a further thirty-five by the end of 1979, which included people such as Rory McGrath, Rob Grant, Doug Naylor and Jimmy Mulville (Hendy, 2007, p187-188).

In addition to a drive to find new talent, the seventies saw a search for new radio formats. Sitcoms were viewed as being far more suitable for television, and were relatively expensive to make in this time of

cuts (Took, 1981, p122). Stand-up too, its roots firmly embedded deeply in the variety tradition, was not seen as fashionable; at the time it was epitomised by ITV television's *The Comedians* (1971-1992), which survived on 'mother-in-law jokes (Hall, 2006, p4-5). To find something that worked, there were many experiments designed to give fresh impetus to the schedule. There were few that we would remember today as many failed to catch the public imagination and they would generally run for just a single series, two if they were lucky (Hendy, 2007, p185). However, one of these experiments did what was required of it: creating a cult. *The Burkiss Way* (1976-1980) was the first radio comedy since *ISIRTA* to get significant press attention, with the critics considering it a successor to *The Goon Show* (Coward, 2003, p72). It was meant to be polarising and it was:

> David Hatch was confident enough to describe the series as "almost a 'Monty Python' of the radio". It had not just captured a young audience, but exhibited many other features of the true cult: an audience that grew by word of mouth rather than official publicity, a language of its own that could be shared among fans, and a preparedness to exclude through provocation and obscurantism those who simply "did not get it". (Hendy, 2007, p190)

So, the new talent and culture in radio Light Entertainment at the end of the 1970s nurtured the environment that created a radio comedy which, as well as being a cult hit like *The Burkiss Way*, was oxymoronically a *popular* cult. Arguably one of the most successful and popular shows ever on BBC radio (books, television show, stage shows and even a Hollywood film followed), *The Hitchhiker's Guide To The Galaxy* was first transmitted on Wednesday 15th March 1978 in the 22:30 slot on Radio 4. Not only was it terribly funny, but it also appealed particularly to the 'geek' listeners, being based as it was on the science fiction genre. The first radio comedy to be broadcast in stereo (Armstrong, 2007), it took full advantage of the BBC Radiophonic Workshop to create a soundscape that created an amazing ambience for the show (Hawes, 2004, p86). *Hitchhiker's* really took advantage of the medium, creating a radiogenic format not fully realised since *The Goon Show*. This had exactly the effect that Lawson Dick had hoped for: bringing in new audiences to the station. What *Hitchhiker's* did 'was to change public perceptions of Radio Four [*sic*]. Several listeners have written of the series introducing them to the network' (Hendy, 2007, p193). So, despite its inauspicious position in the early 1970s, radio comedy ended the decade in triumph, regaining its position as a driver for new comedy, new talent and new ideas (Coward, 2003, p74); but that position was to be short-lived.

2.1.7 1980s

When David Hatch took over as Head of Radio Light Entertainment in 1978 he took on a BBC department that had been rather 'static' and 'traditional', and 'shook it warmly by the throat' (Took, 1981, p169). Radio Comedy was now a place largely populated by young Oxbridge types: as Coward puts it, 'Let's face it, if *Hitchhiker's Guide To The Galaxy* had been written by bloody Martin Smith from Croydon College of Further Education it would never have got on the air' (Coward, 2003, p74). These close connections may have been beneficial in creating the specific erudite comedy of student revues but the inward view meant that trends outside the circle were at risk of being missed. London's Comedy Store had opened in 1979, cultivating the new 'alternative comedy' trend: 'Radio took its time to really wake up to this new movement' (Foster and Furst, 1996, p267). However, while this new breed of comics were not necessarily Oxbridge, they were not exactly worthy of their working class claims either; Alexei Sayle, French and Saunders, Ben Elton, Adrian Edmondson and Rik Mayall all had various artistic and dramatic university backgrounds from London or Manchester (Fry, 2010 a, p209). But, it was not just the red brick wall that got in the way of alternative comics getting onto BBC radio. Channel 4 television launched in 1982 and part of its remit was to provide 'high-quality programmes for minorities' (Crisell, 2002a, p207); and like much commercial

broadcasting, it was keen to attract younger audiences in particular. The channel was eager to take risks so alternative comedy made sense; it employed young comics who would be relatively cheap offering a type of anti-establishment comedy (Foster and Furst, 1996, p267). So, Channel 4 scooped up many of the brightest and best with *The Comic Strip Presents…* series, but BBC television did not want to miss the boat and commissioned *The Young Ones*. This meant that a significant pool of talent had managed to go straight to television without having to bother with the more poorly paid medium of radio. Fry (2010, p207-208) found that the high profile of this kind of comedy at the beginning of the decade actually made him feel that being from an Oxbridge background was now actually detrimental to one's career.

So, in the early 1980s, 'comedy's laboratory' was now television (Coward, 2003, p75), and BBC radio was 'lagging behind' (Foster and Furst, 1996, p267). One of the most popular and longest running of the radio shows of the period was **Radio Active** (1981-87), notable for its Oxbridge roots, including Angus Deayton (Foster and Furst, 1996, p267, Coward, 2003, p75). It wasn't until later in the decade that BBC radio producers really began to take their inspiration once again from the live circuit (Foster and Furst, 1996, p267), but now, finding rather more blurred edges between the Oxbridge and alternative sets (Coward, 2003, p75). Jonathan James-Moore, who went on to become Radio's Head of Light Entertainment, was a great advocate of ensuring that producers were out and about, looking for new talent via stand-up and festivals. The upshot was that by the end of the decade Radio 4 had managed to catch up with television via shows such as **The Cabaret Upstairs** – the radio equivalent of *Saturday Live* (Punt, 2009) – featuring *The Comedy Store* compere Clive Anderson, Jeremy Hardy and, Arthur Smith, and **The Million Pound Radio Show** (1985-1992) including Andy Hamilton and Harry Enfield (Hendy, 2007, p310). However, while Radio 4 had caught up with television, the 1980s had hardly lived up to the standard that was set off the back of the success of **Hitchhiker's**. It was left to Radio 1, which had expanded its remit to include comedy shows in the evening, to be the instigator of something really exciting right at the end of the decade. After a brief sojourn with Patrick Marber's **Hey Rrradio!!!** (Punt, 2009), **The Mary Whitehouse Experience** (1989-90) really found an audience:

> This was a fast paced show that once again mixed-up various elements – stand-up, topicality, sketch humour – and put them together in a more up-to-date package that appealed to a younger audience. Most Radio 1 listeners would probably never have thought of listening to any radio comedy if it weren't for this. (Foster and Furst, 1996, p267-268)

The success of television comedies for younger audiences such as the aforementioned *The Comic Strip Presents…* and *Friday Night Live* had been satisfying the appetite for cult comedy but it was this show – devised by Bill Dare and starring Rob Newman, David Baddiel, Steve Punt and Hugh Dennis – that managed to lure the younger audiences back to radio for comedy. The success of **The Mary Whitehouse Experience** lead to a television version and ultimately, a famously huge live performance at Wembley, causing comedy to be talked of as the new rock'n'roll (Hall, 2006, p90): a rock'n'roll song sung, once again by Oxbridge graduates.

2.1.8 1990s onwards and changing fashions

At the end of the 1980s, a couple of Radio 4 shows made the transition to television but **After Henry** went to ITV (Street, 2002, p129), rejected by BBC television for being too middle-class, and **Whose Line Is It Anyway?** was picked-up by Channel 4 when BBC television was too slow to respond. Clive Anderson, its host, was told by the BBC that. '"Possibly we may have some interest at some stage in talking about maybe making a pilot" the team had to respond by saying "Actually, we're starting filming a thirteen-part series for

Channel 4 on Monday"' (Osborne, 2008, p143). John Birt, Director-General in the early 1990s, was keen to stem this leak and tightened up the links between radio and television in terms of money, talent and resources (Hendy, 2007, p376), incentivised by the possibilities illustrated by ***The Mary Whitehouse Experience.*** This process improvement allowed for a smoother flow for radio comedies to be picked-up by television. Another key change was in the late 1980s Michael Green, the Controller of Radio 4, opened up a late night slot for comedy, allowing the station to trial more high risk material at a time when there would be fewer listeners and when they might be more open to innovation (Elmes, 2007, p230). It was these two changes that went some way to improve the content and profile of Radio 4 comedy, and introduce a period respectively described as 'a resurgence in radio comedy' (Foster and Furst, 1996, p268), 'the heyday of this cross-fertilization' (Hall, 2006, p10) and 'Radio Four's [*sic*] comedy renaissance' (Hendy, 2007, p369). This career path from radio to television, although not new (for example, ***Hancock's Half Hour***, 1950s and ***Hitchhiker's***, 1980s), was more prolific and successful than it had been for a long time. Birt's initiative built a more sturdy bridge that remained stable throughout the 1990s and 2000s, seen through some of these examples of Radio 4 comedies that moved to television during this period:

Title	Radio series dates	Television series dates
On the Hour (as *The Day Today*)	1991-92	1994
Knowing Me, Knowing You (Alan Partridge)	1992-93	1994
Room 101 (Old Radio 5)	1992-94	1994-
They Think It's All Over (Old Radio 5)	1992-93	1995-2006
People Like Us	1995-97	1999-2001
Goodness Gracious Me	1996-98	1998-2001
On the Town with the League of Gentlemen (*The League Of Gentlemen*)	1997	1999-2002
Sean Lock's 15 Storeys High	1998-2000	2002-04
Absolute Power	2000-06	2003-05
Dead Ringers	2000-14	2002-07
Little Britain	2000-02	2003-09
The Boosh (as The Mighty Boosh)	2001	2004-07
That Mitchell and Webb Sound (*That Mitchell and Webb Look*)	2003-	2006-10
Flight of the Conchords (in-house for R2/HBO)	2004	2007-09
Count Arthur Strong's Radio Show! (*Count Arthur Strong*)	2005-	2013-
Genius	2005-08	2009-10
Down the Line (*Bellamy's People*)	2006-	2010
The Bleak Old Shop of Stuff (*Bleak Expectations*)	2007-12	2011
The Cowards	2007-08	2009
I've Never Seen Star Wars	2008-	2009-11

The 1990s, in particular, saw a particular type of comedy become popular, that of genre parody (Foster and Furst, 1996, p268). ***On the Hour*** was a spoof news programme, ***Knowing Me, Knowing You*** a chat show and ***People Like Us*** a documentary. These three, and later ***Down the Line***, in parodying types of radio shows, seemingly '*biting the hand that fed them*' (Hendy, 2007, p369).

'Mocumentary' style shows such as ***People Like Us*** became particularly popular on television, the naturalistic approach's genesis in programmes such as Kelly Monteith's 1970s sitcom, and later *The Larry Sanders Show* and the film *This Is Spinal Tap*. It was the phenomenal success of *The Office*, however, that cemented the fashion of *non-audience* shows on television (Mayhew-Archer, 2010), the critics now disliking the now outmoded laugh track (Mitchell, 2010). Flying in the face of fashion, Radio 4 had not abandoned its

live audience shows, particularly in the 18:30 time slot. Indeed, it could now be argued that since the financial downturn of the late 2000s, audiences are once again more drawn to the warmer styles of audiences shows such as *Mrs Brown's Boys* and *Miranda* on BBC1 (Chortle, 2011a), with the BBC keen on cheering up the nation just as it did during wartime:

> It might not be the cheeriest of times, but the recession has sparked a golden age of comedy on British television. Commissioning Editors have woken up to the fact that viewers want to be entertained by traditional comedies packed with jokes and well-known actors. (Brown, 2012)

On Radio 4 too this is evident, as all of the comedy programmes with the highest mean AI scores are the traditional audience comedy shows that have been around for years: ***JAM*** (1967-), ***Clue*** (1972-), and ***The News Quiz*** (1977-). All of these shows have been running for years and part of the appeal is their intrinsic longevity, inciting familiarity and nostalgia for many listeners. All of them have been at risk of cancellation as key talent has died (Kenneth Williams and Peter Jones, Willie Rushton and Humphrey Lyttlelton, Alan Coren and Linda Smith respectively) but all have managed to replace lost talent and keep current enough without losing any of the warm familiarity that makes them popular.

Alternative comedy did not go away in the 1990s but merely blended in with what we have today, becoming part of the mainstream. Its legacy can in part be seen in the 'lecture style of stand-up which has endured on radio more than on television. For example, Radio 4 shows such as ***Jeremy Hardy Speaks to the Nation***, ***Mark Steel's in Town***, ***Mark Thomas: The Manifesto***, ***Mark Watson's Live Address to the Nation***, ***Tom Wrigglesworth's Open Letters*** and ***Susan Calman Is Convicted*** allow shows focused on one theme or topic. This sanctions Radio 4 comics to do stand-up in a slightly more intelligent way, without resorting to the popularist television styles of *Live at the Apollo* and *Michael McIntyre's Comedy Roadshow*. These television shows, slick-looking but cheap to make, have become ubiquitous and as a result are beginning to become unfashionable:

> *I suspect that we're heading towards the death of stand-up. I think that the Michael McIntyre Roadshow and the Live at the Apollo... we've seen them now. They'll play out on Dave forever and a day. I wouldn't be at all surprised if the big television cheap fillers may lose their currency after a while.* (Raphael, 2012a)

In that respect, Radio 4's inclination to continue to approach stand-up in a more oblique way seems the correct move. On the other hand, in terms of sketch performance, Radio 4 had gone back to basics. ***Sketchorama*** is taking sketch groups off the circuit and cherry picking sketches from them to pull them together in one programme. Although this approach has none of the complexity of something like ***On the Town with the League Of Gentlemen,*** which pulled its sketches together, themed through their locale, or ***Goodness Gracious Me,*** which had a strong thematic thread, it does allow the listener to hear a wide range of talent. Where these new, young, relatively untried sketch groups would be unlikely to get a commission for a full series, they get the chance to showcase their best bits. This echoes earlier programmes such as ***The Cabaret Upstairs*** and, indeed, the very nature of variety that was the first type of comedy to be heard on the radio.

A further legacy of alternative comedy is that of political correctness, which itself caused a reaction resulting in a kind of meta-bigotry, an acceptable route for arguably *tasteless* humour. Epitomised initially through characters such as Alf Garnett, it has been heard more recently on Radio 4 through comic characters and comic personae's jokes, such as:

- Alan Partridge (***Knowing Me, Knowing You***) – 'Please don't write in saying that's sexist – it's not.'

- Gary Bellamy (***Down the Line***) – 'Some people have been very angry, but often that's a reaction that takes place when you don't understand something.'

- Jimmy Carr (***Chain Reaction***) – 'I'm what they call a plastic paddy. I've got an Irish passport, Irish parents, born in Ireland, but I speak and present myself in this way because I was raised and educated in the Home Counties. Which just goes to show what you can do when you apply yourselves.'

- David Mitchell (***The Unbelievable Truth***) – 'Often things we thought might be true turn out not to be. For example, there's absolutely no truth in the rumour that the last entry in Anne Frank's diary reads: "Today is my birthday, Dad bought me a drum kit."'

All these jokes are told by characters that have established themselves as being rather ignorant, or in the case of Carr and Mitchell, fostering comic personae that are rather crueller and more cutting than they might be in real life. In the case of characters, clearly the performer is just playing a part and is unlikely to hold the same beliefs as the character they are playing. In regards to the comedians, it can be more of a problem as, for example, Jimmy Carr has the same name as his on-stage persona, so it can be easier for audiences to think that he might hold racist or sexist views as heard in his jokes (Mitchell, 2010). The comic characters once again allow comics to say things that might have seemed unsavoury in earlier years, aided by racial or gender stereotypes being replaced by social stereotypes (Hall, 2006, p166); so we are hearing the same jokes but 'shifting our taboos to soothe our consciences' (Carr and Greeves, 2006, p196). Un-PC jokes from the 1970s were spurned by the 1980s alternative comics but have been revitalised and are now used under the guise of postmodern irony; we are *allowed* to find them funny again. We may have never stopped enjoying them amongst friends, but you might be more likely to now hear them broadcast. Comedy, its style and content, is cyclical (McGowan, 2008). Just as the Green Book censored controversial topics 50 years ago, 'If today, the Green Book looks to us ludicrously old-fashioned, who's to say that 50 years from now, our own values and our own panics and hysterias won't look equally ludicrous to our successors'. (Dominic Sandbrook: *Auntie's War on Smut*, 2008).

> Comedy's future is all about its past. The scarcity of certain types of comedians will remedy itself, as new trends are sparked off by successful acts and backlashes against the status quo set in. Comedy is – and always will be – a world where styles drift in an out of fashion. (Hall, 2006, p20)

So, for radio comedy, we can see a number of areas where over its 90-year lifespan its cyclical nature is present:

- Jokes – Subjects we might have enjoyed but were then thought unacceptable can once again be expressed via comic persona.

- Styles of comedy – Fashions exist in comedy and they come around more than once. We are seemingly entering a period where stand-up may begin to feel old fashioned while variety and knock-about sitcom could be all the rage; this is a situation which might have seemed impossible in the 1960s after the decline in music hall, and the newer subtlety of programmes like Hancock's Half Hour and the sophistication of satire. 'Music-hall has been proclaimed dead and buried on innumerable occasions. The reports have been grossly exaggerated' (Nathan, 1971, p16).

- Missing the boat – Radio was not in the vanguard of satire in the early 1960s, nor alternative comedy in the early 1980s. Does the growth of social networking and self-broadcasting mean that a similar risk exists in the early 2010s? Will BBC radio miss the next big thing in comedy, whatever that might be?

- The pull of the medium – At the beginning of radio's existence, it was resisted by talent as it was seen to be detrimental to careers. This changed as it was seen that it could create huge stars, but more recently radio is no longer able to offer many performers a career solely in the one medium due to low budgets.

Only retrospection will give us a clear view of successful Radio 4 comedy shows of the 2010s but the following are worthy of consideration as shows that have been prominent thus far:

SITCOMS
Ed Reardon's Week
Cabin Pressure
Clare in the Community
In and Out of the Kitchen
Count Arthur Strong's Radio Show
PANEL SHOWS
Just a Minute
I'm Sorry I Haven't A Clue
The News Quiz
The Unbelievable Truth
Dilemma
COMIC LECTURE / STORY TELLING / STAND-UP
Tom Wrigglesworth's Open Letters
Mark Watson's Live Address to the Nation
Mark Steel's in Town
Meet David Sedaris
Bridget Christie Minds the Gap
SKETCH
Sketchorama
John Finnemore's Souvenir Programme
INTERVIEW
Chain Reaction
My Teenage Diary

2.2 Comedy's Place on Radio 4

2.2.1 Radio 4

Radio was hit hard as a primary form of entertainment when television's popularity grew in the 1950s, resulting in a huge decline in radio audiences 'that some thought would prove terminal' (Crisell, 1994, p27). The average evening audience dropped from nearly 9 million in 1949 to 3.5 million in 1958 (ibid). However, it has been radio's status as a secondary medium that has seen it survive in the following decades (Shingler and Wieringa, 1998, px), despite expectations to the contrary, and today 91% of the UK population listens to radio at least once a week (RAJAR, Q4 2013).

BBC Radio 4 began in 1967 when Frank Gillard, Director of Sound Broadcasting 1964-69, restructured BBC Radio (Briggs, 1995, p577-578). This change was driven by a need to introduce a popular music station - Radio 1 - that would service the audience soon to be left bereft because of the new Marine Broadcasting Act of 1967, which took pirate stations off the air (Crisell, 2002b, p127). Radio 4 inherited both speech and music programmes from the Home Service and also from the Light and BBC's overseas services, and in doing so was rather a mishmash of content (Hendy, 2007, p14, Crisell, 2002b, p127). Over the following years different Controllers took varying approaches to updating the station. Through the 1970s, the 1980s and into the 1990s, Radio 4 was euphemistically alluded to by its Controllers in terms of an ancestral estate, with its upkeep and alterations to be done merely with trowels and feather dusters, not shovels and sledgehammers.[67] Tony Whitby talked of 'weeding here a little, weeding there a little' (Hendy, 2007, p274), and Clare Lawson Dick described her job as 'a constant weeding out of tired or out-of-date programmes' (ibid, p273). David Hatch said of Radio 4 that 'to be its Controller was like inheriting a long-established country estate that had to be handed on intact' (ibid, p299), and Michael Green described his changes as simply 'moving the furniture' (Elmes, 2007, p110). James Boyle, who became Controller in 1996, was the

---

[67] Full list of BBC Radio 4 Controllers: Hendy, 2007, p405.

first to bring in the wrecking-ball. He found that Radio 4, having been maintained but never fully renovated, had become tangled undergrowth for the listener and was a veritable 'warren of programmes' (ibid, pxi). He felt that the schedule's gradual evolution over 30 years was due mainly to sentiment and inheritance rather than any specific planning, and that it would never be created from scratch as it was (Barnard, 2000, p40). Although research in the late 1980s and throughout the 1990s indicated that listeners didn't want change (Hendy, 2007, p294-5), Boyle commissioned extensive research to give him the background to effect a 'radical re-engineering' (ibid, p392) of Radio 4 from April 1998. The goal of this update was for the station to improve its appeal to its existing listeners while also attracting new audiences (Elmes, 2007, p206). The new schedule aimed to simplify the timetable for the listeners by having consistent start and end times for programmes and by reflecting the moods and activities of those available to listen (Hendy, 2007, p392). Boyle also oversaw the separation of production (supply) and commissioning (buying) into distinct functions (ibid, p289). Up to this point, department heads had managed their own business, but the introduction of Commissioning Editors meant that there was a more detached overview at network level, with individuals being able to compare in-house and independent productions without a vested interest in the in-house business.

Since the major schedule changes initiated by Boyle's extensive 1997 research, any further updates for Radio 4 have been relatively minor. Mark Damazer's biggest modifications were arguably the removal of the UK theme tune in 2006 and the children's programme *Go4it* in 2009. Gwyneth Williams, the incumbent Controller, is again treading carefully around the metaphorical flowers and family portraits; her biggest change to date was arguably the extension of *World at One* to 45 minutes from 30 minutes in November 2011. Williams has, however, been instrumental in the biggest change in Radio 4's comedy offering in the past 17 years. She facilitated the introduction of the Sunday 19:15 comedy half hour that replaced Damazer's *Americana* that in turn had replaced *Go4it*.

Despite Boyle's attempt at simplification for the audience, Crisell (2002b, p130-131) claims that Radio 4, alone in the BBC's radio offering, retains elements of the Reithian principle of serendipity. Still a relatively complex schedule, to a certain extent, a listener who chooses Radio 4 allows the station to curate their consumption of broadcast audio speech. Few people make appointments to listen to specific shows on radio over and above their routine (Raphael 2012a), so they will hear what the programmers choose for them. Chance hearings can be seen as a positive:

> I just turn on Radio 4 and listen to what's there, I don't ever look in the Radio Times or
> whatever, I don't have to (Male, Radio 4 Listener, ABC1, 40-55, Surrey).
> (Counterpoint Research, 1997a, p11)

Nowadays, Radio 4 offers the following:

> The remit of Radio 4 is to be a mixed speech service, offering in-depth news and current
> affairs and a wide range of other speech output including drama, readings, comedy, factual
> and magazine programmes.
> (BBC Executive, 2010, p3)

Its varied genres means that, unlike some other services, it does not aim to target a certain demographic group, instead offering programming to 'everyone in the UK interested in intelligent speech radio' (BBC Executive, 2010, p14) – a 'very broad audience' (Pilgrim, 2009). Radio 4, along with its sister station Radio 4 Extra (previously BBC7), have a 'de facto monopoly' on national speech radio in the UK (Aitkin, 2007, p15).

2.2.2 The Radio 4 Audience

An important group for the network are the ABC1 35-54 year olds – younger than the core ABC1 55+ listeners. They are known as the 'replenisher' audience (BBC Executive, 2010, p17), so called as they replenish the total as the older listeners expire. Share of listening within this group has been broadly in decline throughout the twenty-first century – the reach of 20% in Q1 1999 has dipped as low as 12.1% in Q1 2012 (RAJAR). This led to a dire prediction of a declining audience in a 2006 report from a management consultancy:

> Radio 4 is expected to lose listeners in all scenarios. (BBC A&M Report, 2008)

The anxiety continued:

> In the last decade a cause for some concern has been the declining rate of uptake among these "Replenishers". (BBC Executive, 2010, p17)

However, despite the replenisher decline, Radio 4's reach remains significant, attracting 21% of the adult population in any one week – around 11.2 million people listening for at least 5 minutes – and taking 12.5% of all linear radio listening (RAJAR, Q4 2013). The mean age of the Radio 4 audience is 55 (RAJAR, Q4 2013) and has remained static over the past decade despite the aging UK population and declining replenisher audience (BBC Executive, 2010, p16). In fact, it has changed very little since 1967 when it started at a mean of 53 (Hendy, 2007, p267).

2.2.3 The Audience for Radio 4 Comedy

RAJAR figures for Q4 2013 (BBC Audiences Portal) show that Radio 4 comedy alone, through its origination and repeat slots, reaches a total of 5.5m listeners in any one week, around half of all Radio 4 audience. Radio 4 comedy origination slots attract audiences as follows:

Figure 3 – Summary of Radio 4's half hour comedy slots – reach in 000s. (RAJAR, Q4 2013)[68]

Slot (28')	Day	000s listeners (1/2 hour reach)[69]
11:30	Monday	989
	Wednesday	843
	Friday	771
18:30	Tuesday	1,330
	Wednesday	1,210
	Thursday	1,080
18:30	Monday	1,520
	Friday	1,220
23:02	Tuesday	678
	Wednesday	615
	Thursday	575
19:15	Sunday	521

Information in figure 3 indicates that there is a considerable audience for each of the slots by any broadcaster's standards. Repeat and non-linear figures will add to the overall Radio 4 comedy listening. For example, the Monday 18:30 slot, where *I'm Sorry I Haven't A Clue* and *Just a Minute* TX,[70] gets a repeat on Radio 4 on Sunday at noon which attracts a further 1,210,000 listeners. The total reach for the two R4

---

[68] In this chart and figures 4 and 6, 18:30 slots are separated into Mondays and Fridays versus Tuesdays to Thursdays. This is because the Monday and Friday slots contain the core Radio 4 comedy programmes, all of which get narrative repeats the following weekends. For programming purposes, they are treated differently to the other 18:30 slots.

[69] These are average figures across a quarter; RAJAR reports at this level of granularity.

[70] TX is a widely used term to denote transmission or broadcast. Similarly, RX is used for recording, although less commonly used.

TXs is smaller than the sum of the two slots at 2,440,000 (ibid), as there is a small amount of duplication across the two broadcasts. This slot also get repeats at Monday 07:30-08:00 and 22:00-22:30 on R4 Extra where the reach totals 290,000 (ibid). The BBC's Audiences Portal also gives figures for listening on iPlayer (live and catch-up). The episode of *Clue* which originated on 16th December 2013, for example, was available online for 13 days, during which time it was heard through iPlayer over 122,000 times (BBC Audiences Portal: accessed 10/04/2014). Differences in measurement methodologies across platforms do not allow us to add the figures together for a total reach figure.[71]

'Radio' Comedy?

While Radio 4's main routes of delivery are via FM and LW radio, digital routes such as IP, DTV and DAB[72] are growing and there is additional incremental listening via non-linear methods such as podcasts and iPlayer. These might, arguably, not be called 'radio'. Thus, discussion should ideally speak of 'audio comedy' instead. However, for this study, the term 'radio comedy' is used as it is more familiar for the reader, albeit done so with the awareness that the focus is the product rather than its route to the listener.

2.3 Features of Radio Comedy

Having the quality of a particular aptness for radio is known as being 'radiogenic' (Chignell, 2009, p93), although, in light of the proliferation of delivery systems, the term *audiogenic* might be considered more appropriate if non-linear listening were included.

Some comedies have specific reasons for being on the radio rather than any other medium. The first radio comedy programme on the BBC, heard in 1924, was *A Comedy of Danger*, which was set in the pitch-darkness of a coal mine where miners were trapped (Crisell, 1994, p157). This programme was designed to take advantage of the 'invisibility' (Shingler & Wieringa, 1998, p80) of radio and its listeners were told to listen to it in total darkness to fully appreciate the nature of the setting for the story. However, over the years, BBC radio comedy has spanned the spectrum of being particularly apt for radio, for example, *It's That Man Again* (*ITMA*), *The Goon Show* and *The Hitchhiker's Guide to the Galaxy*, to presenting comedy that has actually been developed for other media and is merely broadcast on radio as a secondary outlet. Examples of this latter type include: early variety broadcasts direct from music halls,[73] television shows adapted for radio such as *Steptoe and Son* and *To the Manor Born*, and television pilots broadcast first on radio as a cost saving exercise to ascertain viability (for example *Happy Mondays: Pappy's Fun Club* broadcast on Radio 4 in 2008 but paid for by BBC television).

Early radio research found that people prefer to hear a joke rather than read it (Cantril and Allport, 1935, p99, p221), but this does not mean radio is the perfect medium for humour. Radio comedy lacks the visual cues of television and the group atmosphere of live performance. However, the enduring popularity of some radio comedy shows such as *JAM* and *Clue*, running for decades and attracting audiences exceeding that of some digital television stations, prove that comedy can be very successful on the radio.

Radio comedy can be seen as both deficient for its invisibility and successful because of it (Shingler and Wieringa, 1998, p81), seen by some as lacking and by others as 'the most effective humour delivery system yet devised' (Coward, 2003, p7).

---

[71] Figures cannot be simply added together as there may be duplication.

[72] IP is listening to radio over the internet, DTV is radio heard through digital television and DAB is a form digital radio.

[73] See from p15.

> There is something odd about the success and importance of radio comedy. Why is something so inherently visual… a success on an invisible medium? To take this point further, comedy has been a vital ingredient in radio's development and success but even highly visual forms of comedy, including ventriloquism, have worked well… Some comedy clearly works well on radio; the simply narrated joke for example or the sitcom based on strong and familiar characters and a good script. But comedy can also exploit the invisibility (or 'blindness') of radio; the iconic 1950s radio comedy, The Goon Show, being a good example. (Chignell, 2009, p13)

The lack of visuals can be used to its advantage, allowing the imagination to fill in the blanks.

> There are some kinds of humour which can work only – or work best – in a purely spoken medium. The absence of a visual element allows ideas to work with the full range of the human imagination. (Ross 2003, p90)

> *I think audio is a really good way of consuming comedy… better than television for stand-up, and for lots of other forms, just because it's more immersive.* (Bennett, 2011)

Radio is often consumed alone (Crisell, 1994, p11-12) but theories posit that the ideal way to experience comedy is in a group (see from). For radio, the presence of a studio audience at a radio recording can replicate the group experience for the listener, their laughter being a proxy companion that diminishes the detrimental effect of listening alone by narrowing the distance between the performer and the listener at home (McQuail, 1997, p117).

While there may be certain elements of comedy that are particularly suitable for radio, the fact that a range of comedy programmes have moved from radio to television and vice versa[74] illustrates that comedy is a versatile genre.

2.3.1 Radio Comedy as a Genre

Genres are a type of categorisation providing shorthand for the industry and allowing recognition of the distinctions between different types of radio programmes, providing conventions and aiding production, commissioning and scheduling processes (Barnard, 2000, p107). The way that the genres are split can be based on a number of criteria and may vary based on the requirements of the users. Programmes can be grouped based on their style, approach, production techniques or purpose (Starkey, 2004, p239). For example, if 'radio panel shows' were to be considered as a genre, they would share production techniques regardless of whether the aim was to be comedic, such as *Clue,* or informative, such as *Gardeners' Question Time.* Radio 4 programming, however, is generally segmented based on its *purpose*. This can be seen in a variety of aspects of Radio 4: the way the programmes are commissioned (by a Commissioning Editor specifically for Comedy), by the service licence requirements[75] (a minimum number of radio comedy hours being a requirement, BBC Executive, 2010, p35), and how it faces its audiences (for example, BBC iPlayer genre segmentation).

Radio comedy is not an immutable term. The BBC production department which is responsible for the majority of the comedic output on Radio 4 was, in the early years, called 'Variety'. But this department also encompassed a wide range of programme types under its umbrella (BBC Handbook, 1956, p83):

---

[74] See .

[75] See from .

The personality-type show –	e.g., Peter Ustinov in ***In All Directions***
The broad comedy show –	e.g., ***Take It From Here***, ***The Goon Show***
The domestic situation comedy –	e.g., ***Life with the Lyons***
The act type show –	e.g., ***Variety Playhouse***
The light-dramatic show –	e.g., *Journey into Space*
Quiz programmes –	e.g., *Twenty Questions*
Interest programmes –	e.g., *In Town Tonight* (an early chat show format)
Musical programmes –	e.g., the BBC Show Band

At the end of the 1950s the 'BBC Variety' Department became 'Light Entertainment (sound)' (BBC Handbook, 1958, p240), but despite the name change it generally covered the same sub-genres (ibid, p90-91). Only in 2008 was the name changed to 'Radio Comedy'. Over these decades, the department ceased to produce programmes unrelated to comedy, such as purely music shows or chat shows without a comedic element. Yet it still retains *Quote, Unquote* which is quiz rather than a comedy, remaining under the Department's section merely as a relic.

In Radio 4's commissioning guidelines, there is no definition of what exactly radio comedy is. The best it offers is an indication of what might be expected of each of the slots, differentiating the requirements between them. For example, the 11:30 slot on Monday, Wednesday and Friday has the following information (BBC Radio 4 Commissioning Guidelines Spring 2014, p20):

> This slot introduces a lighter note to the mid-morning schedule.
> It is the main slot for situation comedies, entertaining light dramatisations and comedy dramas. We will also broadcast sketch shows and other new formats. Shows recorded in front of an audience bring energy and warmth to this time of the day and repeat well at 18:30.
> This round the slot is only open for the following submissions:
>
> - Audience situation comedies.
> - Audience sketch shows with experienced performers and writers.

There is no definition given for radio 'situation comedy', nor for any other sub-genre, and the other slots have even less prescriptive requirements than this 11:30 summary. The implicit message is that only those with proven understanding of the genre will gain commissions:

> Therefore, to stand a chance of being successful, your company or department will need to be able to demonstrate substantial and considerable experience in radio comedy and/or television comedy.
> (BBC Radio 4 Commissioning Guidelines Spring 2014, p15).

From the listeners' point of view, 'radio comedy' appears to have a wider definition than the specific requirements of Radio 4's commissioning process. Research in 2005 (BBC Audience Research, 2005) found that most people thought of 'presenter-led banter and wind-ups' when asked about radio comedy. A BBC report (BBC Executive, 2010, p32) also revealed this mix of understanding: the results of asking respondents which station was best for 'comedy' highlighted that Radio 4 was the leader at 13% in 2009/10 but also found that 10% said 'Any commercial', 9% said Radio 1, and 8% said Radio 2. This showed that DJ banter and chat (stations other than Radio 4 have limited or no 'traditional' style radio comedy) are still considered, by audiences, as radio comedy. This varied understanding of the term is not limited to the public. For example, the Radio Academy Awards (the most high profile of all UK radio prizes) have the following criteria for the radio comedy category:

> **D4: Best Comedy**
> This category is for any programme or series in which the prime objective is to make listeners laugh. This could be a one-off programme, a regularly scheduled programme/series or a limited-run series. Judges will be looking to reward comedy programmes with high production values and a true sense of what will make their target audience laugh.
> (Radio Academy Awards, 2012)

This is a relatively broad definition of radio comedy.[76] Radio 4 definitions of comedy (occasionally called radio 'comedy and entertainment' [BBC Radio 4 Commissioning Guidelines, 2012, p46]) do not include the type of DJ banter that many of the general public and indeed the Radio Academy include in 'radio comedy'. *The Adam and Joe Show*, for example, has won the award for best comedy show (in 2010), but the format and style - mixing comedy with music, chat and news - would exclude it from the Radio 4 comedy portfolio, which while allowing a spectrum of styles, tends towards a more structured format focused on comedy material alone. An improvised panel show before an audience, such as *The News Quiz*, might be at one end of the spectrum, while a tightly scripted narrative recorded in a studio, such as *Brian Gulliver's Travels*, could be at the other.

2.3.2 Sub-genres of Radio Comedy

Within the genre of Radio 4 comedy there are huge variations in style, approach and production techniques, albeit with the common goal of amusing the listeners. Radio 4 comedy reaches that objective via a number of routes. As genres can be split across varying criteria, so programmes within a genre can be split into sub-genres (Starkey, 2004, p247). This kind of categorisation is used by the Commissioning Editor for Radio 4 comedy to ensure that there is a mix of 'types' of comedy across the output (Raphael, 2012b). Segmenting into sub-genres is a difficult process in radio comedy as the programmes be segmented a number of ways, such as by production technique (for example, panel show versus sitcom) or style (for example, surreal versus satire). Even if categories are agreed upon, there is much subjectivity in deciding into which group a programme can be placed. Steven Canny (2011) explaining *Brian Gulliver's Travels*:

> [It is] *effectively a series of connected sketches. I have tried to create the illusion that it's not. As a producer, my job is to make it seem like a consistent narrative. So, I've always called it a sitcom. Caroline* [Raphael, the Commissioning Editor] *and Jane* [Berthoud, the Head Of BBC Radio Comedy] *disagree and don't call it a sitcom, rather a sketch or broken comedy. Bill* [Dare, the writer] *has no particular feeling about it.*

Indeed, programmes can be described as a mix of more than one sub-genre, for example, *On the Town with the League of Gentlemen*, described as a 'mixture of sketch show and sitcom' (Coward, 2003, p83) or, similarly, *Radio Active,* a 'hybrid of the sitcom and the sketch show' (Starkey, 2004, p167). Sometimes, programme content can be a deliberate fusion of genres, the very nature of the output intended to challenge the perceptions of the listener and subvert the conventions of an established type of radio. Parodies like *Knowing Me, Knowing You with Alan Partridge*, *On the Hour* and *Down the Line* are comedies in the style of factual content - a chat show, a news show and a phone-in show respectively.

2.4 Comedy and the BBC's Public Service Role

Because of the unique way the BBC is funded, radio comedy has long held a 'privileged position in the UK' compared to the rest of the world (McWhinnie, 1959, p96).

2.4.1 Culture

Radio 4 has a commitment to provide comedy as part of its output. In any financial year, one condition of its service licence demands that there must be at least 180 hours of original (i.e., not including repeats) comedy broadcast (BBC Executive, 2010, p35). This delivery is part of Radio 4's Statement of Programme Policy (SoPPs). Part of the reason for comedy's presence in the schedule is to fulfil a cultural requirement relating to the public purpose of 'Stimulating creativity and cultural excellence' (ibid).

---

[76] There are further technical requirements.

You cannot operate the BBC without Comedy at its heart. (Davie, 2009)

The Trust's audience research for Radio 4's last service review found that comedy is the *key* genre in fulfilling the creativity requirement:

> Comedy Programming – although somewhat hit and miss in terms of popularity, is frequently offered by listeners as evidence of entertainment and creative output, and often acknowledged as the starting point for new, popular titles that have made their way onto TV, including *Mock the Week, The Day Today* and *Little Britain*. (BBC Trust, 2010, p4, p44)

However, whilst radio comedy is recognised for its cultural influence, there are aspects that challenge this:

1. In terms of the SoPP delivery, the cultural requirement for comedy on Radio 4 at 180 hours per year is far lower than that of drama and readings (for example) at 600 hours indicating that it is of lesser importance.

2. Radio comedy is part of our culture but perhaps only because of the way that BBC funding allows genres to continue when it might not be financially viable in the commercial arena. Bernard (2000, p109-110) argues that radio comedy may only continue to exist in the UK due to residual loyalty and tradition. In the US for example, while it has had a tradition of radio comedy, today it does not have the same level of penetration that is found in the UK. But, can it be said that its comedy suffers because of a lack of radio comedy now? Long-running television comedies developed since the US's decline in radio listening have managed to achieved global appeal, for example, *Mash*, *Cheers*, *Frasier, Friends* and, more recently*, The Big Bang Theory.*

3. Even within the UK, radio comedy today has only a fraction of the impact that it used to (Chignell, 2009, p16). It is unlikely that there could ever again be a radio comedy with the penetration into the public consciousness that **ITMA** had.[77] Taking a more recent example, **Little Britain**, which became a huge hit even in the US, only really made an impact on public consciousness once it went to television.

4. Radio comedy as an area of study and discussion remains under the radar for the most part and the medium gets nowhere near the academic or critical attention that television or film attracts (Chignell, 2009, p17; Shingler and Wieringa, 1998, Pxii).[78]

   > We don't read or hear about the radio medium very much either – it rarely makes the front pages, rarely arouses the same sort of heated debates over say, violence or sex or sensationalism, that television seems to engender… in society as a whole it is largely ignored. (Hendy, 2000, p3)

5. In-depth books on the concepts and practicalities of radio comedy do not exist, whereas they can be found, for example, for radio drama, such as Lewis's *Radio Drama* (1981) and Crook's *Radio Drama: Theory and Practice* (1999).

2.4.2 Entertainment

Another benefit derived from radio comedy is that of entertainment, one of the triumvirate of the BBC's aims which also includes to inform and educate.

> *I think we gain enormously by having comedy on BBC radio… Part of the BBC's mission is to entertain and comedy is curiously good at that.* (Davie, 2009)

Furthermore, if Radio 4 were an unrelenting diet of news and factual output, it would be indigestible (Hendy, 2007, p216), and comedy allows listeners to tune in with a sense of 'pleasure and anticipation, rather than a sense of duty' (Tony Whitby: ibid, p78).

---

[77] See from p18.

[78] See from p9.

However, while comedy may be a key aspect of the entertainment value for Radio 4 audiences it is recognised that it can be polarising for many listeners: 'They love it or they hate it' (Williams, 2011).[79] There is even a requirement for humour to *aim* to be challenging to some of its listeners. It is a necessity of the network to develop new comedy and, in turn, it is accepted that not every listener will find all of it entertaining: '*Ultimately there needs to be some element of risk to get something new and different*' (Berthoud, 2010). Still, the polarising nature of comedy could be detrimental to the overall perception of the station were too many comedies a literal and metaphorical *turn-off* to some of the listeners. Listening figures are not the only measure of audience approval, aggregate appreciation is also a BBC measure.[80] The aggregate AI for a BBC radio station consists of the total responses across all the programmes. Thus, in theory, the more people who listen to comedy programmes, the more comedy appreciation responses there will be, increasing comedy's influence. Comedy can incite some of the highest and lowest AI scores (see p225), so the mix of comedy programmes could influence the station's reported total.

Despite its importance in the mix of output, comedy is not necessarily seen by listeners as a 'cornerstone' of Radio 4 entertainment. When asked to pinpoint the programmes that defined the Radio 4 schedule (albeit in 1997), no comedies were seen to be key. Comedy was of lower priority than news and other factual programmes (Counterpoint Research, 1997a, p34). Entertainment was seen to be primarily represented by such programmes as *Book at Bedtime* and *Desert Island Discs*.

In autumn 2011, Williams introduced a new comedy slot to the Radio 4 schedule, Sunday nights at 19:15. This replaced *Americana* which had in turn been introduced by Damazer to replace *Go4it*. For the year Q4 2011 to Q3 2012 the comedy programmes actually achieved a lower average audience reach than *Americana:* 525,000 versus 558,000. Even taking into account a drop across the whole network for the same time periods, there was a -2% like for like decline (RAJAR). This behaviour suggests the audience preferred the factual content to comedy.

2.4.3 Audiences

Typically, the spread of television in the 1950s led to a decline in the popularity of radio comedy. In the US the growth in pop music around the same time meant that commercial stations were able to attract bigger audiences with music rather than comedy, music also being preferable as it is cheaper to produce (Wertheim, 1979, p383-384). While Britain too saw declining radio comedy audiences, the BBC, as a public service broadcaster, did not need to ensure that all of its output was fully cost effective and thus radio comedy survived the critical period of the 1960s when commercial radio abandoned traditional styles of comedy. For example, it is likely that had **The Clitheroe Kid** (1958-1972) had been on a commercial station rather than the BBC, it may have been cancelled much earlier than it actually was as a BBC show; over its 14 year run its audience dropped from 10.5 million to 1 million (Foster and Furst, 1996, p206).

However, even since its heyday, comedy has been viewed as a key attraction in pulling in new, younger audiences to Radio 4. Clare Lawson Dick, Radio 4 Controller from 1975-76, found comedy to be a great word of mouth driver to get people talking about the station and, in turn, invite new people to listen (Hendy, 2007, p186). Having seen what could be done in terms of cultural impact by shows such as **Hancock's Half Hour** or **Take It From Here**, she was keen to try to focus on comedy as the genre that could increase listenership. Innovative comedy could, once it found a new audience, enable a broadening of the perception

---

[79] See from p48.

[80] See from p88.

of Radio 4. The aim was to use comedy as a point of entry, such as was consequently found with ***Hitchhiker's***: where listeners wrote in to say that it was this programme that had introduced them to the network (ibid, p193).

> The typical listener to **The Hitchhiker's Guide to the Galaxy** was 31 years old and "more upmarket" than even the typical Radio Three listener. (ibid, p268)

Boyle's 1997 research indicated that to bring in new, younger listeners, new talent and new formats were key (ibid, p392), a view that persists today:

> There're so many programmes I know that younger people would like but are missing. For example, *More or Less*, *The Philosopher's Arms* and select comedy… A lot of people turn on for Radio 4 comedy. (Williams, 2011)

Nowadays the huge audiences of early radio comedy do not exist, but the genre of radio comedy continues developing, mainly on Radio 4, with some shows still garnering linear audiences of over 2.4m in a typical week (RAJAR, Q4 2013).[81]

However, the belief that Radio 4 comedy can bring in unique audiences – 'It's a real turn on time. A lot of people turn on for Radio 4 comedy' (Williams, 2011) – is generally anecdotal and must not go unchallenged. Research indicates that the listeners for comedy actually 'aren't too different to Radio 4 listeners because comedy is such a big part of Radio 4' (BBC Audience Research, 2010, slide 2). Also, the view that comedy can bring in younger people to Radio 4 (see) is not apparent in audience listening figures. In fact *all* of the comedy slots attract older listeners, on average, than the station as a whole, which has an average listener age of 55:

Figure 4 – Summary of Radio 4's comedy slots – mean ages (Q3 2012 – BBC Audiences Portal – RAJAR)

Slot (28′)	Day	Mean age
11:30	Monday	61
	Wednesday	60
	Friday	59
18:30	Tuesday	57
	Wednesday	56
	Thursday	57
18:30	Monday	57
	Friday	57
23:02	Tuesday	61
	Wednesday	62
	Thursday	61
19:15	Sunday	59

However, these mean ages could be hiding significant detail. Might the late night (23:02) programming, created as '*the innovation slot*' (Berthoud, 2010), appeal to a younger audience, which is hidden within the mean?

---

[81]

Figure 5 – Radio 4's comedy slots on a Wednesday – age profiles – percentage mixes (Q3 2012 – BBC Audiences Portal – RAJAR)

Age Groups	Radio 4 total	Wed 11:30	Wed 18:30	Wed 23:02
15-24	5%	3%	5%	1%
25-34	9%	5%	6%	4%
35-44	13%	7%	11%	8%
45-54	18%	15%	21%	16%
55-64	21%	27%	23%	22%
Above 65	34%	43%	35%	48%
mean age	55	57	56	62
Age Groups – varying ways to segment younger audiences	Radio 4 total	Wed 11:30	Wed 18:30	Wed 23:02
15-44	27%	15%	22%	13%
15-34	14%	8%	11%	6%
15-24	5%	3%	5%	1%

The data in figure 5 shows the audience mix for the three comedy slots that TX on a Wednesday (the only day during which three slots are allocated to comedy). It indicates that for each of the measures of a younger audience, the late night slot does not attract a younger skew than the Radio 4 total. It is not even younger than the other two slots earlier in the day. To pilot comedy that is intended for transmission on BBC1, BBC2 or BBC3 television could be perceived as odd or jarring for Radio 4.

> *It's always struck me as an irony that Radio 4's image of its audience being 55 as an average age, also was the channel where you were launching Stewart Lee and Richard Herring way back and more recently Tom Wrigglesworth and Sarah Millican.* (Schlesinger, 2011)

What these figures don't tell us, however, is the ages of those listening through iPlayer, making it a possibility that the total audience *could* be skewed younger. However, because non-linear listening remains relatively low, it would still make little difference to the overall figures.

Lost Audiences

Listening figures are not the only measure of audience approval, the BBC also measures programme appreciation.[82] Comedy can incite some of the highest and lowest AI scores (see p225), and while cult comedies may attract a cult audience, they can also cause others to turn the radio off or over to another station. Pulse verbatim responses can illustrate the strength of feeling created and provocation to stop listening to comedy that is not liked:

> ***The News Quiz***   27-Apr-12   Male 51 (respondent 617760)
> Appreciation score – 1
> 'This show is bad enough already with its forced humour and a "script" for Toksvig that is written by 6 people, but when I heard that an impressionist was going to be on the show, thereby shoehorning more opportunities for crap thinly disguised as "satire" into the show, I turned off immediately'.

> ***Fags, Mags and Bags***   23-Aug-12   Male 60 (respondent 719314)
> Appreciation score – 1
> 'Awful – turned over to Classic FM'.

> ***The Secret World***   25-Sep-12   Male 60 (respondent 719314)
> Appreciation score – 1
> 'Just rubbish – turned to another radio station'.

> ***Andrew Lawrence: How Did We End Up Like This?*** 25-Oct-12 Female 58 (respondent 607011)
> Appreciation score – 1
> 'His voice, as soon as I heard his high squeaky voice I turned it off'.

---

[82] See from p85.

Appreciation score – 2
'Noisy and brash. The sound jarred and I switched off'.

2.4.4 Financial

Radio 4 has a large audience but at a relatively high cost compared to other BBC radio stations, being comparatively 'extraordinarily expensive' (Hendy, 2007, p6). Figures for Radio 4's cost for 2012-13 (BBC, 2013a, p8):

Content	£91.1m (including news output which is managed separately)
Distribution	£9.8m
Infrastructure/Support	£21.2m
Total	£122.1m

To place this in context, the total cost is more than double that of Radio 3 which is a mere £54.3m.[83] Total cost, however, is not a measure of value for money. The total listener hours can be compared to costs in order to calculate the cost per listener hour (RAJAR, Q2 2012-Q1 2013):

Total listener hours 2012-13	Radio 4 – 6445m	Radio 3 – 652m
Cost per listener hour 2012-13	Radio 4 – 1.9 pence	Radio 3 – 8.3 pence

Radio 4 may cost more than twice as much as Radio 3, but it gives a far lower cost per listener hour as it has a much bigger audience.

'Comedy is the area of Radio 4 output which is perceived to have the greatest potential commercial value' (BBC Executive, 2010, p66). For example, **Band Waggon**, **Hitchhiker's** and **Knowing Me, Knowing You with Alan Partridge** have had film versions and there has been a DVD of '**I'm Sorry I Haven't a Clue**: *Live on Stage*'.

Within Radio 4, comedy is a relatively expensive genre (National Audit Office, 2009, p16), commissioned at £11,300 for a 28' programme. Other genres for programmes of the same duration tend to be cheaper, for example, documentaries at £8,300 or arts programmes at £6,200 (Radio 4 Commissioning Guidelines, Spring 2014). The total cost for comedy across the whole of the Radio 4 schedule is around £5m per annum[84] – around 6% of the £91m content total.

Whilst radio comedy is a relatively expensive radio genre, it is cheap compared to television comedy. Radio hourly costs can be as little as one fiftieth of the equivalent television costs (Hendy, 2007, p6, p184): Radio 4's mean hourly rate being £21,000 (National Audit Office, 2009, p16) versus BBC television's £110,000-£600,000 (BBC TV Commissioning, 2012). Because of this, it is sometimes used as a means of piloting television comedy (Smith, 2011a). A television sitcom pilot might demand a set to be built, location filming, costumes and makeup; all these elements would be unnecessary for radio. In recent years, television comedy has been funding five radio comedy pilots per financial year, each broadcast on Radio 4.

As cost efficiencies are sought across the whole of BBC radio on an ongoing basis, comedy output is always considered for reductions. Even the relatively recently introduced 19:15 comedy slot will be subject to a decrease in the origination level, from 26 episodes per year to 16, from 2015/16; this would be a saving of around £110,000 versus 2014/15. The Radio 4 programmers are not currently considering a similar reduction in news output, for example, showing the lower importance of comedy in its programming.

---

[83] Radio 4's cost is however a fraction of that of the BBC's flagship television channel, BBC 1, at £1463.2m (BBC, 2013a, p8).

[84] See .

2.4.5 Development

Comedy is often held to be polarising (see p48) and difficult to predict in terms of audience appreciation, so Radio 4 comedy's relatively low cost compared to television can be seen as an advantage when managing risk:

> *It's a unique resource that the BBC has for developing comedy.* (Mitchell, 2010)

> *It's very rare, a unique thing having the BBC Radio Light Entertainment department* [*sic*]. *There's nowhere else in the world it exists. I don't think you can overestimate its contribution and value to comedy in this country. It's quite remarkable.* (Lock, 2011)

In the US, for example, there is no such culture of contemporary radio comedy providing an intermediary step between live performance and broadcasting:

> *In the US, comedians tend to develop through stand-up but* [it] *is not necessarily the best pathway. It's convenient in the sense that it allows the networks, the television executives, to retain power. If you come up through the stand-up route, you never have to deal with production. You never have to learn how to produce either in terms of creating a scripted half hour or the literal production realities that exist. You go from standing on a stage by yourself with a microphone to being on a set somewhere where you have to hit a mark and you don't even necessarily know what a mark is. You don't know the first thing about production if you come up that way.* (Black, 2011)

Comedians still value the role that BBC radio comedy plays in their career development. In a survey of 105 comedians during 2012, 70% still thought that BBC radio was an important element of their career development.[85] Many recent winners of the Edinburgh Comedy award will make their way onto Radio 4: Bridget Christie, Humphrey Ker, Adam Riches, Tim Key and Jonny Sweet have all won best show or best newcomer in the past 5 years and all have been heard on the network.

Later in a comic's career, Radio 4 continues to allow development opportunities. Well-established comics can be lured to radio (despite the relatively low pay) through the convenience of not having to learn lines, rehearse endlessly, dress up, or be covered in makeup (Fry, 2010a, p332; Mayhew-Archer, 2011). An example of this from 2013 can be seen in the one-off *Sketchorama: Absolutely Special*, when the *Absolutely* team were enticed to Radio 4. For their trouble, the programme went on to win the BBC Radio Best Scripted Comedy (Studio Audience) award in January 2014.

The relationship between BBC radio comedy and BBC television comedy is often seen as a one-way route with programmes starting on radio and moving to their more lucrative sister medium, television:

> An escalator from Radio 4 to BBC2. (Elmes, 2007, p230)

> *Radio comedy is absolutely the seedbed for most other things.* (Reynolds, 2011)

Others see the connection to be more symbiotic, for good or ill. For example, *That Mitchell and Webb Sound* ran on Radio 4 from 2003 to 2013, overlapping with the sister television series *That Mitchell and Webb Look* during 2006 to 2009. Some of the early radio content was reused for the television but only sketches that were 'cherry-picked' for their suitability for the visual medium (Mitchell, 2010).

> *I always felt that it was a two-way thing and that that ideas should come back. Certainly talent should come back and try new ideas. It was a two way street.* (Schlesinger, 2011)

> Sibling media… intermittently been tempted by incest. (Lawson, 2012)

The opportunity for radio comedy to move to television has existed since television began to be taken seriously as a medium. When the coronation, in 1953, drove sales of television sets it also started the trend of

---

[85] Details of the survey can be found in the appendix – p285.

movement from the audio to the visual, including *Hancock's Half Hour* in 1956 (Street, 2002, p98; Barnard, 2000, p113). Many types of television comedies have had a birthplace on radio, but Schlesinger (2011) claims he found that sketch and panel shows can make the transition more readily than sitcoms. Supporting this is the 2003[86] voting result for the UK's favourite sitcoms; none of the top 10 had started on radio. This is the case down to number 18, where Red Dwarf is seen (the characters first appeared in *Son of Cliché*), and next at number 30 with *Hancock's Half Hour*.

Examples of radio shows that have moved to television:[87]

Title	Radio series dates	Television series dates
Hancock's Half Hour	1954-59	1956-60
Just a Minute	1967-	various
The News Quiz (as *Have I Got News For You*)	1977-	1990-
The Hitchhiker's Guide to the Galaxy	1978-2005	1981
Radio Active (as *KYTV*)	1980-88	1989-93
After Henry (went to ITV)	1985-89	1988-92
Up The Garden Path	1987-93	1990-93
Whose Line Is It Anyway?	1988	1988-98
The Mary Whitehouse Experience (Radio 1)	1989-90	1990-92
In and Out of the Kitchen	2011-	2015
Sparkhill Sound (*Citizen Khan*)	2011	2012-
Nurse	2014	2015-

One issue may be that some consider radio comedy as merely a stepping-stone to television rather than an entity in its own right. One radio comedy producer seemed to be dismissing the genre as unimportant when he advised his colleague: '*Let's not worry about it, we're just making shaped air*' (to Canny, 2011). The view of its merely intermediary status – '*TV is the more powerful medium* [vs radio] *in the country and that's how you get your comedy to a wider audience*' (Mitchell, 2010) – seems confirmed by its financial reward; many talented radio comics are enticed to television by its greater payments for content. This means that programmes popular with the Radio 4 audience, when they become successful and move to television, would no longer be heard on the network, contrary to the wishes of the Radio 4 programmers.

A significant new risk for radio comedy development is the increasing ease with which comics can now *self-broadcast* using non-traditional broadcasting means. Clarke (2011, p21) notes that the relative 'ease of making short, tailor-made online content' can appeal to new and established comics alike. The greatest talents may see Radio 4 comedy as something they can bypass on their way to stardom. '*I think radio has to be aware that it is in danger of becoming less relevant both to performers and listeners*' (Herring, 2010), meaning that increasingly it could be argued that 'Radio 4's role is unclear' (Armstrong, 2007). Indeed, while Mullone (2013) writes of how Radio 4 may currently still be 'a crucial rung on the ascent of any professional joker', he goes on point out that 'the formats of BBC radio comedy seem stuck in the past', claiming that the allowable formats and ideas are too prescriptive to nurture creative talent. To bypass BBC radio, podcasts allow comics complete creative freedom unfettered by Radio 4's comparatively dictatorial requirements. Meanwhile, in television land, comics are even attempting to fund pilots of filmed content through *Kickstarter* campaigns, [88] allowing them creative freedom of normally highly prescriptive platforms, again excluding the need for piloting new formats through radio.

---

[86] http://www.bbc.co.uk/sitcom/winner.shtml [accessed 04/04/14].

[87] Further examples included in the list on p30.

[88] See for example: https://www.kickstarter.com/projects/690037994/the-buzz-a-tv-comedy-show-with-mark-dolan [accessed 20/01/2015].

The 23:02 slot was introduced in the mid-nineties '*primarily to innovate*'; 11:30 and 18:30 both were seen as too mainstream to bear too much innovation (Berthoud, 2010). Hendy (2006, p169) claims that, due to its smaller audience, the evening allows for more 'specialist or intricate' programmes and supports shows that can appeal to a 'minority audience', allowing 'programme-makers and schedulers greater freedom.' However, there is no evidence of listeners wanting to listen to challenging comedy near bedtime. The 1997 research indicated that listeners to late night radio want something undemanding (Counterpoint Research, 1997c, p6):

> Ideal Radio:
>
>> Able to concentrate on continuous narrative.
>> Still intelligent, but not demanding.
>> Pleasant drawing-out of last energy.

If the slot is being used to test reactions to new comedy, the listeners at this time might not be at their most receptive to innovation. Arguably, if Radio 4 is aiming to trial new ideas with a small (and low risk) audience, the Sunday evening 19:15 slot is smaller than late night. However, if the primary aim is actually to hit a younger audience, 18:30 would be more effective. Feedback from research undertaken by The Trust indicated that while there was indeed a desire for developing innovative, risk-taking comedy shows (BBC Trust, 2010, p46), the ideal platform for this type of programming would in fact be Radio 4 Extra (at the time known as BBC7):

> There was also some feeling that more could be done to build on associations with Radio 7 as a platform for 'new' comedy, which was a particularly interesting potential development for younger audiences. (BBC Trust, 2010, p61).

Unlike other BBC radio departments, Radio Comedy comes within the corporate structure of BBC 'Vision' (television), rather than 'Radio and Music', ensuring necessary financial support to develop pilots. Consequently, the link from BBC radio to BBC television is stronger in comedy than in any other genre. However, the risk with this approach is that the objective is split between creating a programme suitable for the Radio 4 audience versus creating a format that is a potential television show. What works for radio won't necessarily be suitable for television in every instance (Caulfield, 2009, p83).

> *There's always a bit of a tension within radio about how much the comedy should be tailored just to the radio audience and how much they should use it as essentially an unofficial testing ground for television, and there will always be that tension. It's in the interests of the BBC and Britain in general that some radio comedy time is given over to some experimental programmes.* (Mitchell, 2010)

In addition to piloting on radio, television can simply adapt formats that have been seen to be successful on radio, the thinking being that a proven format programme, albeit from another medium, would have a lower risk of failure than an idea with no track record. McWhinnie (1959, p175-176) pointed out, however, that although benefitting from an established demand, the programme to be adapted must also be suitable for that new medium. For example, **Hitchhiker's** was a cult hit on Radio 4, but its television incarnation was arguably not as successful:

> If he wanted a scene with a million singing robots or to crash a starship into a sun, he could do so… Such freedom does not apply to a visual medium. So Douglas [Adams] couldn't just adapt the radio scripts; he had to re-imagine the whole adventure visually. (Webb, 2003, p206)

> Ruined by the cheap-looking sets, ridiculous 'aliens' and dubious prosthetic heads. (Hawes, 2004, p86)

Radio shows such as **Hitchhiker's** may not work on television as it is hard to convert epic audio worlds to visual media. Likewise, very small, intimate moments on radio might be boring to look at on television:

It can be murder to convert a radio sketch to television. You could have a lot of great verbal jokes but it doesn't work just to point a camera at them being said. You end up trying to crowbar in visual moments. (Mitchell, 2010)

For example, Logan (2009) argued that the decision to transfer **Down the Line** to television was done for reasons other than its suitability for the medium. As a spoof radio show it was not an obvious move to a visual medium. This commentary was prophetic as the television version, *Bellamy's People*, did not prove successful and was not recommissioned. Likewise, Lawson (2012) lamented the television adaptation of **JAM**, arguing that it is pointless to create a visual version of a show that had nothing to add visually: '**Just a Minute** is perfect radio because it is entirely verbal, awarding auditory attention.'

2.5 Comedy and Polarised Audience Responses

BBC Radio 4 comedy's job is to amuse the listener (Damazer, 2010), and in doing this it must appeal to their sense of humour. However, not everyone finds the same things funny: 'a sense of humour is as unique as a fingerprint' (Helitzer, 2005, p33).

> Comedy is more likely to divide listeners than any other genre: what is witty for some will fall flat for others - and for others still, may be offensive. (BBC Executive, 2010, p40)

While any genre could be subject to differing tastes, comedy seems to provoke particularly extreme reactions:

> Charlie Higson: 'If comedy doesn't make people laugh, it seems to make them angry.' (Bradbury and McGrath, 1998, p154)

> Paul Whitehouse: 'Somebody made the point recently that they don't like The Archers, but they don't ring up saying, "Can you take this off the air?" They simply don't listen for that half an hour. I'm one of those – I switch the radio off as soon as I hear that theme music, and I urge anyone who doesn't like **Down the Line** to do the same. Read a book or go to bed.' (Moss, 2006)

> Jon Naismith: '*Obviously, comedy is a polarising genre.'* (Naismith, 2010)

> David Mitchell: '*Every joke's going to offend someone.'* (Mitchell 2010)

> Gwyneth Williams: '*I think that of Radio 4 comedy, people are very critical. They love it or they hate it.*' (Williams, 2011)

> Jonathan Lynn: 'People get inexplicably angry and abusive if they expect to laugh at your show and then they don't find it funny.' (Lynn, 2011, p45)

> Caroline Raphael: '*When you piss people off, you really piss them off. They are really vociferous if they don't like* [Radio 4 comedy]. *And remember, if they don't like it they're going to not like it for six weeks on the trot. It invariably will offend somebody.*' (Raphael, 2012a)

Whilst these views are arguably unsubstantiated, data relating to 18:30 Tuesday comedies from a sample pilot study[89] provides clear evidence that Radio 4 comedy *can* elicit extreme responses, with the lowest possible score of 1 being the third most popular selection of the pilot examples.

Must a public service broadcaster always strive to avoid annoying people? Should comedy creators aim to keep everyone happy or is it acceptable for some of the listeners to sometimes hate the output? It depends on the type of comedy. Programmes that are 'universally adored', such as **Clue** and **The News Quiz** (Bolton, 2011), are broadcast when the biggest audience is available, whereas comedies that are higher risk, and more likely to have the potential to provoke dislike, can be found in the late night slot where there is a smaller audience and a heritage of edgier content (Raphael, 2012a). As discussed (see from p45), part of the BBC's role is in broadcasting programmes that take risks and may cause division. Shows such as **I'm Sorry, I'll Read That Again** were said to have attracted a 'cult audience' (Briggs, 1995, p805).

---

[89] See p157.

Mat Coward: 'When I was a kid, every cool person I knew loved ***ISIRTA*** and every prig hated it. (Grannies, generally, thought it rather noisy).' (Coward, 2003, p71)

David Hatch: told his colleagues in a review board meeting in 1980, 'Comedy programmes tended to divide the Radio 4 audience into old and young, and the old audience always wrote in to complain about the programmes aimed at the young.' (Hendy, 2007, p189)

Generation divide is a common feature of comedy polarisation. Younger people can be particularly attracted to programmes that their elders can find a 'noisy, punky and youthy mixture of the aggressive and the silly' (Coward, 2003, p76). The first radio comedy show to do this was ***The Goon Show*** (Bradbury and McGrath, 1998, p13), deemed 'incomprehensible and irritating' (Bolton, 2011) to those excluded from the cult. A later example dividing the generations was that of ***The Burkiss Way***, which was supported by Radio 4 in its early days as it attracted much sought-after younger listeners (Hendy, 2007, p190), despite the fact that the majority thought it 'tediously juvenile' (BBC Audience Research Department, 1978, p32).

Comedy can be *particularly* satisfying for those in on the joke in the knowledge that others simply don't get it. Divisive responses can actually be the aim for some:

Nick Doody: '*Everyone liking your stuff, I think it's George Carlin who said that's the definition of mediocrity.*' (Doody, 2011)

Steve Bennett: '*Because they're comedians they want a reaction, that's their job. If someone really hates them, I think there's a strange sort of satisfaction in that.*' (Bennett, 2011)

Jane Berthoud: '*Some of our output I would expect to be polarising. It kind of has to be*'.

One current example of a Radio 4 comedy that provokes extreme responses is ***Count Arthur***. Its recommission drew discussion on Radio 4's *Feedback* programme (2012a and 2012b), with contributors illustrating the most divergent positions, for example:

'I really think is the least funny thing I've ever heard.'

'I find it boring and silly and absolutely not amusing.'

'It has such a poor script that I couldn't believe it was being broadcast.'

…contrasting with:

'It just made me fall-over. I had to actually sit down. [laughs] I'm just laughing thinking about it.'

'Count Arthur is one of the funniest, if not the funniest comedian I have ever heard or seen.'

Polarised responses are clearly a significant element in comedy appreciation with evidence that comedy incites extreme responses found through BBC's own (unpublished) research:

We looked at, for example, programmes and genres that delivered the greatest range and standard deviations… and felt that, for example, Radio 4 comedy was attracting the biggest range of scores. So that it had the highest and the lowest. (North, 2010)

One thing we have looked at with AIs is the distribution of scores… I think that generally comedy is a bit more divisive than other genres. (Collins, 2010a)

Comedy Programming …somewhat hit and miss in terms of popularity. (BBC Trust, 2010, p4, p44)

However, if people tend only to consume media that they enjoy, it seems unlikely that people might choose to listen to a radio comedy that they dislike so why would there ever be negative appreciation responses? Menneer (2014) suggests that the nature of the genre may make it difficult for a listener to predict whether they might like a programme or not. This means that some people listen to it in hope of enjoyment but are disappointed:

I would guess that the principal reason for some lowish scores is that their audiences are less able to predict whether a particular programme will appeal. To a greater extent they can be caught by surprise.

2.6 Radio 4 Comedy Commissioning

2.6.1 The Numbers

There are a finite number of hours in the Radio 4 schedule and only a small proportion of the output is dedicated to radio comedy. The exact output is bound by the following criteria:

Statements of Programme Policy (SoPPs)

- Part of Radio 4's service licence is that it delivers at least 180 hours of original comedy per year, around 3.5 hours per week. (This is actually exceeded each year).

Output Guarantees (OGs)

- The output is divided up very specifically across the slots and between particular suppliers. For comedy, output is allocated in 28' portions.[90]
- The difference between the OGs and the total is the 'Window of Creative Competition' or WoCC. This is business that is not allocated to any specific supplier and is therefore open to competitive bids from both BBC departments and independent suppliers. This is expected to total 10% of the eligible hours[91] across all of Radio 4 but is not specified at genre level.

Indie Quota

- Independent suppliers are guaranteed to get at least 10% of Radio 4's eligible hours. Some of this comes from comedy output.

The OGs are distributed as follows:

Figure 6 – Radio 4 Comedy Output Guarantees April to March (2013/14). Number of 28' equivalents slots

Slot (28' equivalent)	BBC Radio Comedy Department	BBC Radio Drama Department	Window Of Creative Content – WOCC	Total 28' Origination Slots
Mon, Wed, Fri: 11:30	38	24	60	122
Tues-Thurs: 18:30	82		34	116
Mon and Fri, 18:30	80		24	104
Tues-Thurs: 23:00	38	12	31	81
Sun 19:15			26	26
TOTAL	238	36	175	449

Generally, each of the slots costs £11,300 while the Mon and Fri 18:30 programmes are a little higher. This means that Radio 4 pays over £5 million per annum for comedy programmes.

2.6.2 The Process

Comic performers, writers and producers work together to come up with ideas for radio comedy shows, but there are many factors involved in whether a concept makes it to broadcast. In essence, producers offer ideas to the network 'gatekeepers' (McLeish, 1999, p279) who attempt to balance the output for the station.

In the earliest days, radio comedy was selected based on what was popular in the music halls at the time and broadcast unedited to the audience (Crisell, 2002a, p39). Radio comedy output was controlled by

---

[90] Exact duration = 28' including announcements or 2x14' including announcements.

[91] Eligible hours excludes World Service, news and presentation output.

the incumbent Directors of Variety on a seemingly casual basis: ad hoc commissions could take place over casual drinks (Took, 1981, p22-23). After WWII, Heads of Departments and producers would meet with network Controllers on a weekly basis and discuss programme ideas (Briggs, 1979, p715; Briggs, 1995, p388-389). This basic idea of pitching to the Controller was maintained up to the late 1990s and Schlesinger (2011) remembers that as a producer that '*we used to troop up to the Controller once every month and talk about ideas and he'd either say yes or no.*'

In 1996, in an attempt to improve transparency of costs and efficiency of processes, Commissioning Editors were introduced as a filter between the departments and the Controller (Hendy, 2007, p289). Initially the Commissioning Editors were allocated dayparts to manage, this being see as the most audience-facing approach (McLeish, 1999, p286). The process was changed 18 months later so that they were in charge of genres instead, as it was '*... so much easier from the outside. The talent know where to go, who to talk to. We're not playing one commissioning person off against another*' (Raphael, 2012a). The process of suppliers pitching to Commissioning Editors who then, in turn, propose a range of genre-specific programmes to the Controller, has been in place since this time. Offers were originally managed as hard-copy documents but nowadays proposals are managed through a web-based system called Proteus.[92]

For Radio 4 comedy the decision-making primarily takes place in the form of a commissioning round twice a year. The Commissioning Editor will publish guidelines for the suppliers that indicates both what kind of programmes are being sought and an indication of the amount of business available. Each of these aspects will also reflect any overarching, relevant network strategy.[93] Initially suppliers are expected to submit a number of short pitches, known as 'pre-offers', which are whittled down to a smaller number by the Commissioning Editor. Those accepted for the next stage will then get the opportunity to be resubmitted as fully-worked proposals. These 'full-offers' are in then evaluated and a final list drawn-up for agreement with the Radio 4 controller.

It takes four to five months from the point when the Commissioning Editor publishes the guidelines to the results being communicated to the suppliers. Exceptionally, suppliers can offer ideas outside of these rounds which increases the speed of the process.

> *The doors are always open, particularly after Edinburgh or if it's just a brilliant idea or talent is available. We have always commissioned between rounds:* **Down the Line** *is a classic example. I don't think it was ever put in for a round. It was just an idea that was pitched, and there are others like that.* (Raphael, 2011)

Programmes are generally commissioned around twelve to eighteen months ahead of TX but can be further in advance in order to secure talent or allow for development, or closer to TX to allow reactivity. Competition is fierce in the commissioning rounds. For example, in spring 2013 only 46 programmes, out of 221 ideas pitched for the comedy slots, were commissioned, or shortlisted for later consideration – i.e., 79% of initial proposals were rejected.

---

[92] https://ext-proteus.external.bbc.co.uk/proteus-web/login.action [accessed 04/04/14].

[93] More information on the round, deadlines and the guidelines can be found on the Radio 4 commissioning page – http://www.bbc.co.uk/commissioning/radio/what-we-want/radio-4.shtml [accessed 04/04/14].

Figure 7 – Radio 4 Commissioning Process Summary[94]

Figure 7 – Radio 4 Commissioning Process Summary
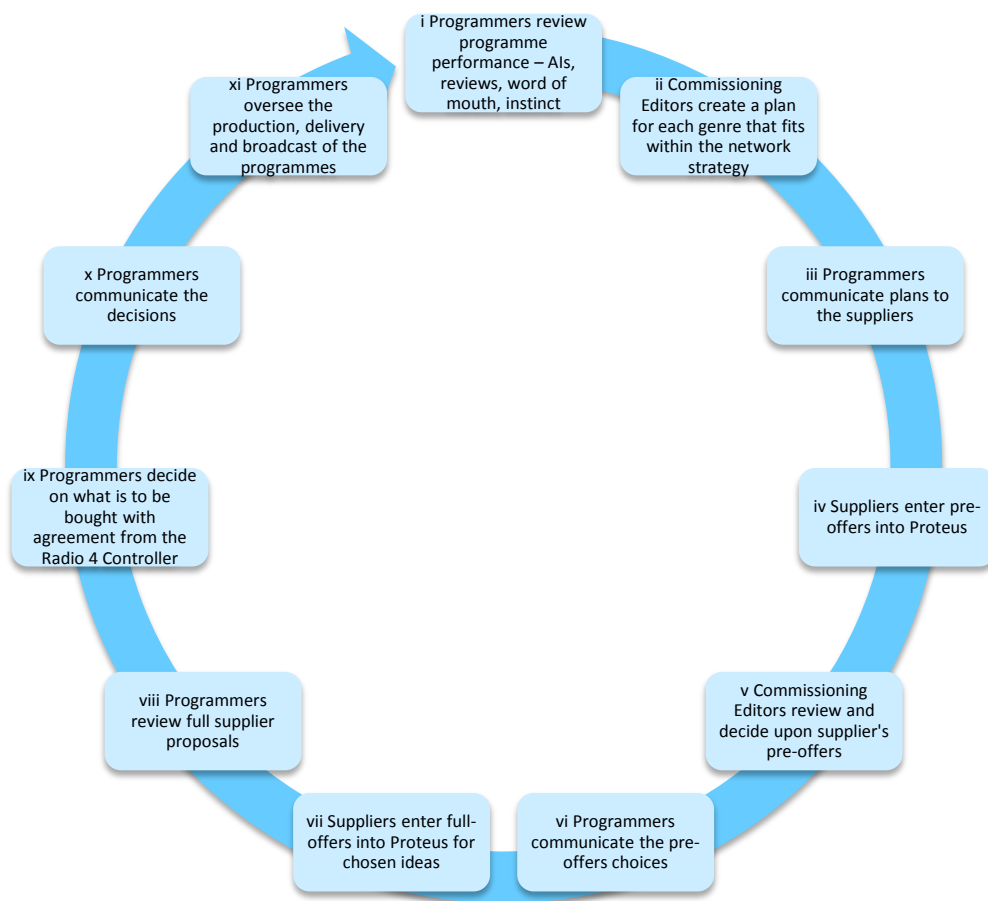
2.6.3 Limitations of Radio 4 Comedy Commissioning

i The Process

Academic studies of the BBC have found that many aspects of the commissioning process are a product of the organisation's culture and history - 'simply because this is the way that things "are done"' (Hendy, 2007, p5). Views on Radio 4 comedy commissioning appear to depend on one's relationship to the process. Internally the Heads of Radio Comedy find the commissioning process for Radio 4 very clear (Schlesinger, 2011; Berthoud, 2010), although Mayhew-Archer (2011), who stood in as the Acting Radio 4 Comedy Commissioning Editor during 2010/11, found the process '*painfully slow... overly complicated and time consuming*'. Independent producers appear to find it complex and inflexible - Jon Thoday, MD at Avalon talent agency, claims that the twice-yearly commissioning rounds are not suitable for radio comedy as they preclude reactivity, and that without speed to market radio loses its advantage over television, risking that talent will bypass the medium altogether (Armstrong, 2007).

In the commissioning round initial pitches are often done via the pre-offers procedure[95] whereby an idea must be summarised in 200 words ahead of a full pitch. There is a feeling from some within Radio 4's supply base that this is insufficient to communicate comedic ideas effectively:

> *The fucking nonsense of pre-offers. You have to write a sentence about it. They'll* [Radio 4 Commissioning] *look at that sentence and go 'no, I don't want that show' so they can go home at five o'clock. Isn't that offensive? Fuck them!* (Anonymous BBC Radio Comedy Producer, 2010)

---

94 Proteus is part of BBC Radio's programme proposal management system.
95 Point iv in figure 7.

As Silvey (1974, p101) puts it:

> How could programmes of kinds which people had never yet heard be effectively described? Would anyone have thought **The Goon Show** attractive if all they know about it was a written specification? It takes a lively imagination to conjure up the taste of a new dish from reading the recipe.

Raphael, Radio 4's Commissioning Editor for Comedy, explains that the purpose of pre-offers is just initial *'checks and balances... clash checking for subjects'* (Raphael, 2012a).[96] She is aware that there is a risk that a good idea can unnecessarily be rejected at this stage. However, she has found that, when producers challenge a rejection it is often because they have written a proposal that does not accurately represent the idea. If the idea is good, Raphael is happy to revise decisions should the producer be able to *'fight their corner'* (ibid).

Mayhew-Archer (2011) argues that in an ideal world the creative producer or talent, the person who has the passion about the project, would pitch their idea directly to the Controller.[97] Schlesinger (2011) agrees that indeed, *'you can get a wonderful sense of someone's enthusiasm, tone, vision and imagination from meeting them and talking about an idea'.*

> *One of the barriers to good programme making is the number of stages between the people who might come up with the idea and the person who can actually commission it. If you devolve more actual programme decision making power to senior producers then it's much easier for the person with the idea to meet the person who can green-light it.* (Mitchell, 2010)

However, Schlesinger (2011) points out that while someone's pitch can be very convincing if they can produce a cogent, impassioned argument: *'ultimately, until you see a script you can't really make a judgement as to whether something's good or not, whether they can write or not'.*

Dissatisfaction with the process is, of course, observable mainly from those on the receiving end of rejection.[98] Lock (2011) found rejection communication *'much worse'* in BBC television versus BBC radio and claims this could be easily improved. Mayhew-Archer (2011) believes that as the rejection message tends to come from the producer, rather than the person who has made the decision, the exact reason tends to be obscured along the non-official communication routes, often the resulting feedback being the not very useful: *'not quite right for the slot'.* Schlesinger (2011) explains that, *soft* feedback is often used as no one wants to hurt anyone's feelings, so explanations such as *'it's clashing or the slot isn't available or we've run out of money'* are too often routinely (and unhelpfully) employed. Schlesinger, when Head of Radio Light Entertainment, attempted to create an open dialogue:

> *I always believed in straight talking and honesty and getting people in the room. To get the writer and the producer in the room with the Controller, whether it's good news or bad news, is the best way.* (ibid)

In July 2010, Goddard produced a study of the BBC's commissioning of independent radio productions. He summarised the feelings toward the commissioning process, not confined to Radio 4, but it being the primary focus. He asserted that there was a general feeling of dissatisfaction, one not limited to suppliers of greater or lesser resources nor, indeed, those less successful in getting commissions (Goddard, 2010, p63-64). Naismith (2010) claims that, as an indie supplier involved in this study, it was the first time ever that his thoughts on the process were sought. A culture of holding back on criticism of the process appears customary: a very senior member of the Radio Comedy Department would only admit anonymously that they were at all critical of the process (Anonymous BBC Producer, 2011):

---

[96] Point v in figure 7.

[97] Bypassing points iv to viii in figure 7.

[98] Occurring during point x in figure 7.

I think at times it felt a bit too rigid. I think that judgements could often be made on criteria that felt convenient to apply rather than necessarily helped really ascertain what would make a show tick. I do understand there's a lot of people submitting a lot of ideas…

Criticisms are seldom aired due to fear of being seen as rocking the boat. Another BBC radio comedy producer highlighted a satirical piece on YouTube as a reaction to the process. However, this was done anonymously and only shared among those of a common viewpoint. (See for a transcript).

The Radio 4 commissioning process has a number of restrictions that may impact upon the 'correct' decision being made. For example, regulatory requirements such as the SoPPs, OGs, and indie minimums,[99] mean that ideas might not be commissionable merely due to from *where* the idea is coming from and *when* it is pitched rather than the idea itself. Financial restrictions also apply as there is generally a fixed budget of £11.300 per half hour programme. While that can be increased slightly in exceptional circumstances any extra money would be limited. Timing is also an issue as any idea would be considered in light of what has been already bought and scheduled, whether there is any available money, and the comedic zeitgeist. Furthermore, although there are no specific quotas in Radio 4 comedy, the Commissioning Editor and the programme makers may have personal targets in terms of ensuring what they regard as a reasonable mix of gender, ethnicity and regionality across its contributors. For example, Raphael (2012a) has admitted to favouring the commissioning of female talent to counteract the dominance of the '*boy bands of sketch comedy*', and Berthoud (2010) has identified a lack of black and Asian contributors on regular long-running series, and has found it a slow process to get more ethnically diverse comics on air.

ii Objectivity in Decision Making

'The Commissioning Editor's job is primarily to commission and review programmes' (BBC Radio 4 Commissioning Guidelines Spring 2012, p9), Raphael specifies: 'My job first and foremost is to get comedy that will be enjoyed by the Radio 4 audience' (Dowell, 2008). Part of that process is creating a range of comedy to service the wide range of tastes of that audience. '*That's why there's a commissioner, so they've got an overview of what is in the mix, and they can make choices*' (Schlesinger, 2011). However, can one person make objective decisions about such a subjective area?

Probably not, but I do commission things I don't like. There are things I'd have taken off years ago, I just don't get them, but I know that they're right for the Radio 4 audience. I know enough about the audience and I do look at the AIs and look at the feedback we get coming in. After a series you can tell whether something's hit the mark or not. (Raphael, 2012a)

Thus, Raphael states that her personal tastes are of lesser importance than any information taken from how the audience reacts. She continues:

*I'd have taken off **Clue** a long time ago. Gyles Brandreth, I'd shoot. Quote, Unquote would have just come off although it barely counts as a comedy. I think The Write Stuff is tired and we'll run out of writers again. I'm not that obsessed with **When the Dog Dies** but we've got Ronnie Corbett and he means a lot, like Stanley Baxter and June Whitfield. It's nice to showcase these people. I can tell whether they're well written or well produced enough, they just don't make me laugh.* (ibid)

Audience feedback and reviews for Radio 4 programmes are only available for shows *after* they have been broadcast. For new ideas the Commissioning Editor has to identify which will be successful merely from a synopsis, treatment, script, occasionally a pilot, or from a live performance. Mayhew-Archer (2011), the only

---

[99] See from

person other than Raphael to work as Radio 4 Comedy Commissioning Editor,[100] found the job of decision-making very stressful with a considerable volume of work, including feeling duty-bound to attend recordings in the evenings. He found that having sole responsibility for selecting what was to be commissioned a very 'lonely' role: '[I feel] *perpetually tense… mostly I've found it absolutely terrifying'*. Mayhew-Archer (2011) explained that, whereas the Radio Comedy Department PDG (Programme Development Group) enabled the discussion of decisions, as Commissioning Editor he did not have a similar luxury. '*Try as one might… I defy anyone to be completely objective'*.

- Knowing What's Going to Work

William Goldman famously claimed of the movie industry that 'nobody knows anything' (1985, p39). He wrote that it is near impossible to predict what audiences will like and cites an experienced studio executive who said of the films to which he had given the go-ahead: 'If I had said yes to all the projects I turned down, and no to all the ones I took, it would have worked out about the same' (ibid, p41). Deciding what comedy to commission, deciding what is funny, may be even more difficult than other genres. Given the view that perhaps no one really knows what is going to be funny, even those who might be the most experienced, expert opinion can be wrong. Many who might be expected to *know* what is likely to be funny admit to uncertainty. John Watt, BBC Radio's Director of Variety 1937–45 throughout the wartime period, wrote on this matter:

> You know there just isn't a foolproof formula for churning out big radio successes. I know. I've used the same one several times; once it's a winner, the other times a flop… Every year we put on at least twenty different series of shows. Out of that we're really lucky if we land one big like **Band Waggon** or **Garrison Theatre**. You can never tell with the public. (BBC Year Book, 1941, p72)

Michael Standing, BBC Radio's Director of Variety (1945-53), was praised for backing great shows (Took, 1981), but he admitted that he backed poor ones too, however no one remembered those ones:

> I complimented [Standing] on his bravery and foresight in promoting such programmes as **Take It From Here** and **The Goon Show**, and backing them in their early days before they were established and, indeed, when they were nearly taken off the air. "Ah," he said, "but I also showed the same 'foresight' backing other shows that never made it, and in fact wasted a lot of people's time and a lot of the Corporation's money." But fortunately successes are remembered and failures are quickly forgotten. (ibid, p3)

One of Radio 4's most successful comedies, **Hitchhiker's**, was not considered an obvious "winner" by Con Mahoney, a highly experienced, 'old-school' Head of Radio Light Entertainment during the 1970s, (Hendy, 2007, p76). When Mahoney heard the pilot he had to ask the producers, Simon Brett and David Hatch, whether it was funny or not; only when they had both reassured him that it was did he feel happy to commission the series (Webb, 2003, p108; Hendy, 2007, p191).

Jon Pidgeon (2008), another Head of Radio Light Entertainment, explained how the Controller of Radio 4 took a lot of convincing about the idea for a show that went on to become a success:

> Simon Nicholls also piloted Dave Gorman's **Genius**, prompting a meeting where, while I ground my teeth and contemplated last straws and camel's backs, Radio 4 Controller Mark Damazer appraised it as if we were relaunching *Tomorrow's World*. Unconvinced by the sound of an audience rolling in the aisles, the Spurs supporter earnestly declared football featuring three teams on a triangular pitch to be unworkable: unarguably so, like every single one of **Genius**'s loopy ideas, but funny, which was our aim. A second pilot was ordered, Damazer got the joke, and now **Genius** too is destined for TV.

---

[100] Prior to Raphael, the Radio Light Entertainment Department (as it was then known) would pitch directly to the Controller of Radio 4.

Although the Commissioning Editor will be the main filter for which comedies are broadcast on Radio 4, their choice is not really made entirely alone. Usually they can only select from programme ideas on offer: ideas that have already been through a process of selection. Ideas originating from the Radio Comedy Department, for example, have already been discussed though the PDG meetings:

> *Refinement through PDG, essentially we build-up a slate then we look across it and see what should be offered. Conversations with Caroline* [Raphael] *about what she's interested in and what she's not. Looking at the guidelines in terms of what the priorities are that round and what are not and then trying not to offer too much so that we're not putting too much effort into too big a slate, and not offering too much choice. Choice is crippling and in the old days commissioners were getting four projects for every one. It's just a nonsense.* (Canny, 2011)

Thus, the Radio Comedy Department can manipulate the final mix offered to the Commissioning Editor. For example,) the Head of Radio Comedy was of the opinion that there shouldn't be too much sitcom at 18:30 (Tues to Thurs). She could influence that by minimising the number of sitcoms offered in that slot from her department (which makes up over 70% of that scheduling slot):

> *I'm not a commissioner but what I can do is offer or not offer. So, if you don't offer too much you can't commission too much.* (Berthoud, 2010)

A Commissioning Editor does not have final sign-off on what's bought for a network, that responsibility belongs to the Controller. When agreeing programmes, Controllers, as the network gatekeepers, have a specific set of responsibilities that they must consider over and above those of the Commissioning Editors (McLeish, 1999, p279), however, for the most part the Controller will defer to the greater, in-depth knowledge of the Commissioning Editor. On occasion however, Controllers have to be strongly convinced of programme ideas; they must be prepared to put aside their personal tastes in favour of what might be best liked by the Radio 4 audience.

There is an argument that decisions of this nature - judging what the Radio 4 audience might potentially like – cannot be objective and, without the benefit of audience research, will always come down to instinct. Thus McQuail (1997, p65) claims that decision makers cannot solely rely on historic audience data to make decisions, they need to anticipate and lead audience tastes and interests and, through intangible elements such as skill, intuition and luck, make the right decisions. They need 'a deeper knowledge of what makes audiences tick' (ibid). If decisions on programming always come down to a judgement call, then even the best commissioning processes may not always result in the best outcomes:[101]

> The roll call of the rejected contains names as luminous as Paul Merton, Tommy Tiernan, and the Flight of the Conchords. No commissioning system is flawless, everyone has blind spots, but what should have been a conduit to creativity was all too often a cul-de-sac, the default response a feebly justified negative, and there was hardly a series I cared about that I didn't have to bat for. (Pidgeon, 2008)

- Production

Even if a decision maker could be completely objective about an idea, there is no guarantee that the idea could be fully realised exactly as envisioned. Many aspects of the development and production processes can impact upon the final programme. The creative team behind a comedy idea will generally want as little interference as possible between their vision and the final result.[102] For Radio 4 comedy, after the point of commission, the talent will work mainly just with the producer, fewer people than they would have to deal with than in television (Mayhew-Archer, 2011). This following section briefly describes the role of the radio

---

[101] Kahneman's *Thinking Fast and Slow* (2010) covers the area of decision making extensively and explains why experts are not always right.

[102] With self-broadcasting that might be possible, albeit within the confines of far lower production budgets.

comedy producer in steering an idea between conception and delivery, and illustrates a number of demands upon them, some of which are conflicting:

Realising the original vision:

- Lock (2011) considers it the producer's job to '*get the closest to what that particular talent wants*'.

- '*My instinct increasingly is that if you've got someone talented, go with it. By and large, most hits are by and large the product of one person's very clear vision.*' (Mayhew-Archer, 2011)

- '*With shows that don't work, the problem can often be found in too many voices influencing the shape and the tone of the series. Lack of clarity of vision. The best shows come from a very small group of people who've got a clear vision of how the show should be executed.*' (Schlesinger, 2011)

Making programmes the producer believes in:

- 'How can you expect other people to like the show if you don't like it yourself? You've got to be the first person who says "I think this is brilliant".' (Lloyd, 2009)

- '*You have to make it for yourself. You have to make something that you think is funny and just go on your instincts and hope that your instincts are similar to those of your audience.*' (Naismith 2010)

- '*You can't make a show that you don't like yourself… They could be asked to take on a show that they wouldn't necessarily choose but in that case there's a good chance that they may well end up shaping the show in a slightly different way so it better fits their likes and dislikes.*' (Berthoud, 2010)

Making programmes for the audience:

- A producer's control can be too restrictive – Denis Norden felt the producer of **Take It From Here**, Charles Maxwell, was the *proxy audience* and so they wrote for him and not for themselves or the audience (Took, 1981, p74).

- Constraints include the 'underlying notion of the need to interpret audience desires in preference to the producer's own.' (Hendy, 2000, p110)

- Particularly at the beginning of their career, most producers will learn their trade on existing shows that they might not have a '*taste and tendency*' for. (Canny, 2011)

- '*I want the Radio 4 audience to like the shows*'. (Dare, 2012)

Restrictions of the broadcaster:

- 'The radio producer will never have total artistic control over his or her programme' due to having to sit within the expectations of the station and confines of the commissioning guidelines, the schedule, budgetary restrictions and production efficiencies. (Hendy, 2000, p70)

- Creating something that can air: '*Overall, that thing a producer does of bridging the gap between where the writer's job ends and making the thing actually work.*' (Doody, 2011)

Making decisions that one can later justify:

- '*There has to be somebody the buck stops with and I think in the event of something like a complaint, since the buck stops with him* [the producer] *he gets to make sure the buck is the shape he wants it to be in the end.*' (Doody, 2011)

This intermediary role and the associated processes can dilute the original vision. Lynn (2011, p74-75) claims that whenever he was persuaded to let production dictate the shape of the development of comedy 'the result has not been good'. Similarly, Black (2011) has found that in comedy production (albeit in US audio, television and film) it is inevitable that things will change from the original vision, and one has to just accept it:

There's something about the writing process, when you finish everything is perfect. You can imagine everything executed perfectly, and then inevitably production fucks it up. The realities of life fuck it up... What you set out to do is almost always going to be slightly different than what you end up making, just because the process is so collaborative you have so many people involved, it can't not be... I've gotten good at controlling what I can control to the extent that I can control it, and letting the rest go... Hopefully you end up with something you can live with.

Fortunately, the non-visual nature of radio and its slighter production process means that the vision might be more readily realised (Lewis, 1981, p173; Raban, 1981, p81; Tydeman, 1981, p1). The very visualization process in television can be dictatorial in its realisation. The image of Zaphod Beeblebrox in **Hitchhiker's** would have been infinitely better in the imaginations of the Radio 4 listeners than the television incarnation (Starkey, 2004, p180). Even with recently improved television production on shows such as *That Mitchell and Webb Look*, compared to **That Mitchell and Webb Sound**, Canny (2011) says: '*I've always thought they work much better on radio because they're quite high-concept a lot of those ideas. They're brilliant images on radio and on telly they're always a bit shoddily realised.*'

Radio also has the advantage, compared to TV, in that there are simply fewer people involved and thus the writer can have much more input into the whole process (Mayhew-Archer, 2011). Writer John Morton, creator of **People Like Us** and **The Sunday Format**, was amazed about how accepted it was in radio that writers get involved with the development, even casting and the production itself. On television, 'the higher the stakes, the more marginalised you get as the writer' (Bradbury and McGrath, 1998, p173). This can be seen as a benefit for radio if one has the view that, '*You're not going to get a hit by legislating through committee*' (Naismith, 2010).

But even in radio and even when the writer is performing in the show, there can be a chasm between what they thought they had created and the final product:

*The first time I listened to the **Bigipedia** pilot I was massively disappointed because I had it all in my head and of course the execution's never going to sound like that, because there's pesky things like actors and real people. It wasn't until I listened to it with other people in the room and they were falling about laughing I thought, "OK, this is good"... I wasn't prepared for it not to sound exactly like the thing in my head, which is really stupid and naïve*. (Doody, 2011)

This can be even more pronounced when the writer has such a completely different vision for the final project compared with the producer's that there is real disagreement. An example of a second series that called for diplomacy from its writer:

I'll talk to him [the producer] *beforehand and say 'These are the things we disagree on' and be more specific about the casting. Very specific about genders, about the age, and just suggest actors: leave less to him... Diplomatically it would be hard to say 'I didn't like working with x'... I mentioned to Jane* [Berthoud] *that I found it pretty difficult but I think I'm lumbered with it. I think I can manage it better and I'll be more prepared and we'll work together better.* (Anonymous Radio 4 Comedy writer, 2012)

iii Audience Research

RAJAR measurement gives data on audience ratings, but it is limited in granularity (see p81). The BBC Pulse survey gives further information about programme appreciation, but also has its limitations (see p115). This restricts the usefulness of audience research data to Radio 4 programmers in evaluating comedy shows. For example, appreciation is only measured for programmes that have been broadcast and pre-testing[103] is deemed too costly for radio. Where pre-testing has been used on occasion for television, it has sometimes proved to be misleading. Mayhew-Archer (2011) found that when pre-broadcast research was done on his

---

[103] Pre-testing in broadcasting is getting feedback from a sample audience prior to broadcast – more typically utilised ahead of movie releases.

Office Gossip television comedy, the outcome was '*this show works*'. However, the eventual viewing figures were very low. Mayhew-Archer's view is that the research had been able to show that nobody was offended by the programme, but had failed to identify that there was a lack of passion for it. Research which asks an audience how much it *likes* a new comedy can be a poor predictor of long-term performance. For example, pre-testing was carried out for the television version of *Dad's Army*, as Tom Sloan, the Head of Comedy at the time, felt that some might be offended by the subject matter. Those shown the pilot episode almost universally disliked it. *Dad's Army*, creator David Croft, was able to 'suppress the evidence', concealing the results from the Commissioners. Had this appreciation evidence been more widely available, the BBC might have decided not to proceed with one of the most enduringly popular British television sitcoms of all time (McCann, 2002, p81-82).

> *I think Geoffrey Perkins gave the example where if you say to an audience: 'We're thinking of two sitcom ideas at the moment, one is set in a shabby, genteel hotel on the south coast*[104] *and the other is set in a nudist colony' the audience will probably think the nudist colony is going to be hysterical. That's an extreme, reductive example.*
> (Schlesinger, 2011)

It is only after a series or pilot has been broadcast that audience research can be utilised for Radio 4 commissioning decisions. Raphael (2012a) often offers AIs as supporting evidence when proposing to recommission or decommission a series.[105] In turn Williams (2011), while an advocate of professional instincts, will also be influenced by convincing research:

> I think that if you do have a very clear judgement that something's right, then you will need a lot of counter evidence to make you change your mind. However, as someone with a journalistic background who values evidential programming and reporting, of course you've got to look at all the evidence and assess it and think about it. So, genuinely, I'm extremely interested in audience research that helps the programmes.

The uses of AI's within the BBC commissioning and production is described in detail in chapter 4.

2.6.4 Further Measures of Success

Once a programme has been broadcast, there are a number of routes that stakeholders will use to evaluate the 'goodness' or success of the show. Where television would use overnight viewing figures, radio must turn to other sources. Audience research is used by some but is not necessarily available or even desirable to all those involved. In addition to RAJAR quarterly data and the Pulse survey (i.e., AIs, verbatims etc…) detailed in following chapters, discussed below are further measures used by programmers to evaluate Radio 4 comedy shows:

i Chain of Command Feedback

One's success will in part depend on the feedback given from the next person up in the chain. For a writer or performer, success might have come at an early stage when their idea is chosen to be pitched by a producer. For the Producer, it is a triumph that the PDG gives them the go-ahead. The Executive Producer will be pleased to have a commission from the Commissioning Editor and, in turn, they will appreciate agreement from the Controller. Just to make it through the process can be seen as an attainment. Once the programme has been made and heard, any positive feedback that makes its way back down the chain will been seen as a measure of success. Recommission is the ultimate validation.

---

[104] I.e., Fawlty Towers.
[105] See from p88.

If it gets broadcast, that's a success. All I ever wanted. (Doody, 2011)

I aspire to keep working in comedy and the way to keep working in comedy is when I do a thing people ask for more of it. That's the most important measure for me. (Mitchell, 2010)

And ultimately whether it's recommissioned as well. Or, whether it's promoted. For example, **The Secret World** *started at eleven* [23:02] *and got repeated at eighteen thirty which I think is a promotion.* (Dare, 2012)

If the network are happy is a big one. We need to please the people we're selling work to. Feedback from the Commissioning Editor and the Controller. Things that Gwyn [Williams] *notices are really key for us.* (Canny, 2011)

Typically, in television, there can be many more people involved in the decision making process including research, marketing, production, scheduling, finance, commercial sales, online, compliance and top management (Gross et al, 2005, p83). For Radio 4, these functions tend to be involved after the programme is agreed, with just the Commissioning Editor and the Controller (with advice from scheduling) taking the call.

ii Instinct

For all those involved in the process, as instinct is involved in the creation of a comedy show, so is it prevalent in evaluating its 'goodness'. 'Experts' will tend to place their own intuition above other evidence (Kahneman, 2011, p11-235).

If I listen to it and it fills me with joy. (Doody, 2011)

Underneath it all, you know yourself whether it was a good show or an average show. (Simpson – Galton and Simpson, 2012)

...it's completely unscientific, and my gut instinct. (Raphael, 2012a).

However, it is recognised that those involved might be a bit too close to a project to be objective, or at least certainly not at the time; a truer view might only come with a retrospective eye:

You need to believe in what you're doing at the time. Everything you do, you need to think 'this is brilliant'. In terms of whether it's good, that takes a long time for you to completely know. (Mitchell, 2010)

Gross et al (2005, p83) write of how, in US television, the days of 'the fabled gut-instinct' are gone as even with a pilot, a positive reaction from executives is just the beginning of a process that involves pre-broadcast assessment through research. This is a course of action that is not open to the budgets of BBC radio.

Instinct must be tempered as comedy is highly subjective. Is it even possible for anyone in the position of Commissioning Editor to be completely objective about selecting comedy?

Probably not, but I do commission things I don't like. There are things I'd have taken off years ago, I just don't get them but I know that they're right for the Radio 4 audience. ...I know enough about the audience and I do look at the AIs and look at the feedback we get coming in. After a series you can tell whether something's hit the mark or not. (Raphael, 2012a)

Thus, Raphael identifies that her personal tastes are of lesser importance than any information that might be taken from how the audience reacts.

iii Live Audiences

The advantage of many radio comedy recordings is that they are done in front of a live audience, so the producers and performers can get an instant response. Laughter and apparent enjoyment is a key criterion for the success of a radio comedy.

I believe that the Light Entertainment producer is in a better position than any other producer, because he actually goes out in front of the audience and says, "Welcome to the Paris studio; this is what we're going to do". At the end he has to go up in front of them and say, "Thank you very much", and on his way back to the cubicle, he is abused if it's awful. So he knows, you know, what the audience thinks! In Light Entertainment you do face the audience live, all the time, and that's actually a jolly good discipline. (David Hatch: Elmes, 2007, p226)

Unlike any other radio show we do get immediate feedback because we do so many of our programmes to an audience in the first place. So it may not be the absolute Radio 4 audience but it's still a potential Radio 4 audience. (Berthoud, 2010)

We may think we know, and with enough experience we may often or even usually know, but there can be no certainty until it has been played in front of an audience. So to understand 'funny', or comedy, you have to examine the audience's response. (Lynn, 2011, p11)

The first, the obvious template of your success is the studio audience. Invariably we worked with a studio audience and if they laughed it gives you a bit of confidence. It doesn't necessarily mean that it's a great show but it certainly is an indication. (Galton and Simpson, 2012)

Reactions to certain elements within a show may also aid the production process. For example, when recording a panel show the recording is often much longer than the slot time, allowing the producer to evaluate the responses to various sections and prioritise the items that were best received for the final edit.

iv Peer Review

McLeish (1999, p293-295) discusses the process of peer reviewing as applicable to radio productions. It includes evaluating the programmes on criteria such as appropriateness for the target audience, creativity, eminence, technical achievements, and whether the programme fulfils its aim: in the case of comedy, whether it is funny or not. Within the Radio Comedy Department shows are considered within a process called programme review: '*We listen to series and we discuss them and try and refine them and work on them. Response from audience is taken so AI comes into that. Any Pulse would be taken into consideration*' (Canny, 2011). One of the Head of BBC Radio Comedy's measures of success for a show it whether it is indeed well received across the department (Berthoud, 2010).

v iPlayer/Online Listening and Downloads

Although non-linear listening is relatively low compared to traditional listening, and rights issues restrict many comedy programmes from being podcast, these platforms will undoubtedly grow and, in the light of no RAJAR overnights, the offline figures attributable to a programme will increasingly become a measure of success. Up to the current time, however, these figures are yet to be made available on a regular basis to anyone within Radio 4 programming.

The potential for these behavioural measures to aid decision making has been illustrated by, for example, Netflix, which uses analysis to develop new creative ideas - *House of Cards* being a successful example, 'using decisions based on a meticulous analysis of the viewing habits of its 44 million subscribers worldwide' (Sweney, 2014). Todd Yellin – Netflix's Vice President of Product Innovation claimed:

We climb under the hood and get all greasy with algorithms, numbers and vast amounts of data. Getting to know a user, millions of them, and what they play. If they play one title, what did they play after, before, what did they abandon after five minutes? (ibid).

While the BBC may not be able to do this yet, particularly as long as it has no demographic information regarding its online viewing, the future may bring this kind of opportunity.

vi Reviews

Reviews are cited by many involved in BBC radio comedy when discussing measures of success:

> *We have a lot of opinion formers writing about* [comedy] *shows both in radio and television. I hate to admit it; I would always want to know what Gillian Reynolds thought of something.* (Schlesinger, 2011)
>
> *I look at the press cuts every morning.* (Raphael, 2012a)

As with audience research, reviews can often be held up when they support an argument and ignored when they say the *wrong* thing. The subjectivity of critique means that bad reviews, ones that do not support the decision-makers' opinions, or those from unimportant quarters, can be brushed aside:

> Talking to Gillian Reynolds – 'This is going to sound like appalling flattery; I do read with great respect what you write Gillian because your love of radio just comes out of every pore of you. And I know from talking to you how much you know. So I do read with respect what you say and it's an extra element of what people are thinking. Of course I pay attention but it doesn't necessarily alter what I do.' (Williams, 2011).
>
> *Very important, at least for certain reviewers. You always wanted to get a good review from Nancy Banks-Smith* [television critic for The Guardian since 1969]. *You never bothered too much about getting a good notice from The Sun because it's a different level of criticism.* (Simpson – Galton and Simpson, 2012)

…and this is understood by the reviewers as they are writing for their readers and not as an aid for the commissioning process.

> *People make up their own minds. And certainly as far as 'professionally' goes, nobody at the upper echelons of the BBC takes any notice of anything any critic writes unless it suits their political purpose. And I speak from long experience and with no bitterness at all.* (Reynolds, 2011)

A reviewer is not employed because they are representative of the listening public; they are in the role primarily because they write well about the subject. Whereas a reviewer for live comedy will be experiencing a performance surrounded by other audience members who in turn may shape the reviewer's opinion, the radio critic has only their own response to go by.

> [Live comedy reviewing] *If you disagree with the room, if the room's clearly hating it and you're liking it or the other way round, then you've got to mention it. To be a good comedy reviewer they'll have needed to see quite a lot of comedy. They'll need to be able to give a coherent opinion, why they like it or why they don't like it. It's got to be an entertaining read as well. It's got to be well written.* (Bennett, 2011)
>
> *What you need if you're a radio critic is a sense of the medium, of its continuous flow, and of the possibilities for innovation for a medium that depends on familiarity… A love of the medium. An ability to make an essay out of a whole weeks programming, to make some kind of an argument.* (Reynolds, 2011)
>
> *Reviewers are a strange bunch.* (Mayhew-Archer, 2011)

…but if they do say the *right thing* in the view of the producer, good reviews can be used as supporting evidence to push for recommission:

> *Do we like it if the critics say it's good? Yeah, that does have some effect but only in as much as we can use it to back us up. If I like it and the critics don't like it, tough. But if I like it and the critics do like it, that's always good for us, particularly if Controllers or commissioners have been in two minds about it, then that's really useful.* (Berthoud, 2010)

The main issue with radio comedy is that it is often ignored compared to television and to get any kind of review, good or bad, is not guaranteed. When Douglas Adams asked Simon Brett what the reviews for **Hitchhiker's** would be like, Brett said: 'This is radio, Douglas. We'll be lucky to get a mention anywhere' (Webb, 2003, p135-136).

…it rarely makes the front pages (Hendy, 2000, p3).

There's nobody who's seriously reviewing any of our output. There's no active pursuit of the work we do… There's no comedy specialist who's listening with any ear so we're in the lap of amateurs. It's the curse of telly and radio. The biggest forms of expression of comedy in our culture, goes without any serious critical analysis. It's bizarre. (Canny, 2011)

vii Word of Mouth and Chatter

Considerations when evaluating a programme are increasingly including the amount of 'noise' around a programme. For example, if it is being discussed on twitter this will indicate that the audience is engaging with the show (Dare, 2012). Measures are being developed, such as the 'Content Power Rating' (CPR), that evaluates programme performance through 'audience delivery, involvement and advocacy' via monitoring of social networking and media coverage (Frean, 2011). Stone (2012) illustrated how engagement with a programme could be quantified by showing the number of tweets using the example of George Entwhistle's conversation with John Humphries at around 08:30 on the *Today* programme on the 10th November 2012 that lead, in part, to Entwistle's resignation (chart below). His contention is that this kind of 'real time measurement' could change the future of radio as it can indicate volume and advocacy of listening: 'It might actually change the type of content that we produce, and certainly it means that radio is a lot more transparent about what works and what doesn't work'.

Figure 8 – Twitter reaction – George Entwistle on Today. Sat 10th Nov 2012 – Stone, 2012
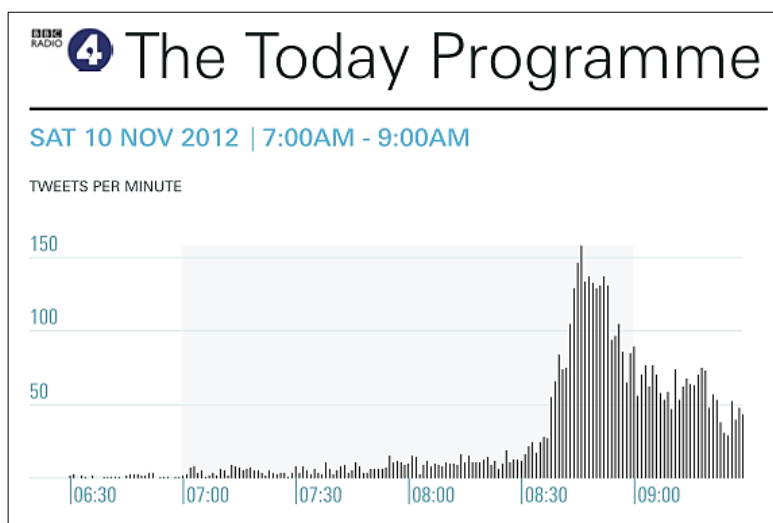
Positive social media reaction can aid producers:

It helped enormously getting good press for it [BBC television's sitcom - *2012*]. *I think our argument was strengthened with the channel substantially because of the reception it got. And in fact, the amount of twitter conversations, a phenomenal amount.* (Schlesinger, 2011)

However, when negative social media response can be detrimental, such as in July 2013, when Shane Allen, BBC Controller of Comedy Commissioning admitted to the Broadcasting Press Guild that Ben Elton's *The Wright Way* was not being recommissioned due to getting a 'barrage' of abuse on social media (Lawson, 2013). He indicated a 'growing concern' about launching a new comedy in the age of immediate viewer reactions via social media.

While measures such as CPR could provide an objective view of the amount of discussion around a programme, direct word of mouth is still a factor in how the decision makers perceive the success of a programme:

> *And then it's that kind of intangible, some things just stick. There are sticky shows. It may be the woman next door, it may be people here in the lift, or standing in a line at a festival, it may be somebody you meet at a dinner party when, God almighty, they finally winkle out that you work for Radio 4. And they say, 'oh that one about the cookery writer or that one about the grumpy old writer', and it sort of happens. Some of these shows just stick and so it's anecdotal.* (Raphael, 2012a)

…recognised also in the pre-internet era as an important factor in comedy:

> It [**The Burkiss Way**] had not just captured a young audience, but exhibited many other features of the true cult: an audience that grew by word of mouth rather than official publicity, a language of its own that could be shared among fans. (Hendy, 2007, p190)

Direct feedback from those one comes in contact with however, could be misleading as all one's friends might hate a show that the audience as a whole may not, and vice versa. There is no guarantee that they represent the audience as a whole:

> Galton*: That's the funny thing about comedy. You might get the world record on how many people watched it but there's sure to be somebody who said "Hmmm…"*
> Simpson*: "And always the ones you listen to.*
> (Galton and Simpson, 2012)

<u>viii Awards</u>

An award is the ultimate validation for a radio comedy show; unfortunately high profile awards in this arena are not common. The most well-known radio honours are the Radio Academy Awards (previously known as the 'Sonys') and these include a category for 'Best Comedy'. The winners and nominees for 2012, listed below, were dominated by Radio 4 shows as it is the primary purveyor of radio comedy.

> Gold – **Mark Steel's in Town** – BBC Radio Comedy for BBC R4
> Silver – **Another Case of Milton Jones** – Pozzitive Television for BBC R4
> Bronze – **Down the Line** – Down the Line Productions for BBC R4
> Nominee – **The National Theatre of Brent's Iconic Icons** – CPL Productions for BBC R4
> Nominee – *Adam and Joe* – BBC Radio 6 Music

While one might assume that the best shows will make the shortlist, there is a cost of around £100 for any show to be entered into the competition, and a view from some that the process has become a '*commercial enterprise*' (Naismith, 2010); it may not be the best radio comedies that win, but the best that have been entered.

In 2012, the BBC introduced a new Audio Drama award in order to, 'celebrate and recognise the cultural importance of audio drama, on air and online, and to give recognition to the actors, writers, producers, sound designers, and others who work in the genre' (BBC Press Office, 2012). Within this is a category for scripted comedy[106] – finalists listed below for 2012 – again dominated by Radio 4:

> Winner – *Floating* – BBC Radio Drama for Radio 4
> Shortlisted – **Cabin Pressure** Pozzitive for Radio 4
> Shortlisted – **Ed Reardon's Week** BBC Radio Comedy for Radio 4

However, as with the Radio Academy awards, only those entered are eligible and for 2013, the BBC Radio Comedy Department missed the deadline and failed to enter any of their programmes for the award.

Further to this, there are on occasion awards given for radio comedies across broader categories. For example, The Broadcasting Press Guild awarded **Ed Reardon's Week** the prize for 'Radio Programme of the Year' in 2011.

---

[106] For 2013, there were 2 categories for radio comedy – comedy drama and audience comedy

Furthermore, working with comics who have won awards outside of the radio medium may still be seen as a mark of success:

> *There's that kind of reflected glow when somebody gets the Edinburgh comedy award and they're already on Radio 4. We will talk about working with award-winning performers and writers; it's just not a radio award that they've won.* (Raphael, 2012a)

2.7 Evidence Supporting the Suitability of Using Radio 4 Comedy to Analyse Appreciation Responses.

- Radio 4 is heard by many people in the UK and comedy is broadcast as part of its service licence commitments. Millions of people listen to Radio 4 comedy every week. It is relatively costly to make.

- Radio comedy is particularly important in all comedy because of its role as a test bed for TV comedy.

- Success of comedy is difficult to predict, thus, additional information that aids understanding of performance is of value. Comedy is thought to be polarising, so understanding the extent of this is important.

Walters was a radio commissioner. Of course. Of course. I should have known.
He had that lifeless, grey, dead-eyed quality that they all have at Broadcasting House.

(Partridge, 2011, p92)

Chapter 3 expands upon the concept of audience measurement firstly by giving a brief history of the development of systematic research at the BBC. It goes on to explain the common currency of measuring size of radio audiences, then considers the limitations of this system, concluding that an alternative method is needed for radio episode evaluation.

3.1 Development of Systematic Research at The BBC

3.1.1 Who's Asking?

In the United States, listener research began as part of market research; there was a vested interest in commercial stations giving people what they wanted in order to increase ratings and consequently revenue (Street, 2002, p59). Without a commercial incentive, during the 1920s, the BBC would merely attempt to interrogate secondary information obtained from a number of sources including: the number of licences bought, genre specific advisory committees, correspondence with listeners, professional criticism and actual times of listening taken from information from a wireless exchange (BBC Year Book, 1932, p105-108). In addition, newspaper competitions were run that incentivised audiences to vote for favourite programmes in return for the possibility of winning a prize (ibid). The first real pressure within the BBC for systematic research to be commissioned came from those involved in educational programmes – the Central Council for Broadcast Adult Education – with proposals being submitted as early as 1929 (Briggs, 1965, p257). A survey was eventually undertaken on behalf of the Council in 1932, its primary aim to inquire into the 'tastes and habits of the ordinary listener' (BBC Handbook, 1932, p161). In attempting to explain to the readers the rationale of the survey, the handbook reads: 'one is left in the isolation of headquarters with the uneasy sense that the service is still inadequate to the potential demand' (ibid, p162). Reith did not advocate this sentiment and was confident that those at the BBC knew better than the listeners. He resisted the push for audience research because, 'he knew that a comprehensive investigation of listener tastes would influence and eventually dictate broadcasting policy, and that worthwhile programmes for minorities would be sacrificed to the ratings' (Crisell, 2002a, p46). Indeed, any research done around this time had no influence upon programming decisions and in one case in 1930 was merely used as material for a comic sketch (Briggs, 1965, p258).

From 1930, there were increasing requests from research and 'Odd scraps of information whetted the appetite' (Gorham, 1948, p59). For example, the relay companies could indicate relative audience sizes for different dayparts based on the load on the system (ibid). Many of those interested in the listeners' habits wanted to identify what the audience wanted, rather than dictate to them (Crisell, 2002a, p45). Val Gielgud, Director of Drama, 'legendary' and intimidating, known across the BBC for his 'sometimes cutting witticisms and his sword-stick' (Silvey, 1974, p25), 'opera-cloaks and beards' (Gorham, 1948, p25), wrote in May 1930 following a meeting of the programme board:

> I cannot help feeling more and more strongly that we are fundamentally ignorant as to how our various programmes are received, and what is their relative popularity. It must be of considerable disquiet to many people beside myself to think that it is quite possible that a very great deal of our money and time and effort may be expended on broadcasting into a void. (Geilgud: Briggs, 1965, p259)

Gielgud and his colleagues were ambitious in the information that they sought, and aspired to gain knowledge, not just of numbers of listeners, but their tastes and habits based on their sociological groups, much like lifestyle segmentation might be used today (Briggs, 1965, p260). But Reith was not alone in being sceptical about pandering to the masses; J.C. Stobart, the BBC's first Director of Education, represented the views of many of the incumbent managers when telling a 1930 Programme Board meeting 'broadcasting is not and should not be democratic' (Briggs, 1965, p260). He went on to write:

> I hold very strongly that the ordinary listener does not know what he likes, and is tolerably well satisfied, as is shown by correspondence and licence figures, with the mixed fare now offered. I cannot escape feeling that any money, time or trouble spent upon elaborate enquires into his tastes and preferences would be wasted. (Briggs, 1965, p261)

So, due to the significant body of resistance, the 1930 discussions did not pay off, but Gielgud was not ready to give up. He thought that understanding of the critical reception of programmes – restricted as it was mainly to press criticism or listener correspondence – was inadequate. In 1934, he went on air to request information about the BBC's drama broadcasts directly from the listeners, asking:

> Write to us and say as candidly, as clearly and as categorically as you can what you feel about the whole question. Are there too many plays broadcast? Are there too few? Do you hear the sort of plays you like to hear?… To some extent, our future dramatic policy will depend on the result of this appeal. (ibid, p263)

Gielgud received 12,726 letters in response of which only 3% were 'critically adverse' (ibid). In this exercise he had clearly shown that there were listeners interested in providing their opinions for the BBC.

From 1930, those inside the BBC advocating audience research had referenced comprehensive foreign surveys of listening habits, particularly one done in 1928-29 in the Berlin region. So, the argument for this kind of survey was strengthened in the mid-thirties as its usefulness was illustrated by the success of Nazi broadcasting. As concluded by *The Times* in 1935, their success on radio was partly due to their 'elaborate measures' taken '(a) to get a picture of the community and (b) engage its interest' (ibid, p257). This proof of efficacy supported the argument made by increasing numbers of BBC managers who now expressed that they wanted more information about their listeners, listeners that by 1935 accounted for potentially 98% of the UK population (Crisell, 2002a, p45).

During the 1930s, commercial stations such as Radio Normandie, Luxemburg and Athlone were beginning to do their own surveys, but their figures were not taken seriously by the BBC (Gorham, 1948, p59) even though it was unable to challenge the data as it had no research of its own. A further driver was the British press, which 'taunted' the Corporation for its lack of research (ibid, p59), for example:

> At present the BBC has to depend mainly for its contacts with the public on the letters it gets from all those who listen to its entertainment. Letters are a better contact than none at all, but after all, the man or woman who, having heard a programme he likes or dislikes, sits down and writes to the BBC to say what he thinks and why is rather an exceptional than a representative member of the audience. Ought there not to be some genuinely representative body like the Committee of Businessmen which I believe is occasionally consulted by the Post Office? You want a Consumers' Council, so to say, to guide the policy of the BBC programmes. (Daily Sketch, 10th July 1935 – Briggs, 1965, p265)

But there still remained as many at the BBC who were resistant to what was seen by then as 'a surrender to an ignorant clamour' (Silvey, 1974, p32). The main advocate for research at this time was the Controller of Public Relations, Sir Stephen Tallents. He lobbied the Heads of Department for support of his proposal for extensive listener research, resulting in the submission of a paper to the General Advisory Council in January

1936 (Briggs, 1965, p266-267). As with the requirements laid out in the early 1930, the information sought remained ambitious; Tallents' paper identified four areas of inquiry – audience <u>habits</u>, <u>preferences</u>, <u>reactions</u> and <u>efficiency of broadcasting techniques</u> (ibid). Tallents was clear however that the proposal was in no way designed to influence programming decisions unduly as 'the surrender of programme policy to a plebiscite would undermine the responsibilities imposed on the BBC by its charter' (ibid). Reith eventually conceded and agreed to a limited experimentation into audience research in March 1936 (ibid, p269).

3.1.2 Size Matters

So, Gielgud et al's requirements would begin to be satisfied finally, despite the fact that there was still a significant faction that considered surveys not to be the way to go, 'dismissed as utopian both by some of those who favoured them in principal and by others who openly declared them dangerous' (Silvey, 1974, p14). Tallents, however, wasted no time after being given the go ahead and set up a Listener Research Group in April 1936 (Briggs, 1965, p269; BBC Year Book, 1939, p55) to enable the buy-in of various areas of the BBC (Silvey, 1974, p24). He then employed Robert Silvey from the London Press Exchange – after informal interview at the Garrick Club – to head up the initiative, resulting in him joining the BBC in October of 1936 (ibid, p17).

The Listener Research Group included the outspoken Director of Drama, Gielgud, and Silvey found him to be a strong ally. This factor, in addition to Silvey's personal familiarity with the genre and its range of output, saw radio drama as the choice of programming with which to conduct the first trial in audience research at the BBC (ibid, p59). An experimental Drama Panel was set up consisting of 350 people based from a convenience sample of volunteers who had expressed interest in the genre. This ran for 3 months and covered forty-seven programmes. The results, evaluated in June 1937, were never thought to be statistically representative of the population and at best gave comparative results between the productions rather than any attempt to measure audience size or quantification of a measure of quality (ibid, p58-61). This initial survey was merely testing the water.

Expanding this research to another genre with a larger sample was the next step and a panel of volunteers was recruited in September 1938 through an advertisement in the *Radio Times*, and press and radio coverage. Seeking respondents who were fans of 'Variety', there was an overwhelming response and 28,000 postcards were received offering the services of 47,000 volunteers. 2,000 of these people were chosen at random and The Variety Listening Barometer was born; it was so called because it was intended '*to measure pressure rather than heat*' (ibid, p79). At the beginning of October 1938 this group of self-professed variety fans logged the programmes they heard against a list of programmes issued regularly by the BBC (BBC Year Book, 1939, p59). There was a worry that respondents might listen exceptionally, so included in the preliminary guide was the direction:

> Please do not feel you must make a special effort to listen because you are keeping a log. What we want is a record of ordinary listening. Therefore please listen just as much or as little as you would have done if you had never heard of this scheme. Remember, we are just as interested in which programmes you did not hear as which ones you did. (Silvey, 1974, p79)

The results of the 12-week initial trial gave 'comparative magnitude of the audiences' (ibid, p80) for the shows and sub-genres, and general findings about listening patterns for specific days and times. Well aware that this band of enthusiasts may well be atypical, Silvey introduced a control group to validate the finding and, as expected, the variety panel's claimed listening to these entertainment shows was far higher than that of the control group. However, it was found that across different shows, the 'relationship was pretty constant

from programme to programme' across the two groups (ibid, p82), and the results were seen to provide 'instructive commentary on the listening figures' (Gorham, 1948, p166).

One of the key conclusions from these pilots, however, was that there was indeed a need to have an understanding of the absolute sizes of the audiences for individual programmes. But, before the question of ratings could be answered, a greater understanding of the general listening population was required: 'at the time, there were still senior officials in the BBC who found it hard to believe anyone dined before 8pm' (Silvey, 1974, p62). The middle-class, London-centric BBC staff had little knowledge of non-urban, working-class life (Crisell, 2002a, p44). So, a survey was undertaken, the broad objective being to find out which genres listeners preferred at different times of the day, the title of the questionnaire being *What do you like?* (Silvey, 1974, p65). Towards the end of 1937, single sided foolscap sheets were distributed to 3,000 randomly selected households from the sampling frame of holders of radio licences. 44% of the households responded and over 3,100 questionnaires were returned. A similar but more extensive exercise was done in the summer of 1938 where the responses totalled 12,700. The resultant information was denoted as '*tastes*' which indicated '…attitudes towards broad categories of output as distinct from particular broadcasts. Similarly, we used the word "reactions" to refer to the way people felt about specific broadcasts' (ibid, p65-66). For simplicity at this stage, the *liking* of a genre was measured purely by the percentage of respondents who said that they *liked* any of the twenty-four categories rather than any *depth* of liking. The results varied from 'Variety' (the most comparable of the segmented genres to what we would today call comedy) being liked by 93% of the respondents compared to just 8% liking 'Chamber Music'. 'Musical comedy' weighed in at a respectable 69% but still well below 'Theatre and Cinema Organs' at 82%. This research not only gave the BBC an idea of the comparative sizes of the audiences for the differing genres, albeit claimed rather than behavioural, but also their composition, having recorded the basic demographics of the respondents (ibid, p62-68).

Off the back of The Drama and Variety Listening Barometers, it was agreed that the principle should be extended to all listening and the next step was the introduction of the General Listening Barometer. The aim now was to, 'estimate the *actual* number of listeners to each programme, not merely their relative sizes as we had previously done' (ibid, p89), through having a properly designed and robust continuing sample. This plan was first implemented with a quota sample of 3,500 homes in December 1939, despite the looming shadow of WWII. Each day 800 people were interviewed to find out which programmes they had heard the previous day (BBC Year Book, 1941, p78), a method of day after recall. The interviewers – employed under contract from the *British Institute of Public Opinion* (BBC Year Book, 1942, p79) – were trained to elicit the most accurate representation of listening patterns, using a printed list of the previous day's schedule as a prompt for the respondents (BBC Year Book, 1941, p78). The output of the survey gave what was called a programme's 'listening figure' (Silvey, 1974, p100), and was in fact the *proportion* of the sample who claimed to have heard the broadcast. The interviewers merely asked whether the respondent had heard the programme so there was no information regarding how much of each programme was listened to at this stage. Nor was there the distinction between *hearing* or *listening to* any particular broadcast. Each listening figure, by inference, gave the approximate percentage of the adult civilian population who would have heard the programme. Percentages (effectively an estimate of reach) were seen to be better expressions of the audience size for distribution, as to use them to calculate an absolute audience size would be to 'lend a spurious air of precision to what professed to be no more than approximations' (ibid, p101). As these figures were based on daily samples of around 800 interviews, there was an element of fluctuation in programme figures. The researchers encouraged small variations to be ignored and instead placed greater reliance on averages over longer time periods. Silvey notes that broadly the results were 'astonishingly consistent' (ibid,

p95) and this was a reassurance to colleagues even though Silvey himself understood that consistency did not prove that the methodology itself was 100% foolproof. The results gave the BBC a Survey of Listening that was implemented continuously, with only one break of ten days due to a war-time bomb. (BBC Audience Research Department, 1953, p7). After the war, this survey was occasionally checked for validity; the primary concerns were that there could be issues with the legitimacy of the sample, and detrimental effects from recall errors:

> Comparisons of the results of these criteria studies and the normal Survey were highly reassuring. They left us satisfied that the use of Quota sampling was fully justified and that our methods of Aided Recall, though capable of improvement, were basically sound. (Silvey, 1974, p94)

Furthermore, it was felt that there needed to be a greater understanding of the listeners by region, so the sample was increased to 600 each for the following areas: London region, West England, Midlands, North England, Wales and Scotland, giving a total of 3,600 respondents (Northern Ireland was added later) (Silvey, 1974, p128). From 1952 the barometer became the Survey of Listening and Viewing as the interviewers included television viewing into their questioning (BBC Audience Research Department, 1953, p7). In 1953 the interviewers were issued with a handbook that set out the best practice for aspects such as fulfilling quotas and the best way to interview (ibid, p35-36). It also gave guidelines on what actually constituted *listening* and it was here made clear that listening only counted if at least half a programme was heard (except for the news) (ibid, p35), and that if the respondent admits to not paying full attention to a show, the interviewer may use a their 'common sense' as to whether it might count or not (ibid, p36).

So, the size of the audience could be estimated with acceptable certainty, but what of their reactions to the programmes?

3.1.3 Goodness me!

In the winter of 1940-41, Silvey became increasingly convinced that the barometer was only giving half the story (BBC Year Book, 1942, p79), and that audience size figures 'should be supplemented by a continuous assessment of audience reaction – what listeners *felt* about the programmes they were listening to' (Silvey, 1974, p113). In 1941 10,000 people volunteered their services having read in the Radio Times or heard on air about the BBC's endeavour for a greater understanding of their listeners. From these volunteers, groups of 500 were recruited to form ongoing panels for various genres, one of which was subsequently called 'Light Entertainment'. Members of each panel were sent tailor-made questionnaires every week for a selected number of broadcasts relating to their specific genre of interest. Each questionnaire was brief and consisted of only four or five questions encouraging open-ended responses. All questionnaires, however, did have one question in common and that was a request for the respondent to rate the programme 'out of ten', i.e., an 11-point scale from 0-10. The Appreciation Index (AI) was expressed as a single figure calculated as a mean from all the scores, and multiplied by ten to differentiate it from any individual score. Hence, the AI at this time could be expressed as a figure from 0-100 (ibid, p113-116).

This system of appreciation measurement yielded AIs which were clustered between 65 and 85, indicating that those listening were generally liking what they heard, albeit from a panel of attested fans of the genre. In an attempt to gain more granularity in this upper range, the marks-out-of-ten were abandoned after the war in favour of a five-point alphabetical scale with corresponding verbal and numerical equivalents (ibid, p116-117):

A+ [worth 100] stands for:

I wouldn't have missed this programme for anything, or…
I can't remember when I enjoyed (liked) a programme so much, or…
One of the most interesting (amusing, moving, impressive) programmes I have ever heard

A [worth 75] stands for:

I am very glad indeed that I didn't miss this, or…
I enjoyed (liked) it very much indeed, or…
Very interesting (amusing, moving, impressive) indeed

B [worth 50] stands for:

I found this quite a pleasant (satisfactory) programme, or…
I quite enjoyed (liked) it, or…
A quite interesting (amusing) programme, or…
A rather moving (impressive) programme

C [worth 25] stands for:

I felt listening to this was rather a waste of time, or…
I didn't care for this much, or…
It was rather dull (boring, feeble)

C- [worth 0] stands for:

I felt listening to this was a complete waste of time, or…
I disliked it very much, or…
It was very dull (boring, feeble)

As with the earlier panel pilots, it was accepted that the views of panels of enthusiasts might not be completely indicative of the population as a whole, perhaps 'an exaggerated version of the facts' (ibid, p115). This understanding still allowed for the results to allow interpretation of the *relative* appreciation between programmes *within* any one genre. The generalisability of the appreciation data was much improved when the post war sample was increased to 3,600, in order to have a more representative sample coving all genres (Silvey, 1974, p128). This change coincided with the change from the 11-point scale to that of the 5-point measure. However, it was not until decades later in 1984 that the panel was recruited on a true quota system. Up to that point the self-selected panel tended to be unrepresentatively middle-class, (BBC Broadcasting Research Department, 1986, p37). However, it was not until 1990 was the system updated to include all network radio programmes, rather than just ones pre-selected (BBC Broadcasting Research Department, 1993, p17-25).

3.1.4 Ready to Report…

During most of the 1940s, 1950s, 1960s and 1970s, the output of systematic radio listening research was amazingly consistent. A daily 'Listening Barometer' report (for some years combining radio and television) was produced that showed the % reach for each programme broadcast (sample size allowing), at regional level in the early days then at national level only. In addition to the size of the audience, the document highlighted the programmes for which a full audience report had been conducted including AI scores when available. The example below from the BBC's archive's at Caversham shows a report from the day of the first TX of *I'm Sorry I Haven't A Clue* on 11th April 1972. It shows that the programme was heard by 2.1% of the adult population in England:

The Φ symbol seen against the programme was the indicator to show that an audience report had been conducted. That document, entitled *An Audience Research Report*, was produced for a number of programmes selected each week from the data collected from the genre panels. The information in these reports (usually between 1 and 3 pages of foolscap), consisted of the following (Caversham Archive Reports, 1955-1978):

- Title and transmission <u>details</u> of the programme
- <u>Size</u> of the audience estimate – percentage reach from the Barometer – and commentary about how that might compare to programmes of a similar type or in the same slot.
- <u>AI</u> for the programme (sometimes known as RI Reaction Index or GE General Evaluation) and similarly commentary about how that might compare to programmes of a similar type or in the same slot.
- The <u>distribution</u> of the AI scores across the various ratings and the size of the sample used.
- Commentary on the <u>general impression</u> of the programme with verbatim examples from the respondents.
- Commentary about what respondents may have particularly <u>liked or disliked</u>.
- Commentary on the <u>answers to specific questions</u> about the programme with verbatim examples from the respondents such as:
  - What they thought of the whole series.
  - How often they listened.
  - Whether promotional action encouraged them to listen.
  - Opinions of elements within the show.
  - Opinions of the performers.

For example, for the same episode of *I'm Sorry I Haven't A Clue* as seen in the previous barometer, the first page of the Audience Research Report details the distribution of appreciations scores and the interpretation of verbatim responses of the respondents to the show:

Figure 10 – BBC Radio Audience Research Report – **I'm Sorry I Haven't A Clue**, 11/04/72 – BBC Caversham Archives
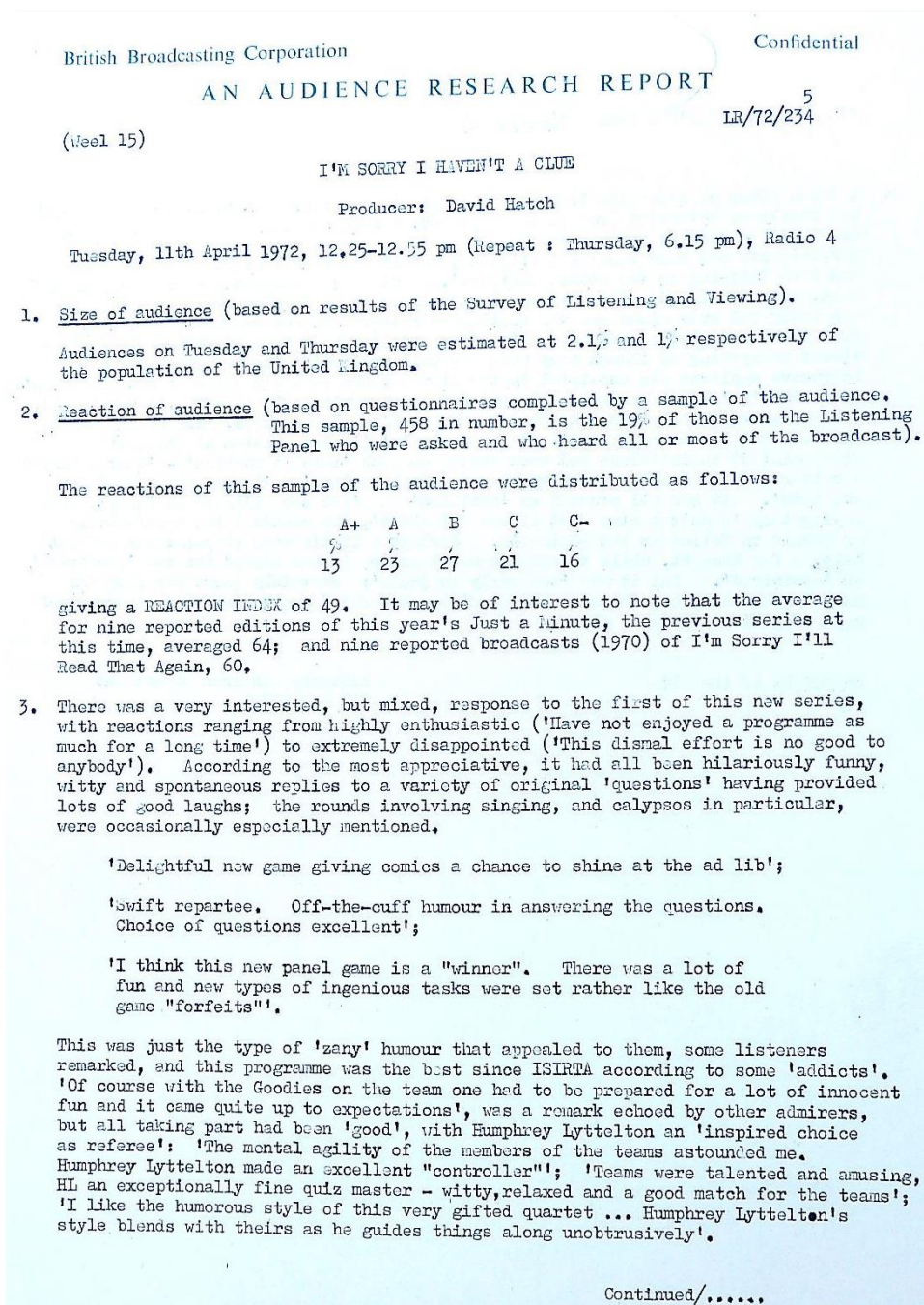
Reports appear to have been conducted for the first and/or last episodes of series and for one-off prominent programmes, although there were no hard and fast rules in this respect:

> It would have been impossible for us to have produced reaction reports on every broadcast, nor was is necessary that we should do so, but from the first we covered a substantial proportion of them – to begin with about five hundred a year which increased after the war to about three thousand. The choice was a matter for negotiation. Producers would ask for the reports and so could heads of output departments. The final choice was made in consultation with the service editors, a consultation in which Audience Research frequently took the initiative. (Silvey, 1974, p118)

There was a generally a time lag of two to three weeks between a programme's transmission and the production of the reports due to the time it took to receive the information back from the respondents and then collate the data. It could be even longer; for example, the report issued for *The Hitchhiker's Guide to the Galaxy* episode 2 (no report was done for episode 1) was completed on 19th April 1978 while the transmission was five weeks earlier on 15th March (BBC Archives, Caversham). During the 1950s, there were 25 full time staff producing the reports constructed from fieldwork via the 1,200 part-time field workers. Based at The Langham Hotel – across the road from Broadcasting House – the team was described by historian Burton Paulu as 'probably the largest department of its type maintained by any broadcasting organisation in the world' (Briggs, 1995, p21).

From 1936, Silvey's department had been known as 'the Listener Research Section' and fell under the Public Relations Division. This was the case until the war when the public relations division was dissolved and research became a department in its own right coming now under the Radio Programme Division (Silvey, 1974, p25). Silvey felt that this structure was not conducive to the most dispassionate, unbiased work which should be expected from a research body: 'It was in principal undesirable that a department like ours should be put under the authority of those it might have to criticise' (ibid, p26). The subsequent restructure after the war accounted for this issue and the research department moved to fall directly under the control of the Director General (ibid).

The reports were available for any BBC staff member who wished to see them – an important aspect for Silvey – but were specifically sent to those most concerned: the producer, the head of department and the planners (schedulers).

> To study these figures, as we did, day by day seven days a week, gave you the feeling of being in touch with your audience much more satisfyingly than depending on letters, which are often written by minorities, or by taking the opinions of the comparatively few people around London whom you can meet yourself. (Gorham, 1948, p166)

It was found that although these reports were no more than broad brushstrokes of how a programme had been received, a particular benefit was found for programmes with smaller audience sizes in that it could be illustrated that they could still be highly appreciated by the listeners who heard it (Silvey, 1974, p118-119).

Thus, Silvey had established a pattern of ongoing, systematic research that was to change little for 40 years, the output of which did indeed answer Tallents' original questions regarding: the habits, indicated by ratings; preferences, indicated by ratings and appreciation; reactions, indicated by appreciation; and efficiency of broadcasting techniques, indicated by responses to tailored questions. The basic structure of Silvey's research was in place for decades with the methodology of the Listening Panel (appreciation measurement) remaining virtually unchanged until 1984 (BBC Broadcasting Research Department, 1986, p37). The broad changes over the history of BBC Radio systematic research are highlighted in the table following:

3.1.5 Summary of BBC Radio Appreciation Measurement Research Development

Figure 11 – Summary of BBC Radio Appreciation Measurement Research Development
(NB: points highlighted in red indicates where the appreciation scale was altered).

Radio Audience <u>Appreciation</u> Measurement	Radio Audience <u>Size</u> Measurement
1936 – BBC Audience research Department set-up.[107]	
1937 – First listening panels set up, comparative programme 'liking' within genres evaluated.[108]	
1938 – 12,700 people surveyed to gain understanding of their 'tastes' over all genres.[109]	1938 – Variety Listening Barometer gave relative sizes of comedy audiences.[110]
1941 – The Listening Panel set up – Groups of 500 respondents per genre were recruited to rate programmes on a numerical 0-10 scale to calculate AIs.[111] From this point until 1990 only selected programmes are monitored in terms of appreciation.	1939 – General Listening Barometer gives daily figures. Set up to sample 3,500 homes in order to estimate absolute audience sizes – around 800 interviews per day.[112]
Post war – To improve differences in AIs between programmes, the scale was changed to a 5-point Likert-style verbal scale giving equivalent ratings with the following weightings: A+ = 100, A = 75, B = 50, C = 25, C- = 0.[113]	Post War – The Listening Barometer became a fully representative sample across all genres & regions using day-after recall interviews.[114]
1950 – Radio AIs now called RIs (Reaction Index) as AI (appreciation index) is reserved for television appreciation and different names discourages direct comparison.[115]	1952 – The Listening Barometer became a Survey of Listening *and* Viewing.[116]
Mid 1970s – AI measurement is now called G.E. (General Evaluation) and is based upon respondents rating across a bipolar scale of 5 points from '*Well worth hearing*' to '*Not worth hearing at all*'.[117]	1973 – Independent radio sets up JICRAR to measure commercial radio.[118]
Up to 1984, the 3000 strong Listener Panel was still recruited through broadcast appeals. The panel was refreshed on a continuous basis and each member served for 2 years. Every member was each sent a weekly questionnaire with 50-60 selected programmes broadcast that week across all the national networks.[119]	1981 – The BBC's Daily Survey limited now to just radio size as BARB takes over television and the survey is improved with new recall aids.[120]
1984 – The new Listening Panel is set-up, consisting of 3,000 representative respondents, including 500 who are specially selected as Radio 3 listeners as a boost for responses to this niche network.[121] They are recruited from the BBC's nationally representative Daily Survey panel.[122] The scale changes to an 11-point scale based on the question '*How many marks out of 10 would you give this programme?*' The result is called the Reaction Index – RI.[123] (Menneer claims that this change was '*intended to engineer greater discrimination in scores between programmes*').[124] Reactions up to 1990 were still only measured for a selected list of programmes.[125]	1983 – BBC Daily Survey changes from 2,000 per day to 1,000 per day but has improved quota sampling. Interviews now take place in the home rather than the street.[126] 60 sampling points (roughly pairs of streets) are selected at random each day and 18 interviews are conducted for each sampling point.[127]
1990 – The scale is updated to a 6-point verbal Likert-style scale: '*with the intention of making the scores more discriminating*'. Poor = 0, Fair = 20, Fairly Good = 40, Good = 60, Very Good = 80, Outstandingly Good = 100. The results were found to be at least as discriminating (having as wide a distribution of mean scores) as the previous scale but through a lower aggregate. For example, when comparing at the aggregate RI for Radio 4 shows common across the old and new system, the old RI from an 11-point numerical scale gave 78 [1989] while the new RI on the 6 point verbal scale returned 65 [1990]. The Listening Panel questionnaire was expanded to incorporate all networked programmes on Radios 1,2,3 and 4.[128]	1988 – The BBC and JICRAR conduct an experiment to compare claimed listening via daily day-after recall to weekly diaries. After reconciliation, weekly diaries recorded higher listening; a theory was mooted that they recorded 'hearing' while the BBC's figures documented 'listening'.[129]
1992 – Radio Opinion Monitor (ROM) set up and administrated by Ipsos-RSL. On behalf of the BBC, they collect Reaction Indices to each programme based on a '*marks out of ten*' evaluation (actually 1-10), recruited from the RAJAR sample. A panel of 4,500 completes a weekly diary once every 2 weeks.[130]	1992 – RAJAR takes over UK radio size measurement and combines BBC and commercial stations.[131]
2004 – Online panel piloted and found to have similar results but with increased sensitivity[132] 2005 – Pulse Panel recruited and administered now online by GfK which provides the AIs based on the 1-10 scale. A panel of 20,000 are used.[133]	

[107] Silvey, 1974, p25-26.
[108] Silvey, 1974, p58-61.
[109] Silvey, 1974, p65-66.
[110] Silvey, 1974, p79. .
[111] BBC Broadcasting Research Department, 1986, p37
[112] Silvey, 1974, p89.
[113] Silvey, 1974, p116-117.
[114] Silvey, 1974, p128.
[115] Silvey, 1974, p158-p162.
[116] BBC Audience Research Department, 1953, p7.
[117] Menneer theorises on the change; 'I guess A+ to C- came to be seen as a rather erudite, middle class type of scale. Not well suited to the less well educated. The verbal scale was intended to be self-explanatory?' (Menneer, 2014).
[118] Kent, 1994a, p17.
[119] BBC Broadcasting Research Department, 1986, p37.
[120] BBC Broadcasting Research Department, 1981, pxi.
[121] BBC Broadcasting Research Department, 1986, p40.
[122] BBC Broadcasting Research Department, 1986, p42.
[123] BBC Broadcasting Research Department, 1986, p42.
[124] Menneer, 2014.
[125] BBC Broadcasting Research Department, 1993, p17-25.
[126] BBC Broadcasting Research Department, 1984, p11.
[127] BBC Broadcasting Research Department, 1990, p47.
[128] BBC Broadcasting Research Department, 1993, p17-25.
[129] BBC Broadcasting Research Department, 1990, p49-57.
[130] Kent, 2002, p251-252.
[131] Starkey, 2002, p45.
[132] Van Meurs, 2008, slide 31.
[133] BBC Gateway Audience Portal, 2011, North, 2010.

3.1.6 BBC's Current AI Collection Process (Pulse Survey)

The BBC Gateway Audience Portal (2011) summarises the *current* process:

> The survey that provides AI scores is known as the Pulse – it has been running in its current format since 2005 and is administered for the BBC by independent market research agency GfK. Scores are drawn from a panel of around 20,000 people who are representative of the UK population. GfK invites them to fill in an online questionnaire every day and on a typical day we can get up to 6,000 respondents telling us about their appreciation of television and radio programmes.
>
> On radio we cover BBC national and digital stations, along with BBC local and national stations and selected national commercial stations.
> We collect other information in relation to how people rate programmes, such as quality, originality, whether a programme is talked about and how much effort is made to view or listen to it.
>
> We also get comments on audience likes and dislikes.
> We ask people to rate all the programmes they have seen [*sic* – radio is also in this survey] on a ten-point scale – we then scale the score up to be out of 100 (so, an average score of 8.2 becomes an AI of 82).

3.2 The Need for Audience Research

The objectives of audience research have been described as categorisation, explanation, prediction, creating understanding, and providing potential for control and evaluation (Walliman 2005, p234). Programme makers and programmers, in particular, need to be able to evaluate their shows (McLeish, 1999, p293) and, if possible, attempt to understand the elements that contribute to the audience's enjoyment of the output (Perry et al, 1997, p388). A clear need for systematic research at the BBC was identified soon after the corporation was established, and broadcasting research been developed over the past 75 years.[134]

Early audience research was found to be of little use in achieving these aforementioned objectives. It either revealed something of little importance or something that could be inferred merely from common sense: 'little more than statements of the obvious' (Crisell, 1994, p201). When Robert Silvey introduced systematic research to the BBC in the late 1930s (Silvey, 1974), he found that those resistant to it argued that it was only telling them what they already knew. He found, however, that what people 'knew' was hardly consistent, thus validating the use of research as an arbiter of diverging opinions:

> We were often told that the findings of our reaction reports were predictable; indeed they very often were, but the predictions of a producer, the head of his department and the editor of the service which had commissioned the broadcast were not always identical – to put it mildly. In such cases the findings of audience research came into their own as outside and independent confirmation of one of the conflicting predictions. (Silvey, 1974, p119)

The value of audience research has always been a matter of contention within the BBC and even today producers admit to using it only if it already supports their beliefs (Oakley, 2009; Berthoud, 2010). In its early years John Reith, the BBC's Director General, had concerns that media research would result in the tail wagging the dog (Crisell, 1994, p22). However, the desire for measurement of media habits and preferences increased with a growing belief that this information could aid broadcasters to manipulate their

---

[134] See from p66 for expanded history of audience research.

listeners (McQuail, 1997, p14; Douglas, 2004, p124), as well as the growing need for commercial broadcasters to develop research to produce a *currency*, allowing 'an open and transparent basis for conducting business' (RAJAR, 2012a).

In the case of comedy, comedic talent involved in programme making can also have an interest in ratings, over and above the desire for the high ratings that lead to recommissions. Attitudes vary along a spectrum, from valuing audience size to it being of no importance to the creative process. Sean Lock (2011), when asked whether he would rather have a large audience or a smaller more appreciative audience said, '*I think everyone wants a large audience. You want as many people to see what you do as possible*.' David Mitchell (2010) answered the same question by saying that while having a large audience is an aim for comedy, it should not be at the expense of watering-down the content just to appeal to the lowest common denominator:

> *There's a real balance there. One of the skills of being a professional comedian rather than a funny mate of yours in the pub is that you find a way of making good comedy without access to the full frame of reference of your friendship group. No one will ever laugh like they laugh under those circumstances. The skill of the comedian is to find things that have wider resonance. So ultimately, the ultimate laugh with the small audience is worth less to me as a professional… But then there definitely comes a point where you feel so diluted and your remarks are so broadly, blandly, tinily [sic] amusing that you feel there's no point in that either.*

Michael Ian Black[135] (2011) sits at the other end of the spectrum to Lock, explaining that in creating comedy, the most important thing is to be '*fulfilled creatively*' rather than garnering a big audience: '*In a way, I don't give a shit if anybody watches it'*. Unlike for television, there is no overnight programme specific audience size information available to radio broadcasters. As much as this can be creatively freeing for those involved in decision making, creation and production, it can also lead to complacency as there is no detail available to indicate if a programme might be losing listeners (Davie, 2009).

3.2.1 Measurement

On the most basic level, measurement is the 'process of assigning numbers to objects, according to some rule of assignment' (Webster et al, 2006, p126) and these numbers in turn are used 'to represent facts and conventions' about these objects (Stevens, 1946, p680). For radio, in particular, the information that broadcasters want to measure is much the same as it has always been. The following questions were put to listeners as early as 1935 by Stanton (Webster et al, 2006, p93):

1. When do they listen?
2. For how long?
3. To which station?
4. Who is listening (split by, e.g., gender, age, economic and educational level)?
5. What are they doing while they are listening – how much attention are they paying?
6. What do they do as a result of the programme?
7. What are their programme preferences?

For BBC radio today, while points 5 and 6 would only be answerable by customised research, systematic research can cover the other questions. Points 1-4 can be interpreted by ratings measurement (RAJAR diaries)[136]) and point 7 is addressed by the BBC's Pulse research which includes appreciation ratings. While these types of measurement have been in place in some form or other for many years, continued fragmentation of media channels and platforms means that audience measurement is becoming more complex as audience habits change (McQuail, 1997, p48; Holden, 2010).

---

[135] American comedian, writer (*Run, Fatboy, Run*), podcaster, with around 2m Twitter followers.

[136] See from p80.

This changing landscape has affected perceptions of the continuing validity of traditional measurement techniques. As Tim Davie (formerly Director of BBC Audio and Music) asked, in the light of proliferation of platforms, '*how good is a diary in capturing the full extent of audio's impact?*' (Davie, 2009). Systematic measurement, typically trend or panel studies (Webster et al, 2006, p116), allow patterns and changes therein to be identified over time. In 1997, McQuail wrote (p67) that media habits change slowly, but Collins (2010b) argues that technology has accelerated the rate of change in people's media consumption. In this respect snapshot yearly surveys are becoming less useful, so systematic research becomes more valuable. There are trade-offs however, as where systematic research measures variables often, the depth of data being measured is likely to be less (Kent, 1994, p11). For example, should broadcast researchers want to answer all of Stanton's questions for Radio 4 it might be possible to conduct a one-off extensive survey to address these points, but it may not be practical to do this every week. However, every week, people can be asked, through the RAJAR sweep, *some* of the questions. Asking *all* of the questions *all* of the time would require a very large budget.

3.3 Radio Measurement

Approaches to radio audience measurement, reflect the challenges posed by the nature of the medium. Radio listening has such a range of levels of involvement that they are particularly hard to measure, with practical definitions of *listening* or *hearing* or even *audience* being ambiguous and open to different interpretations (Barwise and Ehrenberg, 1988, p131). Kent (2002, p250) and Gunter (2000, p113) citing Twyman (1994) and (Barnard, 2000, p188) identify a number of ways in which radio presents particular problems of recording and measurement, compared to television:

- Radio can command various levels of attention, and can be primary, secondary, or even tertiary (subconscious) entertainment.

- As an often non-primary medium, listeners can be unaware or unable to identify or recall the radio stations to which they are listening

- Much listening is done on the move, with consumption often taking place outside the home and even from devices not belonging to the respondent.

- Radio is highly fragmented, and increasingly so. Not only are there many stations but also many platforms via which to hear audio, for example, through FM, LW or DAB, DTV and online through static or mobile devices.

- Radio is more likely to be heard as a continuous stream as opposed to watching specific, individual programmes on television, making it harder for the listeners to correctly identify what they are hearing.

- There is less money behind radio and thus less to invest in audience research.

While these points apply to radio as general rules, the aspects regarding the 'secondaryness', mobility and fragmentation of the media are increasingly relevant to television too. One might also argue that BBC radio stations are likely to have far higher investment in audience research compared to either a small television station or a YouTube channel that could be attracting huge audiences but has small production budgets with little funding for research.

As so many factors are involved in radio listening, no *one* method provides the perfect measurement system. Kent (1994, p6) lists questionnaires, diaries and electronic recording devices as the main methods. Twyman (1994, p9) adds coincidental interviewing to the list. With development in different types of meters, Webster et al (2006, p140) segment audience research methods into more detailed types: telephone recall, telephone coincidentals, listener diaries, household (passive) meters, peoplemeters, portable peoplemeters,

PC meters, and passive peoplemeters. All of these have various pros and cons, and are more or less suitable for different types of radio consumption and the required data output.

One of the key questions to consider when discussing radio measurement is whether one wishes to measure mere *exposure* to the medium or only incidences where the respondent is *aware* of the audio. Metered measurement tends towards the former, whereas diaries reflect the latter.

> *Listening is different to exposure… What you can remember tends to be what you were listening to rather than what you were exposed to.* (North, 2010)

3.3.1 Defining "The Audience"

The differences between measures of exposure and attention relates to the complex question of how media researchers define an *audience member*. While the word may seem superficially unambiguous, even to the layperson it has had a wide variety of implications (McQuail, 1997, p1-2). An audience member can equally be someone sitting at home alone listening to an 18:30 broadcast of **Just a Minute** or someone who has attended the recording of that show along with hundreds of other people. What criteria can they be used to define an audience member at home? Perhaps they have only listened to one minute of the programme? What if they spent the whole time on the telephone to someone else and didn't hear any of the show but were in the same room as the radio while it was on; are they part of the radio audience? In any case, might it only be useful to measure that person as a listener if they claim to remember that experience?

> You may choose to define it conservatively, confining it to those who have given the broadcast their full attention throughout, you can define it generously, including all within earshot, or indeed you can choose any point along its continuum. But whatever your decision, be assured that it is highly relevant to the question of audience size… One system will be found to have estimated the audience for a certain broadcast to be x, whereas another estimates it to be y, and the layman, naturally enough, throws up his hands in despair. But, if they are using the word "audience" in different senses maybe their answers ought to differ, because they are not measuring the same thing. (Silvey, 1974, p179-180)

To decide on how to denote a radio listener, researchers need to pinpoint two aspects (Kent, 1994, p4):

1. What audience behaviour is classed as listening?
2. How long audiences have to continue this behaviour to register it as a listening event?

Even within the UK, different industry standard broadcasting measurement systems use very different criteria for defining an audience member. For example, for a person to be registered as a viewer on the BARB (television) system, they need to be in the same room as the device while a programme is on for at least 30 seconds, *and* to register their presence on the BARB machine (Sharot, 1994, p68). Whether or not they are paying any attention to the screen is irrelevant to the ratings. For UK radio ratings measured through RAJAR's diary method, respondents will count as an audience member if they *claim* to have heard a particular station for at least 5 minutes in any given 15 minute slot (RAJAR, 2012a).

Having a variety of definitions for media consumption makes comparison between different media difficult or impossible, and this is one of the barriers that need to be circumvented in developing any kind of cross-media measurement (BBC, 2012b, p13). McQuail (1997, p149-150) claims that the varying types of audience are too numerous for one definition to apply. The range of dimensions that can apply to an audience can be influenced too by how the audience is perceived overall: whether they might be considered as 'victim' (effect model), 'consumer' (market place model) or 'coin of exchange' (commodity model) (McQuail, 1997, p87).

While researchers might have varying criteria they wish to employ to define media consumption, their chosen definition needs to be clear and understandable by the respondents if they are involved with

registering their use. For example, BBM 1975 Canadian research showed that there was uncertainty on the part of the respondent as to whether they were listening or not: 'There was a tendency for some respondents not to report listening if they were doing something else at the time, if they were not paying full attention, or if they did not control the tuning decision' (Twyman, 1994, p97).

Thus, the tool of measurement used defines what is accepted as an audience member. Should, for example, radio ratings be measured by meters rather than diaries, our definition of a listener would change (Kent, 1994, p5).

3.4 RAJAR

UK radio ratings are measured using respondent diaries administered by RAJAR (Radio Joint Audience Research), a body jointly owned by the BBC and the RadioCentre on behalf of the commercial sector (RAJAR, 2012). The formation of RAJAR in 1992, was intended to provide an '*industry yardstick*' for radio broadcasting (Barnard, 2000, p90), and replace the BBC's daily Survey and the JICRAR's[137] weekly commercial survey (Kent, 1994, p17; Robinson, 2000, p382) - two 'often competing radio research methodologies which reflected the different priorities of the BBC and commercial radio stations' (Robinson, 2000, p382). This type of 'syndicated' research is intended to avoid duplication and arguments about the relative validity of different measurement systems (Kent, 1994, p16). The current contract to conduct the actual research, in place since 2007, is split between two research agencies: sample design and weighting done by RSMB, and the fieldwork, scanning, processing and reporting done by Ipsos MORI (RAJAR, 2012a).

RAJAR surveys over 110,000 adults every year (ibid), covering 50 weeks (excluding Christmas and New Year) and includes over 300 stations, making it the biggest media survey outside of North America (Winter, 2011, slide 9). The survey is technically a 'sweep', which means that while data is presented as quarterly, respondents only participate for a single week (RAJAR, 2012a). Respondents participate as households but fill in individual diaries. Listening is recorded for all from the age of 10 although published figures tend to be for those of 15+ (ibid). To allow segmentation of the data, the following variables relating to the respondents are available to RAJAR subscribers when analysing the data:

DEMOGRAPHICS

Gender
Age
Social grade
Ethnic origin
Region
Working status
Marital status
Household composition
Employment status
Household tenure

---

[137] Joint Industry Committee for Radio Audience Research.

OTHER MEDIA

Media access
Television viewing habits
Internet use
Mobile devices use
Newspapers reading habits
Cinema attendance

3.4.1 Limitations of RAJAR

As with many industry-standard measurements, the RAJAR process attempts to balance the different desires and limitations of all involved (Miller, 1994, p58). The following, sometimes conflicting, requirements impact upon the nature of RAJAR's output:

- Broadcasters – differing requirements of PSB versus commercial stations.
- Media buyers and sellers – concerned mainly with reach and frequency of advertisements across demographic groups.
- Willingness of respondents to provide information – the burden of diary collection limits participation.
- Technical and financial limitations of the data collection and analysis – sample sizes are limited, as is granularity of data.

In the case of BBC Radio 4 comedy, RAJAR provides audience size data at slot/day level for 15 minute intervals over a quarter.[138] A number of aspects limit the usefulness of this information for the BBC, for example:

- RAJAR figures are from a sweep, and are just estimates of size. Errors of sampling and response could be present. Sample sizes from dayparts[139] with small audiences may be too small to be useful.
- Granularity is limited to quarterly data at (mainly) 15 minute intervals over a quarter, thus does not provide minute by minute overnight data as BARB does.
- Diaries measure claimed listening, not actual listening. Recording, recall and bias errors may occur.
- Recordable data is limited and many aspects of listening are not documented. Such unrecorded aspects include attitudinal information, situation, group or solo listening, mood, other activities, levels of exposure, attention, engagement and appreciation.

Some of the most pertinent aspects have been expanded upon below: their effect on understanding Radio 4 comedy is denoted as R4→:

i Quarterly data

Relatively low samples per week mean that figures are reported on a quarterly basis. This would be prohibitively expensive to change (Winter, 2011) and doing so would need agreement from the whole of the radio industry. Quarterly reporting presents a number of issues: firstly, information is very out of date by the time it is presented to the users. Hard copy diaries, although scanned to save time, still need to be checked and the data weighted and analysed (Starkey, 2002, p51). For example, in RAJAR's figures from Quarter 1 2013, the dates of measurement were 31st December 2012 – 31st March 2013. The data was released to subscribers on Wednesday 15th May (RAJAR), around 6 weeks after the end of the measurement period and over 4 months after some of the data was recorded (MacKenzie, 2000).

---

[138] Overnight, the measurement period is expanded to 30 minute slots.

[139] Daypart is a broadcasting term whereby a day is divided into sections, e.g., drive-time or late night.

R4→ The effect upon ratings due to changes in programming cannot be seen until months after the changes have been made. Thus, for example, when a new comedy starts on Radio 4, programmers are unable to act quickly to respond to the programme's performance.

Quarterly figures mean that ratings for any slot are an 'average' of the figures gathered over a three month period. This means that there is no programme level granularity for the data. Performance for any particular week or particular timeslot in that week is not reported. It is possible, however, to get weekly data from RAJAR but it is based on relatively small sample sizes, provided unweighted and at an additional cost to the subscriber (Collins, 2010a). Audience size at programme level is generally unattainable.

R4→ Within the BBC some hold the view that the absence of overnight figures can be freeing for a broadcaster – '*I definitely think we benefit creatively by not being obsessed on a nightly basis with the drugs of overnight ratings*' (Davie, 2009) – while others can find this limitation '*hopeless*' (Raphael 2012a). For example, Radio 4 comedy series seldom run for more than six weeks at a time, often only four. So, in a three month period the ratings are usually representing an average of around three *different* scheduled series. Hence, the data does not reveal which series was most popular in terms of ratings. Any special single programme or exceptional week will be hidden by the average. Thus, in the case of comedy, RAJAR audience size figures cannot show the Radio 4 Comedy Commissioning Editor how many people listened to a particular comedy programme or even series.

RAJAR diaries are not issued over the Christmas and New Year period because diaries are relatively burdensome for the respondents to complete at this busy time. It also ensures that the relevant quarter is not skewed by exceptional listening, which was the case when it was once trialled (Winter, 2011). But this also means radio broadcasters do not have a picture of this exceptional listening and therefore have no aid in programming for this holiday period.

R4→ On Boxing Day 2000, Radio 4 broadcast Stephen Fry's reading of the whole of the first Harry Potter book, uninterrupted, breaking with the established Radio 4 schedule. Having no diaries over this period meant that the network had no audience size data from the event. Being able to measure ratings for such a programme would have been useful in evaluation of the 'stunt',[140] understanding of the size and make-up of the audience in comparison to who would normally be listening. Why make an effort to reach new audiences if there is no proof of success? For example, Radio 4 often runs comedy festive specials but no ratings figures are available to compare the performance against more typical programming.

ii 15-Minute Data

RAJAR diaries segment the day into quarter-hour periods. The respondents are asked to record listening in any 15 minute period where they have been listening for 5 or more minutes. (Between the hours of 00:00 and 06:00 it is half an hour, as small listening figures during this time mean that this time period is of lesser concern to stakeholders). Assuming that the respondent makes no error, this means that any listening of 4 minutes or under will register as 0, while 5 or more counts as 15. Although the underrepresentation scenario may seem unimportant superficially, examples like RDS interrupting another station with traffic information during a long car journey would result in the 60 second bursts never being registered despite the fact that their cumulative total could be more than 15 minutes (Starkey, 2002, p60). Overall, there is an assumption that some listening is understated, as in the RDS example, and some is overstated; listening to a station between 08:40 and 08:50 would mean that the two 5 minute slots fall across two 15-minute periods, so would

---

[140] 'Stunting' is a broadcasting term used for 'block scheduling' that Webster et al (2006, p60) claim can bring in high ratings for broadcasters.

be measured as listening from 08:30 to 09:00. Starkey (2002, p60) argues that the overs and unders may not necessarily cancel each other out.[141]

In a study undertaken by the BBC in 1988, different ways of registering 15 minute listening were evaluated (Twyman, 1994, p96). Diaries were completed by making a quarter hour register in one of two ways: if the respondent had deemed that they listened to 'more than half' (i.e., 8 minutes or more) versus 'any listening' in a quarter-hour. As one would expect, 'any listening' gave higher total listening hours than 'more than half' but the results showed that 'the criteria made little difference'.

R4→ Even if it is accepted that listening for 5 minutes within a 15 minute period is accurate enough to give figures for hourly listening, it must be recognised that it gives us no minute by minute granularity. BARB meters measure second by second and report minute by minute data. If this level of data was available for RAJAR listening, it could allow programmers to have a greater understanding of how audiences come in and out of programmes. For example, it could be possible to take a Radio 4 comedy programme with a low audience figure and see if it was low for the whole of the programme or whether it started off with a high figure which dropped as people switched off or over. Broadcast researchers could make very different respective assumptions about the appreciation for the programme if this data was available - based on the theory that programme appreciation correlates with audience size.[142] Knowing the proportion of the programme that was listened to could be an indicator of appreciation. Also, there might be the possibility of ascertaining whether certain elements of a show were more or less popular if it were possible to see minute by minute when listeners come in and out of shows. For example, programmers might see that there was a mass switch-off at a certain point in a show and this might prompt them to investigate which element caused this, as is possible to do with television data.

iii Exposure, Engagement and Appreciation

Methods such as metering measure mere exposure to media, but diaries go some way to measuring engagement since the respondent can only record an incident in the diary if they are aware of hearing it. Some programmers may think that it is only worth recording incidents that register in the mind of the listener, as diaries do (North, 2010, claims that Jenny Abramsky, former Director of BBC Audio and Music, was of this view). The alternative view is that for radio listening, all levels of attentions – including subconscious – should be quantified. Many advertisers, for example, would wish to know if their messages are being heard, even subliminally (Starkey, 2002, p63).

R4→ Measuring exposure as opposed to conscious awareness of content may reveal listening patterns that are alternative to those currently measured via the diary method. While commercial stations and advertisers may be interested in exposure, as this could be seen to be an aspect of industry currency, PSB might be less so if engagement and value for money are more important criteria. However, subconscious listening might still go on to affect listener perceptions. It is possible that a listener might 'hear' a comic even if they are not listening to a show, and that exposure engenders a subconscious familiarity with that performer – familiarity perhaps being a factor of appreciation?[143]

Diaries record listening of which the respondents are cognisant but make no differentiation between *levels* of attention paid. Even if listeners are consciously aware of the audio, they could be listening to it as their primary activity or merely as background noise. RAJAR diaries do not measure appreciation or any

---

[141] Starkey does not indicate whether the aggregate discrepancy would result in over- or under-reporting of listening hours.

[142] See from p91.

[143] See from p98.

other kind of attitudinal metric. Though it is unusual to include appreciation on a ratings measurement system (Kent, 1994, p16), it is possible and has been done in the Netherlands, for example, (North, 2010).

R4→ Understanding levels of engagement is important for Radio 4 as there are sub-genres of comedy content[144] that may be more suitable for lower attention and vice versa.[145]

3.4.2 Ratings

Despite their limitations (McLean, 2009), ratings are the common currency of the industry and determine the success of a media organisation and its programmes (Webster et al, 2006, p11). Their use by commercial broadcasters as a currency to evaluate media prices (see Webster, 2001, p923) means they will continue to be important in media evaluation (Webster et al, 2006, p110).

However, in linear broadcasting a programme's ratings are significantly a function of its scheduling (in addition to its content, for example). Whatever the appeal of its content, if a radio programme is aired when people are not available to listen, it will not get a big audience (Menneer, 1987, p244; Gunter and Wober; 1992, p70, McQuail, 1997, p114). With on-demand or non-linear broadcasting, such as iPlayer catch-up, the scheduling of a programme is arguably having a lesser impact upon its audience size, since anyone can listen at any time. However, as Shingler and Wieringa posit (1998, p106). some genres of radio content are unlikely to be popular beyond the immediate broadcast as the appeal of radio is partly due to their *liveness* and thus their audience size is likely to remain driven by their scheduling. For example, audiences may not choose to listen to news programmes after transmission as they will be out of date and irrelevant. However it may be hypothesised that most radio comedy programmes, even topical shows, can still be enjoyed after their original transmission.

Ratings can be used by programmers to aid decisions about scheduling and editorial content (Barnes and Thomson, 1994, p78; Webster et al, 2006, Pxvii). However, while it may be important that there is a large audience watching or listening to the programmes, BARB and RAJAR do nothing to inform them about the quality of the audience's experience; both systems are basically measuring exposure and tell broadcaster researchers nothing about attention, engagement or appreciation (Webster et al, 2006, p175). Ratings alone are not always enough to explain everything about the audience's experience (McQuail, 1997, p54).

> Knowing the size of a programme's audience told one nothing about the nature of that audience's listening experience, what it was about the programme that they had liked or not liked or why they felt about it as they did. (Silvey, 1974, p113)

3.5 Summary of BBC Radio Measurement

- Whilst the current RAJAR system provides the industry standard currency, it has limitations such as 15 minute granularity. Thus it does not allow programme level analysis and programmers have to rely on other sources of programme evaluation.

- Audience size measurement is variable depending on the chosen definition of 'the audience', with radio comparatively more difficult to measure than television.

- Audience size does not give information about the attitudes of the listeners.

---

[144] See from p39.

[145] See from p102.

Chapter 04 – BBC AIs and Programme Appreciation

KENNETH WILLIAMS:	*I always listen to your radio programme.*
TONY HANCOCK:	*Really?*
KEN:	*Oh yes, Rosita and myself, we never miss it. I heard last week's. I wasn't so keen on it as some of the others.*
TONY:	*Weren't you? No?*
KEN:	*No. I didn't think it was as funny. I think Ted Ray had the edge on you last week. It must be very worrying for you trying to keep it up week after week, especially with so many good newcomers arriving on the scene.*
TONY:	*(Getting annoyed) Yes, quite.*
KEN:	*And the two lads at the office, they don't like you so much these days either.*
TONY:	*They don't?*
KEN:	*No, no, still they're very fickle. Rosita and I think you were at your peak five years ago. You were very funny in those days. Still I don't suppose you're bothered, eh? You've made your pile I expect?*
TONY:	*Oh, yes, I'm rolling in it. No need to work again for a fortnight.*

Hancock's Half Hour – 'Sunday Afternoon at Home'

After introducing the mechanics of the BBC's programme appreciation measurement process in the previous chapter, this thesis goes on to consider its usefulness to broadcasters, both at station and programme level. Appreciation is then considered in terms of the factors that can affect it, over and above simply the 'quality' of a radio comedy programme. Finally, the limitations of AIs are discussed.

4.1 BBC's Appreciation Index (AI)

Ratings are the common industry currency for broadcasters: for Radio, quantified by RAJAR[146]. These figures indicate how many people are listening but not whether they enjoyed the programme (McQuail, 1997, p54). The BBC measures appreciation separately as part of its ongoing Pulse survey whereas commercial radio does not, so understanding of appreciation is not available for the industry as a whole. Appreciation assessment is a measure that has been consistently employed at the BBC for around 70 years and in this respect it is the sole measure of objective[147] performance at programme level for BBC Radio.[148]

4.1.1 AI Measurement

Currently BBC radio AIs are measured through the BBC Pulse survey, separate to the industry-wide RAJAR diary system. While it is possible to collect appreciation and other attitudinal information within the same survey as ratings data (this has been the case, for example, in the Netherlands since 1987 [Ang, 1991, p144; Mytton, 1999, p124]), it is not common. Partly this is because the task is seen to be onerous for the respondent (Danaher and Lawrie, 1998, p56) but also there is a risk that in evaluating the programmes the respondent could change their behaviour:

> …such a process may well influence the consumer behaviour being recorded. Thus if a respondent indicates certain negative views concerning a programme, then he or she may well be tempted to swap channels next time it is broadcast in order to appear more consistent. (Kent, 1994, p7)

---

[146] See from p80.

[147] See p88 for Mark Thompson's use of AIs as an 'objective' measure, as an example.

[148] See from p66 for an overview of the BBC's development of research.

The BBC's Pulse survey is administered by GfK, with its data then processed and supplied to the BBC by TRP. The survey responses are drawn from a panel of around 20,000 people, representative of the UK population (BBC Gateway Audience Portal, 2011).[149] Respondents are asked to highlight all the programmes of which they have seen or heard at least 5 minutes on a given day. However, they are not expected to respond every day (they are expected to respond at least 10 day in each month) so a typical day will result in around 6,000 people filling-in the online survey, covering both television and radio for BBC national and local stations. The appreciation question is worded exactly as follows (BBC Audience Research, 2009b):

> **In the list below are all the radio programmes that you listened to yesterday. Could you please rate each of these programmes with a mark out of 10, where 10 is the highest score?**

The AI is the mean of the scores given from the 10-point scale running 1-10 (then multiplied by 10 to differentiate it from individual scores) for individual programmes by those rating the show. For example, if a programme has a mean appreciation score of 8.3, its AI would be expressed as 83.

Respondents are also asked further questions (and given a choice of answers) about how they were listening to the programme, including:

> **And how much effort did you make to listen to each of these programmes?**
>
> > **I made an extra special effort to listen to the programme.**
> > **I usually listen to this show as part of my usual routine.**
> > **I did not make any particular effort to listen to the show.**

> **And how much of each programme did you listen to yesterday? Please mark on the scale below where 1 means listened to hardly any and 10 means you listened to all of it.**
>
> > **1 – Listened to hardly any, to…**
> > **10 – Listened to all of it**

> **Where did you spend the most time listening to each of these programmes?**
>
> > **At home**
> > **At work**
> > **In the car**
> > **Elsewhere**

Further optional questions are asked about how the listener felt about the programme, most of which provide Likert-style[150] responses (agree strongly, agree slightly, no strong views, disagree slightly, disagree strongly), including:

> **This was a high quality programme.**
>
> **It's the kind of programme I would talk to other people about.**
>
> **This programme felt original and different from most other radio programmes I've listened to.**
>
> **Was there anything in the programme that you personally found offensive? (Y/N)**

There are also two questions that are open-ended requiring a verbatim response:

> **What did you think about this radio programme? Please write in what, if anything, you LIKED about it?**
>
> **And what, if anything, did you DISLIKE about it?**

---

[149] See p76 for the exact wording used by the BBC to describe the AI collection process - The BBC Gateway Audience Portal (2011).

[150] Likert style scales generally offer respondents a discrete number of answers relating to levels of agreement to a statement.

Demographic information is also collected about each respondent to allow further analysis,[151] including:

Gender
Age
Class
Marital status
Sexuality
Children 15 or under
Main Earner
Nationality
Ethnicity
Country
BBC Region by post code

This kind of evaluation research by survey is categorised by Gunter (2000, p142, p147) as 'off-line', as it is attempting to measure the respondents' reactions to a programme away from the actual experience as opposed to 'on-line', which would be using measures taken at the time of, or immediately after, the exposure.

4.2 How AIs are Used

Ratings tell broadcasters how many people are consuming their media and are the accepted common currency within the industry. However, ratings tell us nothing about how much people 'like' a programme. Appreciation is 'sometimes considered more valid and informative than ratings' as ratings merely, 'are often a predictable function of scheduling decisions, determined by timing and the available alternatives', and thus tell us little about the quality of the programme (McQuail, 1997, p58). For radio programmes in particular, RAJAR ratings are very limited in respect of individual programme performance.[152] Without the luxury of programme-level granularity, evaluation relies on a variety of measures, such as reviews and word of mouth.

Menneer (BBC Broadcasting Research Department, 1985, p41) explains the main ways in which appreciation ratings − albeit in the realm of television rather than radio − are of value to the BBC. They provide:

- A performance criterion for niche programmes where audience size may not be the primary concern.
- A 'morale booster' for programmes not attracting large audiences due to factors outside of the content of the programme itself, such as competition or scheduling.
- A possible means of predicting future audience size for following episodes.
- An aid in the decision making process for recommissioning.
- A qualitative system of judging programmes that can contribute to measurement of the value of public service broadcasting.

He goes on to argue that most routinely they are used as diagnostic tools, informing and suggesting corrective action for the audience size of programmes (Menneer, 1987, p257-258). For example, in the early days of *EastEnders*, static audiences, which otherwise would have caused concern, were mitigated by growing AI scores and justified in the argument that:

> It was merely a matter of time before its following would build up. Had the programme team not had access to AI data during these crucial summer months, their confidence in the enterprise could well have been shaken. (ibid, p261-262)

Had there been no appreciation research, one of the BBC's most successful shows of the past thirty years might have been at risk of cancellation.

---

[151] The full list of variables held on the respondents can be found in the appendix p288.
[152] see p81.

4.2.1 AI Use – Corporate and Regulatory Measure of Quality

The BBC's 2010 strategy 'Putting Quality First' established the initiative of publishing quarterly audience research information:

> The BBC Trust pledged to set new standards of openness and transparency for the BBC, so that the public and the market understands how the corporation spends its money, how it is performing and what it plans to do next. (BBC, 2011a, p2)

For radio stations, published performance data gives audience size from RAJAR (Average weekly reach % and 000s of hours listened per station), but also includes two metrics presented as the 'BBC Radio Quality Measures'. One is an aggregate score for whether listeners found the station's programmes 'original and different', but more prominence is given to the AI as a measure of quality, collected at respondent/programme level and aggregated up to station level BBC, 2011a, p15).[153]

AIs have also been used in another way to give a total 'quality' score for a channel by measuring the percentage of programmes that get a score over a certain threshold. For example, Day (2008, p11) gives the example of 'AIs % programmes scoring 80+' as a measure of BBC programme reporting. Collins (2010a) claimed that this measure is indeed used on an ad hoc basis. It is, however, not a figure that is quoted currently on the BBC's published audience information programme performance: currently the mean aggregate AI has been selected as the appropriate measure.

AI scores can also be used strategically to deflect criticism, with a recent example seeing the BBC coming under a barrage of high profile protests about its coverage of the Queen's Diamond Anniversary. While Mark Thompson (BBC Director-General at the time) admitted that the programmes had not been perfect, he quoted the AI of the programme – which was 82 – as an indicator that the programme was not poor quality, saying: 'With all television and radio programmes, everyone is entitled to their own opinion, but this was a programme which all of the *objective* evidence is that the public thought it deserved 8 out of 10' (Neilan, 2012).

Measurement of appreciation is of particular importance in the context of financial cutbacks; the BBC's 2010 'Putting Quality First' money-saving initiative clearly indicated in its very name that any savings made should not have any impact upon the perceived quality of the content or indeed that of the BBC as a whole.

> The actions that we have taken during the last two years have met the short-term pressures of needing to fund the pensions deficit and lower than planned income from the licence fee whilst sustaining and improving the breadth, depth and reach of the services that our viewers, listeners and readers value. This has, I believe, created a strong platform from which we can begin to meet the future challenge of maintaining the *quality* of our output within our future funding pressures. (BBC, 2011b – pF3)

Comparing appreciation before and after cutbacks allows the BBC to understand if there has been any measurable detrimental effect for the audience.

4.2.2 AI Use – Radio Programming and Production Measure of Appreciation

'The relevance to broadcasters is obvious for programme planning and scheduling' (Twyman, 1994, p102): seen as 'supplementary' (Kent, 1994, p16) data for programmers in their decision-making processes (Gunter and Wober, 1992, p4). Those involved in commissioning decisions use AIs particularly when it comes to looking at recommissions. For the Radio 4 comedy Commissioning Editor, AIs are the *primary* indicator of performance and important in deciding what goes onto Radio 4: '*Most of my business is returning business, and knowing when to recommission or when to kill*' (Raphael, 2012a). With no overnight ratings figures from

---

[153] See p260.

RAJAR[154] and relatively little press coverage for radio (Reynolds, 2011), AIs garner particularly noteworthy significance for commissioning:

> *We always look at them but on radio they're very often statistically small*. (Raphael, 2011)

> *I'll very often do it week by week to see trends as well as the final number.* (Raphael, 2012a)

Raphael uses the AI figures in commissioning round meetings and includes them in her personal notes about each series. She uses the AI score and the number of responses as a quality figure and an indicator of audience size respectively, allowing comparisons between programme performances. Examples of her personal notes from the 2012 to 2014 commissioning rounds indicate the direct effect that the AI can have upon the decision making process:[155]

> This really was not a very distinguished series although AIs not bad. Right to say need more time…

> I want to drop this. Time to stop. Getting AI sample size of 20 and under and in one week did not pick up even that.

> AI average 80. But, not picking up individual AIs. Suggest very few people really like this. But is just not funny anymore. Offer to Jeremy as drama series if that keen on it. Find structure tiresome.

But the AIs are not always the ultimate arbiter. In autumn 2012 a late night programme was recommissioned for a second series after receiving an average AI of just 57 for its first series, whereas a third series of an 11:30 comedy was rejected despite its second series achieving an average of 80.[156]

Regarding scheduling, AIs may sometimes be considered when it comes to the transmission slot. Although particular types (sub-genres) of comedy may fit best at certain times of the day, some programmes could editorially fit in any slot. Good AIs for a programme in a smaller slot could result in it being considered for TX at a time with a larger audience. For example, **My Teenage Diary** first went out in the 23:02 slot. It was very well received, with an AI of 72, and was moved to the higher profile 18:30 time[157] for its second series, where it received an AI of 77. A good AI is also a consideration when choosing which programmes are to be repeated. For radio comedy programmes, only 53%[158] have space in the schedule to be repeated, so it makes sense to prioritise those which audiences seemed to have enjoyed most. For example, **My Teenage Diary** was chosen for repeat in spring 2012 in the Sunday 19:15 slot. However, AIs are not the only criterion for choice of repeats as many factors are involved in comedy scheduling.

Production teams can use AIs to inform editorial decisions and support pitches. Mayhew-Archer (2011) found that, as a radio comedy producer, the only audience research that he had access to were the AIs. However, AIs tend not to be discussed within the department unless they are found to be particularly high or low (Canny, 2011), and producers are not required to seek out appreciation scores. For example, Dare (2012) claimed that while he did see AIs when he was an in-house producer, it was not on a regular basis.

AIs tend to be utilised only when they are supportive of preconsidered arguments: used as illustration to support theses or ignored if they appear undermining. Oakley (2009), writing of BBC television documentaries in which he was involved in the 1990s, explained that although the general

---

[154] See from p81.

[155] It would not be politic to include the names of each of the shows that the feedback relates to as this is taken from personal notes.

[156] See from p59 examples of the range of aspects that are considered for radio comedy commissioning.

[157] See p35 for comparison of Radio 4 comedy slot audiences.

[158] 2014/15 forecast figure.

consensus was that AI scores were thought unsophisticated, it did not stop him from quoting the good ones when pitching the programmes for which a recommission was sought. This kind of attitude toward AIs is also seen in radio comedy, where they are used selectively and considered '*not a science*' (Canny, 2011). Even when research into trends in responses to **The Now Show** supported expectations, Canny did not consider the findings to be robust but rather indicated a distrust of the figures:

> *We did a study across AIs. The lowest scoring AIs were the shows that Marcus, Mitch, Jon, Laura, Steve, Pete were in in a run, and the best AIs were in the shows that had more variety and had newer voices in. I wouldn't leap at those AIs – they're very unreliable – but it's interesting that they reflect my view of it, so I use them accordingly… If you AI at 48 you would be an idiot to just head into another series. If you AI at 90 you would be a madman to drop it or not do it again.* (ibid)

Senior decision makers in BBC radio comedy appear to value their own judgement above the appreciation scores of the listeners:

> *Instinctively I would use it if it helped and I would ignore it if not… It can prove whatever you want it to prove, really. Most people, if it's advantageous to them they'll use it, if it isn't they'll ignore it.* (Mayhew-Archer, 2011)

> *You take it as a guide but nothing very much else.* (Canny, 2011)

> *I don't think, as arrogant as this may sound, that an audience should materially influence or change the essence of a show. If the audience don't come to it and the critics don't like it then you won't do another one, but I don't think that you should be scientifically looking at all the audience research and applying that to a complete rethink of a show.* Schlesinger, 2011)

This view of AIs being of low importance is typical for comedy in particular. Being such a subjective genre,[159] it is seen by many BBC staff to be best judged on expert gut reactions. This has been the case since the introduction of AIs, as BBC execs generally have always placed their own instincts above attitudinal research, considering it intrinsically part of their role to know what's best:

> Personally I paid less attention to the appreciation index than to the listening figures, partly because there is more room for distortion when people are asked to tell you what other people think and partly because it was after all our job to know something about the technical merits and defects of individual items, whereas the figures could tell us when people listened and what they actually listened to. (Gorham, 1948, p166)

However, AIs can sometimes be seen as useful gauges of success, even by creative types:

> *That's another thing, AIs. That's nice, to get good AIs.* (Dare, 2012)

> *The BBC used to have appreciation indexes, the AI. We always thought that the AI was the most important figure. It wasn't as much how many were listening as what they thought of it. And our AIs on **Steptoe** and **Hancock** were very high.* (Simpson – Galton and Simpson, 2012)

Even when AIs are accepted as being good indicators of a programme's performance, a poor score for a programme that is otherwise perceived to be good tends to get rationalised within a defensive position. For example, where a new show is liked and supported within the department, a low AI score is explained away by the argument that new comedy always suffers from low appreciation. If the score is high, that is put forward as being proof that the show is good: '*It's true though, everybody does it*' (Berthoud, 2010).

---

[159] See from p48.

The other thing is recognising that something's rather good but it's commissioned for the wrong slot. There was one very good piece which Paul did with Johnny Sweet that to my mind is not a six thirty show. It doesn't have an audience. It's actually a really clever conceit but it's quite difficult to follow and I don't think it's a six thirty. So, when the AIs weren't very good you go: 'Well, I'm not surprised. Let's put it somewhere else next time'.
(Raphael, 2012a)

However, AIs can give results that have driven decisions that could retrospectively be considered to be wrong. Galton and Simpson (2011a, p229) wanted to write for Frankie Howerd at the end of the 1950s, but Tom Sloan, the Head Of Light Entertainment, brought in 'bloody great ledgers' of audience research that supposedly showed that Howerd's career was over. In truth, Howerd continued to be successful and ultimately a cult figure over the next thirty years. Howard talks of how AIs for *Variety Bandbox*, a decade earlier in 1946, influenced his career:

> 'The first three or four months were disastrous', he admits. 'I wrote my own scripts and the audience appreciation figures were very poor. The BBC said that unless I bucked my ideas up in some way I would have to go. I thought it was the end of my career. I sat down and thought, "Well, there's obviously something wrong, so what is it – the script or me?" And I thought, "It's you. You are being too visual. You are doing what you did on music-hall, you are forgetting that people can't see you. Your timing, your tricks are geared to a visual audience." So I started thinking in terms of vocal tricks only and forgot face-pulling. I also quickly tightened up the scripts. My appreciation figures went up very quickly. (Nathan, 1971, p193)

For comedy, predictions of success would surely be best ascertained by utilising both instinct and research alongside each other.

4.2.3 AI Use - Potential as a Predictor of Audience Size

i Appreciation Level – Relation to Audience Size

While BBC television has BARB overnight ratings *and* appreciation scores to indicate programme performance, BBC Radio is limited to AIs for systematic research at programme level as ratings are only published on a quarterly basis. But is there information about the size of the audience that can be inferred from the AI responses? If people watch what they like and like what they watch, surely the higher the appreciation and ratings (McQuail, 1997, p58)? On a simple level, if this were true, it could mean that it is possible to get an indication of which Radio 4 comedies are comparatively more 'popular' (more listeners) than others, giving a surrogate measure for ratings. Coelho and Esteves (2007, p315-316) claim that with an inability to quantify behaviour, attitudes can be utilised as predictors of behaviour. In terms of available research there is limited data, partly because the figures are not made public and hence it is an area 'largely ignored other than by the Ehrenberg team from the London Business School' (Meenner, 1987, p245).

Whether there is a relationship between appreciation and audience size has been a consideration since the two variables have been measured (Danaher and Lawrie, 1998, p60). Robert Silvey, the developer of BBC audience research from the late 1930s, considering this very point concluded that there was no reason to suggest that the two measures had any direct relationship:

> If the size of a broadcast's audience were predictive of that audience's reactions on hearing it, or vice versa – then there would be no point in measuring both. But it was inconceivable that this could be so. Everyone knew there could be cases of small theatre audiences being delighted with what they saw and large audiences being indifferent or downright disappointed, and the same must be true of broadcasting. (Silvey, 1974, p114)

This view is by no means outmoded. Some later researchers, as discussed by Carrie (1997, p30), have indeed failed to find a relationship between the two measures, arguing that there is no reason for the relationship to exist seeing as they measure two different things. Instead, scheduling (audience availability) and competitive

offerings are far more predictive of ratings than appreciation (Sharot, 1994, p83-84, Menneer, 1987, p244-245). Danaher and Lawrie (1998, p60) reference Windle and Landy (1996) and Menneer (1987, p241) as studies that found '[close] to no relationship between appreciation score and audience size'.

While Carrie (1997) acknowledges lack of evidence in past research, he maintained that this earlier work was not perhaps scrutinised at a low enough level. His analysis of UK ratings and appreciation *did* find that there was a relationship between the measures, albeit with caveats:

> There are systematic relationships between audience appreciation ratings and more traditional television ratings of audience size. When allowances are made for scheduling factors, programme type effects, and variations in audience composition, the size of a programme's audience is positively related to the appreciation viewers have for it. (Carrie, 1997, p3)

Barwise et al (1979, p269) had previously found that correlation did depend on programme genre, the discovery made when segmenting their shows between 'information' and 'entertainment'. Further to this, not only does the correlation exist within genres, it has been found to vary in strength depending on genre. Both Gunter and Wober (1992, p14) and Carrie (1997, p32) discuss studies which find entertainment programmes (an umbrella genre that includes comedy) to have the strongest relationship between ratings and appreciation, as opposed to factual programmes or shows requiring high levels of attention.

> Menneer found a much larger positive correlation between appreciation score and audience size for "entertainment" programmes than for "demanding" programmes (Menneer, 1987). In other words, as audience size varies for entertainment or non-demanding fare, so too does audience appreciation. For demanding programmes, however, appreciation level may remain high regardless of audience size. (Gunter and Wober, 1992, p14)

Carrie (1997, p32) found that there was a strong relationship between AI mean scores and television viewing frequencies: 'Programmes which achieve higher average appreciation scores also have higher levels of repeat viewing'. Not surprising, as one might expect someone who likes a programme to watch it more frequently than someone who appreciates it less.

The relationship works both ways as 'liking' can be shown through consumption levels. Danaher and Lawrie (1998, p54, p55, p63) propose that television programme appreciation can be measured through quantifiable behavioural measures rather than a 'claimed' enjoyment figure, in particular the proportion of programme watched and percentage of viewers watching at least 80% of the programme. They propose that appreciation can be shown with the amount of a programme which is seen, on the assumption that someone more committed to a show is more likely to be enjoying it. They admit that there are limitations to this kind of research, such as measurement of watching in BARB measurement systems being defined only by measuring whether the person is in the same room as a television that is switched on.

Ultimately, there is a complex relationship between appreciation and ratings. For AI to be considered as a useful tool to predict audience size there are many factors which would have to be quantified. 'The relationship [between size and appreciation] is not a simple one' (Gunter and Wober, 1992, p14). A further problem is that, whereas in television the AI can be compared to the television ratings, in radio that cannot easily be done. So at programme level it would be, at best, difficult and costly to validate and, at worst, impossible to analyse and compare the output from two very different measurement systems.

If the relationship between appreciation ratings and audience size could be fully understood and quantified, then one could argue that measuring both of the two metrics would be pointless: 'Indeed there would be no purpose in having AI data if they correlated perfectly with audience size' (Menneer, 1987, p245). There may be a correlation between audience size and how people subjectively rate the content, but

the relationship is complex, making audience appreciation ratings difficult to use as a tool to predict audience size.[160]

ii Appreciation Rating Response Rate – Relation to Audience Size

Despite limitations on estimating audience size using appreciation evaluations, there is another, potentially much more straightforward way of using the AI data to indicate the number of listeners. Analysis done in television found a correlation between the *number* of responses on the BBC's Pulse panel per programme and the BARB ratings figures:

Figure 12 – Correlation between number of appreciation responses and audience size – television (Van Meurs, 2008, slide 32)
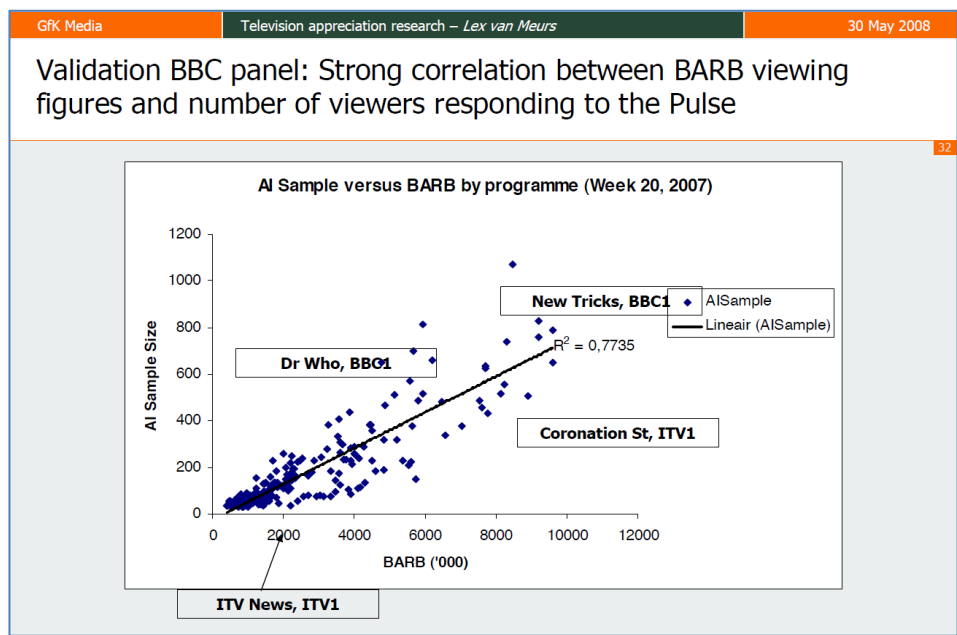
The more people who watch a programme, the greater the number of Pulse responses. This means that the *number of respondents*, regardless of the appreciation scores that they give, can be seen as *proxy* ratings measurement. It is possible to theorise that the same phenomenon would be applicable for BBC radio.

Before the introduction of RAJAR Radio audience size used to be measured on a daily basis. While this level of detail was quantified, it would have been possible to compare the audience size estimates to the number of appreciation responses, to see if they correlated as is currently possible with television viewing. This was mooted by the BBC research team at the time (BBC Broadcasting Research Department, 1993, p23):

> Although the Listening Panel is not designed to measure audience size, the response to an individual programme on the Panel would be expected to be proportional to the number of listeners to that programme as estimated by the Daily Survey.

Today, this would provide daily programme size data which is not available under RAJAR's current quarterly publication system (North, 2010):

---

[160] See from p91.

We [GfK] *can provide you* [the BBC] *with audience size estimates on a daily basis for radio shows because we get 4000 people on radio Pulse... you can say, well 10% of the people doing the survey yesterday said they listened to a programme and we use that for television as a validation. We look at BARB and x number of people watched* EastEnders *yesterday according to BARB. We see a similar sort of proportion of people saying they watched* EastEnders *and giving us an opinion about it. So you'd expect the audience size to be roughly proportional to its audience size as measured by the official measure. And that tends to be the case, with a very tight correlation between the two. With radio you would expect it to be much the same. So 10% of people say they listened to last night's episode of* The Archers*, you'd expect that 10% of the audience was probably listening to* The Archers*. So, you do have on a daily basis an idea of audience size.*

The following paragraph indicates how an audience size estimate for Radio 4 comedies could be calculated:

RAJAR provides quarterly listening figures. For example, a weekly slot (which may have run more than one comedy series throughout the time period) might have an average of 1,000,000 listeners (L). This can be compared to how many Pulse responses the slot has received. For example, for the comparable quarter – say 13 weeks (w), the slot may have received 1,300 responses (n). We can then work out the mean number of responses for each episode (e),[161] in this case 100. The audience size can then be divided by the number of responses to give us the number of listeners represented by any one response (r), in this case 10,000 listeners. Thus, if we want to know the size of the audience, we take the number of responses and multiply it by r. So, if an individual programme had 125 (n2) responses, we would estimate its audience size to be 1,250,000 (L2):

$n \div w = e$	$L \div e = r$	$n2 \times r = L2$
$1,300 \div 13 = 100$	$1,000,000 \div 100 = 10,000$	$125 \times 10,000 = 1,250,000$

So why is this data not used? For a number of reasons:

- Unlike with television, the radio AI response rate data cannot be validated against a standard industry currency such as BARB overnight audience size estimates. At best, it might be possible to validate it against quarterly RAJAR figures. With relatively low sample sizes and limited validity, at best they might be usable only as comparative figures rather than absolutes.

- Establishing the usual response rate may be complex as there could be factors which affect the rate, for example, Van Meurs (2008, slide 28) found that Pulse response tended to be lower at weekends. There could be many more aspects, such as differences in demographics, dayparts or genres, which would need to be established and quantified to eliminate consumption and underreporting bias.[162]

- It could be the case that the respondent's reaction to the programme made them to be more likely to give a response. For example, as panel members are only expected to respond 10 days per month, they might be more likely to make the effort to go online and register their feelings if they have heard something that they particularly liked or disliked that day. Were this the case, as with the previous point, it is a factor that could skew the figures.

- RAJAR is an established standard industry currency and the use of AI responses solely by the BBC as a proxy a ratings measurement could challenge the status quo: '*you obviously don't want to undermine the radio audience measurement currency. RAJARs are important numbers*' (North, 2010).

---

[161] Assuming one episode per week.

[162] See from p151.

4.2.4 AI Use - by Advertisers

Although not currently directly relevant to the BBC as a public service broadcaster with a licence fee to fund it, commercial radio stations rely mainly on advertisements for their income so are of interest to the majority of broadcasters. In turn, advertisers have a vested interest in the appreciation of a programme as it is within this context that their product is placed (Kent, 1994, p16): 'Advertisers' interest is based on the belief that favourable programme attitudes may beneficially affect advertising reception' (Twyman, 1994, p102).[163] Some studies have shown that the enjoyment of a broadcast can influence the effectiveness of an advertisement placed adjacent to it, although the effect might be positive or negative. Danaher and Lawrie (1998, p54) discuss the variety of findings in this area, for example, one study showed that when people saw a more 'enjoyable' television programme the accompanying advertising was 70% more effective, possibly due to the 'Halo Effect' or 'Excitation Transfer Theory' (Cantor et al, 1974). However, the opposite can be found in research where the theory of 'Distraction' has been applied (Perry, 1998), showing well-received programmes can cause detrimental effects upon the advertisements. Danaher and Lawrie (1998, p54) offer a compromise solution, referencing Tavassoli et al (1995), proposing a 'U-shaped' relationship, i.e., programmes that people really enjoy and are immersive, *and* ones that are really disliked, are both unsuitable for advertising. Programmes with moderate levels of appreciation are ideal for the advertisement's message to be conveyed and retained by the viewer or listener. (Appreciation is, of course, just one programme element involved with ad effectiveness.)

While programmes can affect advertisements, ads can in turn influence the audience's perception of a broadcast. Benson and Perry (2006) found that using humorous radio advertisements within a popular breakfast programme resulted in improved perception of the show overall, and increased claimed intent to listen.

Carrie (1997, p19) explains that in the early 1990s BARB audience appreciation measurement was being mooted as a supplementary currency to add to audience size figures. However, this never materialised, possibly either due to the complex nature of the understanding of appreciation or because it was decided not to challenge the well-established currency of audience size (ratings).

4.3 Factors Affecting Appreciation

Appreciation is generally seen as 'difficult to measure' (Kent, 1994, p16) and the UK is one of only a few countries to have regularly collected this kind of data for an appreciable amount of time (Carrie, 1997, p18). In the US, for example, Television Audience Assessment Inc. (TAA) was set up specifically to measure audience reactions based on the 'appeal' and the 'impact' of programmes, measured on a five-point scale (Danaher and Lawrie, 1998, p55). The company, set up in 1980, closed in 1985 due to lack of financial backing (ibid). Currently the BBC asks a sample of respondents to retrospectively rate programmes (that they that they have seen or heard as part of their normal listening) on a 10-point scale and aggregate the mean scores to give a programme total. While the mathematics of calculating a mean is straightforward and easy to understand, the interpretation of the relative values is less clear-cut as there are many foibles affecting its evaluation, challenging its usefulness and validity (Kent, 1994, p16). Tourangeau and Rasinski (1988, p299) propose that when considering answers to any attitude question respondents have to go through a four-stage process. Firstly, they have to interpret the question that they are given, next they retrieve their relevant beliefs

---

[163] Expanded upon from p103.

and feelings regarding the subject, which they then apply to the question and then they finally select a response. All of these stages allow for subjective interpretations.

Considering the many factors that can affect each of these stages, is it possible that appreciation scores can be valid at all, particularly for comedy as such a potentially polarising genre?[164] Can appreciation ratings provide generalised views about which programmes people like and dislike? Despite the numerous variables that literature indicates may affect appreciation, '[aggregate] AI scores are remarkably stable' (BBC Audience Research, 2009b), and even if respondents find it hard to pinpoint why they give a programme a particular appreciation rating, there appears to be relative consistency: '*people seem to often come to the same score* [even if] *their reasons for giving it are quite different*' (Collins, 2010b). Ultimately, regardless of the vagaries of comedy appreciation for individuals and their experiences of broadcasts, large samples *do* appear to provide data that allows broad inferences.

4.3.1 Context of the Experience

How we experience media can affect how much we appreciate it. For example, 'Excitation transfer theory' proposes that the physiological state induced by one programme can be transferred to an adjacent broadcast, inciting a subconscious change to the enjoyment of that subsequent programme (Bjorna et al, 2001, p7). There could be many further contextual factors affecting appreciation: time of day a radio programme is heard, where it's heard, with whom, on what device or platform, whether on headphones or not and with how much attention paid. Since some of these aspects are covered by the Pulse survey a correlation between appreciation and these elements could be investigated.

There is also other metadata which could allow further metrics such as the time of day of the broadcast, to be analysed in relation to appreciation. Unfortunately, these aspects do not cover all the particular elements that we might wish to understand particularly for Radio 4 comedy: for example, we do not know whether listeners are listening alone or in groups. The following points attempt to summarise the general variables that may be involved in Radio 4 comedy appreciation, based on: context of the experience, the exact question asked, the demographics of those being asked, effects of memory and agreement and, more specifically, factors relating to certain comedy sub-genres.

i Context – Expectations

Good appreciation scores are considered to indicate that, for the most part, our expectations are indeed met (Silvey, 1974, p116; Carrie, 1997, p24, p26; McQuail 1997, p74-75). Appreciation of a programme is driven partly by how it delivers against our expectations based on past experience (McQuail (1997, p57, p74) with different genres inciting different expected utilities (Collins, 2010b, Rumble, 2010). For example, what we might expect from a radio comedy panel show - laughter, distraction, and companionship - we might find undesirable in a serious news programme.

> A man who chose the reply 'like very much indeed' when asked to express his feelings towards his neighbour and subsequently chose the same reply when asked about his wife could not therefore be said to have liked them equally. His two answers would not be comparable because he would have applied different standards, liking his neighbour 'very much indeed' as a neighbour and his wife 'very much indeed' in the rather more exacting role of a spouse. (Silvey, 1974, p67)

Many aspects could affect a listener's expectations of a Radio 4 comedy. For example: their familiarity with the station, what they might normally expect to hear in that scheduling slot, what they already know about the programme in terms of sub-genre, the talent involved, what they've heard on the trails or read in listings

---

[164] See from p48.

or how it is introduced by the announcers. Where a programme does not deliver on these expectations, there may be disappointment and the appreciation for the show may be lower than if the listener had had no preconceptions to be dashed. Arguably comedy in particular attempts to illicit very specific responses (enjoyment, mirth, laughter) which if not achieved could result in disappointment:

> *I think that people actually feel offended if you have told them something is funny and if they're not laughing, you've sort of told them that they're wrong. I think people take real umbrage at that. If you think about it, just before a nine o'clock doc or a really intense political thing, we don't say: you're going to find this really interesting, it's very clever. But we do say: now it's comedy and you're going to have a really good time, and you've just let them down. Their expectations have not been met.* (Raphael, 2012a)

For example, one can imagine that a response to a comedy that one finds offensive will garner an extremely low score while it is hard to imagine a nature programme *really* annoying someone to such an extent.

An example of appreciation being affected by expectations can be seen in the example below. While comedies can be heard at 11:30 on R4 on Mondays, Wednesdays and Fridays, documentaries about music and the arts can be heard on Tuesdays and Thursdays. There is no reason to think that listeners are familiar enough with the schedule to be aware of such nuances and if someone were expecting one kind of programme but presented with another, the appreciation for the broadcast programme could suffer:

> *Writing in Three Dimensions: Angela Carter's Love Affair with Radio*
> 16-Feb-12 Female 68 (respondent 613704)
>
> Appreciation rating =   3
> 'swithced [*sic*] off-this to me should be a comedy slot not more of this worship of someone's favourite writers'

The comedy genre is seen to need time to bed in[165] as it sometimes needs to shock and surprise (Carr and Greeves, 2006, p131-189) if it is to innovate (Berthoud, 2010), and this may take some getting used to for the listeners. Berthoud claims that in comedy the greatest reward tends to require the greatest risk:

> *If it's going to be something memorable and brilliant and great for all time, something that becomes part of our culture, then that probably means that in some way it has to be distinctive and different. If it's going to be distinctive and different that probably means that you're trying something that's a bit new, people that haven't been tried before or people that have been tried before but in a format that hasn't been tried before. That's the risk – we're trying something new.* (ibid)

Thus it may be unlikely that a first episode will meet the expectations of a listener as by its very nature it may be unexpected and challenging. Dare (2012) finds that as elements of comedy become more familiar, they are more liked: '*People tend to like things that they hear every week. If you've got a regular character, they will like that regular character whether or not it's the funniest thing*'. This has been seen in appreciation scores as, 'New comedies tend to get higher AIs as audiences get used to the programme' (BBC Audience Research, 2009b). In this respect, the use of marketing and promotion may be an important element in familiarising the potential audience with the comedy ahead of its first broadcast, even if there is just a subconscious raised awareness of the comedy, or elements of it, such as talent, format or style: the 'mere exposure effect' (Kahneman, 2011, p82-83). Likewise, the 'halo effect' (ibid, p66-67)[166] could mean listeners were predisposed towards a new comedy if there were elements of the production that had previously elicited positive feelings.

---

[165] See from p130.

[166] See the next section.

The highest AI scores on Radio 4 consistently come from *I'm Sorry I Haven't A Clue*.[167] North (2010) claims that this is partly due to the fact that it does not challenge expectations, it is a familiar format, it delivers consistently, and as a consequence listeners are '*rarely dissatisfied*' and '*know that it's going to be quite funny in that certain way*'. This safety for the listeners is not available for new shows, '*and it may be that it's too different, it's challenging. You know, people like listening to the slot but they'd rather that it was JAM*[168] *rather than some strange sitcom*' (ibid).

A further consideration for comedy may include the zeitgeist of the genre at the time. It has long been recognised that comedic tastes are expected to reflect the moral requirements of society (Cantril and Allport, 1935, p226). For example, one view is that there was a requirement for broad knock-about comedy in the late noughties in the UK as a reaction to the recession (Chortle, 2011a). Comedy, such as Wartime's **ITMA,** that may be considered brilliant in its time could, based on changing expectations, be seen as poor in retrospect (Took, 1981, p71). More recently:

> *There's a kind of filter, that's not very scientific, which is what else was going on at the time? We got so many more complaints about the comedy around things like 9/11 and particularly around 7/7. Suddenly, anything to do with Christianity, people took exception because of anti-Islamic fears that were growing in this country. People wanted things that were more cosy, people were feeling very unsettled. And I sometimes do go back and think well, that was a really difficult time in this country.* (Raphael, 2012a)

ii Context – Familiarity

The 'halo effect' is a term long established in psychology. It is a cognitive bias which means that when we judge something we might take one element of that thing and apply that trait to the whole (Kahneman, 2011, p82-83, p206). The halo effect can be seen where one is generally predisposed toward the item according to a general impression and consequently rate it more highly (Moser and Kalton, 1971, p359). For example, a listener might rate a Radio 4 comedy highly just because it's on Radio 4 and they are loyal to the station. The halo effect can be augmented by increasing *familiarity*: if we have a higher awareness of something, we are more likely to think more highly of it. If, when learning of the cast list of a film we recognise some names, we might be more inclined to consider that it might be good. However, if we judged the actors known to us as being awful, we would probably have lower expectations of the whole. But generally Robert Zajonc's 'mere exposure effect' (Kahneman, 2011, p66-67) asserts that simply the *exposure* of a stimulus regardless of any knowledge of the 'goodness' of that stimulus, ideally subconsciously, can increase the affection that people will place upon something. In the example of the film, we may have just seen the name of an actor but not in the context of how good or bad they might be. When we hear the cast list, we think to ourselves, 'Oh, I've heard of her', and are likely to be predisposed to favour the film: exposure influencing our view of the whole. Engel (2009) looked at whether increased exposure to television programme information prior to broadcast (through media such as the internet and trailers) increased appreciation scores. He found evidence that the scores were indeed higher, but the increase was variable depending on external factors such as demographic differences.

Familiarity could be an agent if we were to see increasing appreciation through a comedy series and this would be most obvious in a format or style that is novel and demanding, as it might be likely to attract particularly low appreciation initially.[169] For example, in the late 1940s Norden and Muir attempted to create a show that did not just appeal to the lowest common denominator and thus was a more challenging listen

---

[167] See p224.

[168] *Just a Minute*.

[169] See from p130.

than was typical (BBC Year Book, 1949, p35-36). The resultant **Take It From Here** initially garnered 'unenthusiastic' responses (Foster and Furst, 1996, p104), but as audiences began to understand its principles the show became one of the most popular of the period (BBC Year Book, 1949, p35). However, that familiarity was the sole explanation for increasing appreciation in a case such as this is not likely. It must be considered that a show that is very novel may have teething troubles and be improved through its run; Muir and Norden were reshaping the show after each broadcast (Foster and Furst, 1996, p104). Also, one must reflect that if one is measuring the success of the show on appreciation scores, only those who have listened are responding. Increases could merely be the aggregate effect of those who *didn't* like the show at the start *ceasing to listen*, resulting in the mean increasing simply through the mix of the listeners.

While there is a general assumption that '*familiarity breeds* appreciation' (North, 2010), it may not always be borne out in appreciation scores, as regular programmes have also been found to have lower scores than one-off special programmes (BBC Audience Research, 2009b).

– Familiarity as a Risk Management Strategy

Radio 4 provides comedies that have run for decades but does so alongside 'the raw, fresh-from-the-fringe comedians' (Elmes, 2007, p225).[170] In finding new, challenging content there is an element of risk in any genre, but in comedy, where polarised views can illicit extreme reactions, the level of risk is an important consideration:

> Risk-taking is as vital a part of the BBC's mission in comedy, drama and entertainment as it is in other genres. As with all programme making, the greater the risk, the greater the thought, care and pre-planning needed to bring something ground-breaking to air. (BBC Executive, 2009, p8)

Henry Normal (Chortle, 2011b) claims that any new comedy commission is inherently risky, and for Radio 4, where comedy is considered as a key genre in attracting new, younger audiences, to do so it needs to be *seen* to be taking risks. Atkinson (2008) explains how decision makers actually aim to give 'the illusion of something *risky*' but in actual fact, it's very measured. There are a number of ways that commissioning decisions can reflect the need to minimise risk while appearing to embrace it. Where there are elements of high risk as part of a programme idea, the uncertainty can be mitigated by arranging other familiar variables to be lower risk. For example, an unproven comedian can be placed with an experienced producer and script editor (Berthoud, 2010). Alternately, using well-known, established talent in a new format can be a way of adding a familiar element to a show and be a hook for audiences who might otherwise be wary of the novelty.

Ultimately, recommissioning an existing series is comparatively low-risk compared to a brand new idea as for the former, the audience has had time to become familiar with the format, talent and style, and there is audience reaction data available. In contrast, brand new comedy is likely to result in a significant faction of listeners finding some of it unappealing unless there were a concerted attempt to appeal to all: to aim for the lowest common denominator or the 'least objectionable programme' (Webster et al, 2006, p181). This approach, however, is not in line with public service broadcasting policy and certainly flies in the face of the Reithian ethos of 'setting out to give the public what they need and not what they want' (Shingler and Wieringa, 1998, p29). It is also a strategy that McWhinnie (1959, p17-18) claims people would not accept for long. There are those who claim that Radio 4 comedy already pursues this unacceptable, lowest common denominator course. Reynolds has written a number of less than glowing reviews of Radio 4 comedy recently, writing of the station:

---

[170] See from p45.

> For years, radio comedy has nurtured the rest of the media, discovering writers and performers, giving them freedom to grow, pushing boundaries. To judge by what's been on Radio 4 lately, those days are over. Whoever is commissioning comedy these days is playing it safe, putting faith in semi-star names and familiar formats, following trends, not discovering them. (Reynolds, 2008)

Research by the BBC Trust (2010, p46) found that there is a proportion of listeners who felt the same way:

> A few felt Radio 4 could consider introducing or promoting more creative programme formats within its schedule. In particular, some younger, lighter listeners wanted Radio 4 to explore new programmes or formats targeting 'younger' audiences, potentially involving taking more risks.
>
> 'I'd say Radio 6 is innovative, I wouldn't say Radio 4 is…'
>
> (35-54 / Heavy [listener] /London)
>
> 'It's a bit safe isn't it? Radio 4's got a sound, a feel to it. It's a bit predictable.'
>
> (21-34 / Light [listener] /Swansea)

It is clearly a difficult balance for Radio 4 comedy to find the right level of innovation: allowing the potential of the '*fantastic prizes*' of a successful new comedy (Schlesinger, 2011) while ensuring that it does not alienate existing listeners seeking comfortable familiarity. These contradictory requirements have been a concern since the design of Radio 4: 'it seemed highly likely that a crude attempt to modernise would risk the rude displeasure of those who had been most loyal to the old arrangements' (Hendy, 2007, p33).

iii Context – Group Versus Solo Listening

In terms of appreciation, radio comedy has some very specific considerations. For example, while radio listening is often a solitary activity, laughter is more often associated with being in a group; if we are alone, we are unable to confirm our responses (Crisell, 1994, p164). Much research has found that humour appreciation is heightened when amongst other people. This is logical, as 'laughing' and 'enjoying oneself' are characteristically social practices (Hepp, 2006, p200), and presence of others is an important factor when considering the context of the experience of a comedy programme (Lieberman et al, 2009, p498).

Although whole families seldom sit round the wireless these days with rapt attention – if they ever did (Gray, 2006, p252) – we can assume that people still do hear the radio together on occasion, such as at the breakfast table or in the car. In the UK we do not measure this factor; RAJAR diaries do not ask respondents to specify whether they are alone or with others when they record their listening (RAJAR, 2012b), and nor does the Pulse survey that measures appreciation. North (2010) admits that it would be an interesting addition to include this variable. Radio research in the Netherlands, however, has looked into this aspect and found that there is a general pattern to solo listening; it is most likely to happen during early mornings and late evenings (Mytton, 1999, p124-125) and group (i.e., more than one person) listening occurs more in the daytime. It was also found that older listeners are more likely to listen solo than younger people. This research, unfortunately, did not consider whether this factor had any impact upon appreciation.

Some experiments in the mid-twentieth century showed that the larger the audience the more likely they are to laugh, explained partly by the theory that the laughter of others becomes a source of mirth itself (Martin and Gray, 1996, p222). Carr and Greeves, (2006, p181) theorise that in groups we become aware that we are being judged by our response and this can explain the altered levels of laughter and smiling:

> When jokes are told in public, people edit their responses unconsciously and continuously. All of us are much more likely to laugh out loud when we're part of a group rather than when we're alone – we are signalling that we get it, that we are part of a group with a shared sense of humour. …we understand, even if only subliminally, that when and how we laugh can give something away.

Thus, there is an argument that the response of an individual to humour when influenced by a group is not indicative of the true *quality* of the stimulus. If a respondent is influenced by the responses of others, is their evaluation objective (Cantril and Allport, 1935, p100, Chapman, 1976, p157, Martin and Gray, 1996, p222)? Incidences could be explained by actions such as 'feigned funniness' (La France, 1983, p5) or the adoption of 'ironic' amusement, particularly if the content is particularly bad (McQuail, 1997, p97).

However, regardless of the effect on response while in the group, surely when the respondent comes to evaluate the programme they have no reason to be influenced by others, and can give an objective opinion at that point? Lieberman et al (2009, p498), summarising research in this area, where 'social facilitation/situational cueing' is used with humour appreciation, state that there are no consistent findings. Sometimes group experience of comedy influences the appreciation of humour evaluation, sometimes it doesn't. Martin and Gray (1996, p222-223) explain that the inconsistency of findings in this area are due to both the wide variance in methodology in humour research and the validity of the experimental situation. The way in which humour evaluation is conducted is often not comparable to the way in which audiences would truly experience such stimuli and furthermore, the actual content has been known to be of low quality or poor suitability for the respondents. Martin and Gray claim that this lack of validity results in mere 'experimental artefacts' (p223) rather than anything informative about genuine responses to humorous content. The benefit of the research undertaken for this thesis is that the evaluations (appreciation ratings) are based on normal radio comedy listening rather than an experimental design.

- Studio Audience as a Proxy

Early radio comedy was seen as merely a conduit for existing art forms. In its primitive form it was just a 'live relay' of a theatre performance (Crisell, 1994, p164-165), broadcasting both the production and the audience's reactions.[171] Hence, from the very beginning of the life of radio comedy, a laughter track has been present in the homes of listeners: a *proxy* group sharing the experience.

> This studio audience really becomes a kind of broker in the transaction between performers and listeners. It is the agent of the performers because it encourages the listeners to laugh aloud, making them feel they are part of a large assembly and thus able to give vent to a public emotion. (ibid)

A laughter track, however, is not always desirable. Cantril and Allport (1935, p100), pioneers in radio research, asked over 1,000 listeners whether they found that humour was improved by having 'studio' audience laughter on comedy broadcasts; 61% said that the laughter was an improvement (39% said no). Reflecting on the fact that preference may play a part, there is no simple direct correlation between audience laughter and improved appreciation.

Martin and Gray (1996) found that a radio comedy with studio laughter was more appreciated than one without, but a study by Lieberman et al (2009) found that for a specific American television sitcom, the reception to a laughter track depended on which episode it was heard with. They write (p497) that the episode that did not benefit from a laughter track had a 'narratological richness' and 'more complex story structure' with 'higher levels of satire' than the other episodes used in the experiment. This is a concept discussed by Crisell (1994, p165) who posits that sometimes the listeners at home can actually feel excluded from the programme, particularly if the laughter seems to be unrelated to the audio: 'in this case the presence

---

[171] See from p15.

of the studio audience is counter-productive'. Thus, some types of comedy may benefit from studio audience laughter but others may not.

So when is a studio audience required for a Radio 4 comedy show? Broadly, the choice is made based on the prediction of whether or not the performance will work in front of an audience (Schlesinger, 2011), often those with a higher joke-count being more suitable (Mayhew-Archer, 2011). The long-established use of studio audiences to create a laughter track continues to be employed on Radio 4 comedy, particularly in the 18:30 slot where it is understood that 'The act of sharing laughter with the studio audience enriches and energises the listening experience at this busy time of the evening' (BBC Radio 4 Commissioning Guidelines Spring 2012, p53-55). But, as the growth of social media allows audiences to communicate with each other during broadcasts, might that group experience be shared in a new way?

iv Context – Time and Day

For live comedy, there is an understanding that people are more receptive to humour later in the day (Berthoud, 2010), but Radio 4 offers comedy early in the day: between 11:30 and 13:00 five days a week. Are audiences more receptive to comedy at certain times even if broadcast rather than experienced live? Day of the week may also be a factor as, if their routine is, say, a Monday to Friday work week, people may be more open to having fun and being entertained towards the end of the week and at the weekends (Bennett, 2011).

The level of appreciation at particular times on particular days of the week may be related to the sub-genre of comedy. A raucous show could be more appreciated late on a Friday night (ibid), versus something more thoughtful and whimsical on a weekday teatime (Mitchell, 2010). We might assume that what is *required* by the listeners from Radio 4 comedy on a Sunday evening might be different from what they want to hear on a Monday at 11:30am, and this might influence their appreciation of a show. It could be the case that the relationship between the format of the show and the daypart could affect appreciation as the secondary activities of the listener could dictate the 'utility' required of the programme. Mytton (1999, p124-125) found that attention paid to radio varied throughout the day while the BBC's 'share of ear' research found that, for Radio 4 specifically, listeners were only listening with attention 41% of the time (BBC Audience Research, 2009c, slide 23). If it was a time of day when one could give significant attention to the programme, for example, when driving, the listener might have a greater appreciation for a programme with a complex narrative. Conversely, if the listener were engaged with a demanding activity and just wanted the radio as background companionship, perhaps a 'dippable'[172] (Raphael, 2011) show might be more appreciated. The expectations of the listeners would be different at various times of the day.[173]

Time and day have been found to impact upon how likely it is that listeners are listening in groups or solo.[174] A dippable panel show like *Clue* can be more appreciated when we can laugh along with others or are unable to give our full attention to the broadcast, whereas a complex programme like *Bigipedia* might require more attention and perhaps solo listening to appreciate it fully.

The time of day could relate to the mood of the listener and in turn affect appreciation. The BBC's 1997 research sought to identify the mood of the Radio 4 audience for different dayparts in order to satisfy the listener's needs though the schedule (for example, Counterpoint Research, 1997b, p6). The mood of the

---

[172] I.e., a programme that a listener can dip in and out of, without needed to give full attention.
[173] See from .
[174] See .

listener is not currently measured in either RAJAR or Pulse research, but is something that is under consideration:

> We had a really interesting discussion about mood and whether or not we could have a mood monitor. Perhaps as part of the AI score on the panel page. It would be quite fun to say 'what mood were you in at these times of day?', so that we could then start to map that. There is some [internal] research on levels of mood and happiness throughout the day. I think we were going to have a look at those sorts of patterns and see whether that reflected on AIs. But all you can see is correlation. You can only say 'well, this seems to be up and this seems to be down', but you can't say whether one is driving the other. If you're feeling in a bad mood, do you choose to watch television that puts you in a worse mood or do you make some compensation for that? Is it the television that's making you in a bad mood in the first place? I can imagine for some daytime TV that might be the case! So all we can really do is observe. (North 2010)

GfK attempted to look at how appreciation varied across dayparts for German television (Preunkert. 2009, slide 3-12). Although differences were indeed found – for example, aggregate appreciation scores were found to be lower in the mornings before 10:00am and differences seen between weekdays and weekends – it was not fully understood whether there was any intrinsic difference in people's scoring or if it was purely driven by content mix (North, 2010).

v Context – Adjacencies

Some research into commercial advertising has identified that adjacent content can affect an audience's response to comedy. Other programmes and interstitial items that are experienced around transmission can effect evaluation, for example, religious content experienced just before humorous content was found be detrimental to some people's sense of humour (Saroglou and Jaspard, 2001, p33). Schachter and Wheeler (1962), looking specifically at humour, explain that if the initial stimulus of the audience member is ambiguous (for example, an advert or trailer), the attribution of the aroused state will be transferred onto his or her immediate environment (post-exposure), an effect that Cantor et al (1974, p. 812), citing Schachter (1964), describe as 'misattribution of arousal'. Bjorna et al (2001, p7) reference Zillmann (1971), stating that:

> The intensity of an emotional state will depend upon the level of excitation at the time (and) the level of excitation is a function of stimulus specific arousal and of residual excitation, due to incomplete delay, the past excitement is still present. The residual excitation will potentially produce an "over-intense response to more or less immediately subsequent stimuli".

This was illustrated regarding humorous radio content by Benson and Perry (2006) where they found that a light-hearted radio show was enhanced by the placement of funny adverts around and within it. However, this kind of immediate context of listening can also have a detrimental effect upon programme appreciation. Perry (1998, p2) identifies what he called the 'distraction hypothesis' where he found that humour in television advertisements resulted in a sitcom being rated with decreased programme enjoyment (p15). While there are no advertisements on Radio 4, there *are* interstitial elements between the programmes – trailers for other programmes and continuity announcements. Whether having a positive effect on programme appreciation through *excitation transfer* theory or a negative one through *distraction* theory, an observable effect is possible. Humour in the interstitials is not the only kind of stimulus that can change the liking for the adjacent programme: Perry (1998, p17) mentions music and advertisement quality as a stimulus, and Bjorna et al (2001) identified energy levels as a differentiator.

For Radio 4, there are often complaints about the number and frequency of trails;[175] heavy listeners are likely to hear some trails a number of times. Should one be found particularly annoying, and be heard

---

[175] BBC Trails are on-air advertisements for BBC programmes – on Radio 4, these are around 30" long.

directly before a radio comedy programme, it could have an impact upon the perception of the show (distraction theory). Such stimuli can cause physiological *priming* effects which impact upon humour appreciation. An experiment was carried out where some students were asked to evaluate some of Gary Larson's 'The Far Side' cartoons. They all did so while holding pencils in their mouths, some sideways, forcing their lips into a smile and others just at the end, making their face into a frown. Those in the first group rated the cartoons funnier than those in the second; just changing their face shape changed the way that they evaluated humorous stimuli (Kahneman, 2011, p53 -54).

On Radio 4 the buffer between the trail and the programme is the announcer, and it is easy to imagine that the way an announcer introduces a Radio 4 comedy could affect listener perception. Continuity announcers '*are your friend, guiding you and hand-holding you through the schedule, signposting what's coming-up*' (Hubbard, 2012). Their way of 'selling' a programme to listeners could be an important factor in preparing them for what they are about to hear.

vi Context – Platform

It has been found that appreciation of television programmes that have been viewed either on digital platforms or through non-linear routes is higher than via traditional channels (Collins, 2010a, Thickett, 2010). This is attributable to two aspects. Firstly, as they are able to be more selective in their viewing, television audiences are watching more of what they like and therefore total appreciation is on an upward trend. Secondly, even the same programme viewed on two different channels will achieve a higher appreciation rating when viewed digitally. The widespread understanding of this phenomenon is that rather than BBC1 feeding a set menu for an evening, choosing between more channels gives the viewer a more autonomous experience: an a la carte menu. Viewers are more likely to rate highly programmes actively chosen, since otherwise they would be admitting that they had selected badly (consumption bias).[176] Radio appreciation is increasing, but not on the same '*trajectory*' as television because, up to now, there is far less time shifting and fewer digital alternatives (Collins, 2010a).

4.3.2 The Question

We have seen how the Pulse appreciation question asks the respondents to 'rate' each of the radio programmes that they heard,[177] but it has not always been worded thus.[178] For many years it was based on a five-point alphabetical scale (A+ A B C C-) with corresponding verbal equivalents ranging from '"I like it very much indeed", through "I have no strong feelings" to "I strongly dislike it"' (Silvey, 1974, p67). The wording of the question brought to the fore the element of 'liking' into the programme appreciation rating. Changes in the BBC's rating system meant that in the late 1970s, the ***The Hitchhiker's Guide To The Galaxy*** radio programme, for example, was still calculated from a 5-point scale, but the wording of the question asks the respondent to rate how 'worth hearing' each programme was.[179] Whether the change between these methods of questioning was fully validated at the time is not clear, but it is safe to assume that there *could* be differences between the results when the wording or the scale of the question is changed substantially. Even if the wording is the same, different respondents can interpret it in different ways. To 'rate' a programme (deliberately ambiguous), one person may relate it more to enjoyment, another to quality, entertainment value, or any other aspect they consciously or unconsciously assign to it. If the interpretation is

---

[176] See from p152.
[177] See p85.
[178] See from p75.
[179] See p75.

different, will that necessarily result in respondents allocating different scores, and if they do, does it matter (Crisell, 1994, p200)?[180]

Carrie's 1997 (p58) research indicates that the questions used in television appreciation at the time asked people to rate programmes on a 6 point scale for how '*Interesting and/or Enjoyable*' they were. This allowed respondents consciously or subconsciously to attribute the most appropriate word to the relevant genre, for example, a comedy programme would be evaluated on its enjoyableness while a documentary would be allocated a score based on how interesting it was. Much work was done to validate the wording (ibid, p61) and it was found that people were able to use it effectively to discriminate between programmes of different genres. Ultimately it has been found that, whatever the wording of the questions, there is a certain level of consistency with the expected range of answers. Barwise and Ehrenberg (1988, p50) demonstrate this to be the case by referencing work done in different countries, with British, Canadian, North American and West German respondents giving appreciation ratings resulting in equivalent means, largely between 60 and 80, regardless of the exact wording or scale of the question.[181]

The current methodology of the BBC involves conducting the Pulse survey online, but previously it was done with pen and paper (also on a scale of 1-10). North (2010) explained how initial comparisons of online versus paper survey response to appreciation ratings in Holland had resulted in very different figures for the same programmes. To further investigate this phenomenon, GfK ran an online survey in parallel to the paper one and found that the differences were due to the demographic mix of those online rather than any other factor. Once the data was correctly weighted, the two measurement systems were found to be acceptably close. Van Meurs (2008, slide 31) illustrates the results of this validation, undertaken for 2 weeks in 2004. He found that there was only a 2% difference between the two measurement systems; over 90% of the programme AIs were within 10% of one another. He also found that the online system had a slightly wider spread of responses, which he attributed to improved immediacy of response.

It is worth noting that, as with any survey, aspects such as question order, ambiguity, use of jargon, inbuilt bias, and many more attributes can affect the responses (Moser and Kalton, 1971; Fricker, 2008).

4.3.3 Demographics of the Respondents

i General Rating Differences

Across different demographic groups, variations in the 'average' level of appreciation allocated have been observed. The research department at the BBC gave the following guidelines for those attempting to interpret appreciation scores (BBC Audience Research, 2009b):

- Women score programmes higher than men by an average of 2-3 points.
- Respondents aged over 65 score higher than other respondents, typically by 3-4 points.
- The higher the social grade, the lower the AI tends to be. On average ABs award AIs 2-3 points lower than DEs.
- Younger people tend to give lower ratings so look for an average that fits with your target audience.

Carrie (1997) did extensive analysis of television AIs in order to try to understand differences in appreciation. He found that 'a variety of demographic variations' was confirmed (p24), indeed including higher scores given by women, older people and lower income social groups: similar to the rules of thumb

---

[180] Discussed further from p120.

[181] Exceptions to this generalization were found in South Africa where the score averaged about 60 as there was only one government-controlled channel available, and Wales where an average of over 80 has been excused as being due to poor translation of the scale from English to Welsh. (Barwise and Ehrenberg, 1988, p50).

given by the BBC researchers above. This was, in turn, a similar pattern to previous research such as Menneer, 1987 (p254).

These are very broad brushstrokes and may be misleading for a number of reasons. These analyses were conducted over a number of programmes so are the variances due to absolute differences between 'types' of people or due to the weighting and mix of the programmes watched or heard?[182] Do women rate more highly because they are more generous with their scoring generally, because they really do appreciate programmes more than men do, or is it because they watch more programmes within genres that achieve higher appreciation?

> *To be able to say something general about a particular demographic on a total level would be quite difficult. It would probably be because something like, if more women spend more time watching soap operas and soap operas are the most popular form of television, then the average AI for women will be higher than for men… It's the* <u>*composition*</u> *of the choice of programme.* (North, 2010)

The finding that younger people tend to be less satisfied with broadcasts is not always the case. In the early 1970s, it was found that 16-19-year-olds gave the highest ratings for BBC television and radio compared to other age groups. One theory to explain this was that despite a more cynical attitude, they have a smaller frame of comparative reference and were too young to feel that 'things are not what they used to be' (BBC Audience Research Department, 1976, p16). Looking to the future, as appreciation tends to be higher via digital platforms[183] and younger audiences are more likely to be using newer technologies, their approval may again increase (Thickett, 2010).

Although Carrie (1997) identified differences for appreciation based on demographics, he also found that 'these differences are smaller than supposed' (p27), and concluded that 'little of the overall variation in average programme appreciation scores, for programmes classified by type or by channel, is due to audience composition effects' (p28). Hence, rather than the audience composition being the *primary* factor in AI variation, he concludes that the appreciation score truly 'reflects something qualitative about the programme itself, rather than just something about the differences in audience composition and the appreciation scoring styles of the particular individuals who watch it' (p28). Ultimately, the question of whether certain groups rate more highly intrinsically or do so because the content is more suited to them is a question which has never been fully answered. For example, when considering gender:

> *Interpretation of these differences is, however, a matter of conjecture. Is it the case that men are more demanding of television standards than women? Alternatively, are women simply more appreciative of the contribution of television towards their lives than men? Or could it be the case that men actually are less well served by broadcasters than women?* (Menneer, 1987, p255)

Size of audience and appreciation score have a relationship (albeit complex)[184] so how does that link to the audience make-up? Where analysis has been done at a low level (genre specific), broadly, the higher the AI score, the higher the audience. Can this be also true when we consider highly-targeted programming? If people are becoming more selective about what they're watching because they have more choice, we might expect that niche programming is more highly appreciated because only those who are really interested will watch or listen to it. 'Programmes with specific, targeted audiences tend to do better – because they have come specifically to listen to the programme they are rating' (BBC Audience Research, 2009b). McQuail (1997, p58) considers that appreciation could actually be particularly low should this same programme be

---

[182] See from p126.
[183] See p104.
[184] See p91.

placed in a prominent slot as with a wider audience the *average* appreciation might be lower as the majority of people may be very dissatisfied as they are not served by the programme, even if the target audience are suited. In a broadcast environment where there is limited choice the viewers may not consider the targeting of the programmes as the primary focus, rather selecting the 'least objectionable', indicating that appreciation is likely to be middle of the road rather than extremely good or bad (Barwise and Ehrenberg, 1987, p63).

The distinction between *niche* and *popularist* programmes may have less of an impact in radio. Compared to television audiences, it is commonly thought that radio has more loyal listeners; people switch on Radio 4 and just leave it on partly due to inertia,[185] and are unlikely to make appointments to listen. For example, the average time spent listening to Radio 4 is higher than watching the BBC's flagship television channel, BBC1 (BBC, 2012a, p5, p7):

Average Weekly Time spent per user: BBC1 = 07:35 hours,
                                        Radio 4 = 11:56 hours.

Radio 4's remit is not to offer niche programmes for niche groups, rather it is to offer programmes that the 11 million listeners each week will find of interest. 'Radio 4 does not target any particular demographic group; it aims to reach everyone in the UK interested in intelligent speech radio' (BBC Executive, 2010, p14).

Audience 'loyalty' has a relationship with appreciation. Both Barwise and Ehrenberg (1987, p69) and Carrie (1997, p206) found that the more television episodes watched, the higher the appreciation. The distinction can be marked (Barwise and Ehrenberg (1987, p69):

Frequency of Viewing (out of 5 weekly episodes)	1/5	2/5	3/5	4/5	5/5
Average [Mean] Appreciation Score	52	61	68	80	87

Danaher and Lawrie (1998, p61) reference a number of publications - mainly by Barwise and Ehrenberg - in relation to loyalty, saying that where audiences are larger, the programmes have higher loyalty and higher appreciation, the 'double jeopardy' effect (p61).

Carrie (1997) found further distinctions, such as the overall amount of television that viewers watched. He segmented his responses into light, average and heavy viewers and found that those who watched over 200 programmes over the 5 week measured period (i.e., gave appreciation scores for those programmes), gave higher AI scores than those who watched fewer programmes (p83):

Weight of Viewing	% of Individuals	Average' AI Score
Light	25%	69
Average	40%	71
Heavy	35%	74

Carrie found this to be true regardless of the way that weight of viewing was classified: 'The key main finding was highly consistent: heavier viewers of television tended to be somewhat more generous on average in their audience appreciation scoring patterns than were lighter viewers' (p85). This is interesting because one might assume that lighter watchers might be more discerning and therefore more appreciative of what they select to watch. Perhaps instead this indicates the seemingly obvious point, that people who like television more, i.e., are more appreciative of it, watch more of it?

---

[185] See from p124.

ii Response Spread

Another area of audience segmentation which Carrie (1997, p89, p95) identified as having a correlation with appreciation was with how the respondents actually allocated their scores: whether they were 'consistent' versus 'regular' versus 'varied' appreciation scorers. His research on the 6-point scale[186] showed that respondents could be segmented into groups where there were differing levels of variation in the scores given over the measured period. Where respondents tended to stick to one appreciation score (consistent = 80% of the time they gave the same rating) the mean AI was much higher than among those who varied their responses. Regular scorers (regular = limited to two points for at least 80% of the time) were not as appreciative as Consistent scorers but were more so than varied scorers (varied = the remainder of the viewers). (Carrie, 1997, p89):

Category	% of Individuals	Average' Score
Consistent Scorers	16%	78
Regular Scorers	59%	72
Varied Scorers	25%	67

This makes sense as we understand that people are likely to watch what they like and give relatively high scores – aggregate AIs for BBC television = 83 and BBC Radio = 81 (10-point scale, BBC, 2012a, p6, p8). So where scoring is more varied, there is more space at the lower end of the scale than at the upper end. Those who give more varied scoring, Carrie found, tend to be male, 16-34's, ABC1 light viewers, and their total mean scores were indeed lower than those who score more consistently (Carrie, 1997, p95).

iii Rating Differences in Comedy

Gender differences

Enjoyment of comedy is highly subjective and many academic studies have identified independent variables that contribute to the variances found within audiences. Gender appears to be a factor in the results of some studies but findings vary greatly in this area. Madden and Weinberger (1982, p9), for example, concluded that gender *is* a consideration in humour appreciation despite conflicting findings. Sanner et al (2001, p5) discuss research that indicated that 'women process humour with more emotion while men process with a more analytical standpoint' due to physiological differences relating to brain hemisphere processing. In other studies however, for example, Jackson and Jackson (1997, p4), no significant variation between humour ratings of the genders was found, explained by the 'emerging male viewpoint of greater equality between the sexes'. Sanner et al (2001, p.15) also found no conclusive difference between men's and women's responses and proposed that reactions to humour were becoming more 'androgynous'. Benson and Perry (2006) also found no significant gender differences when measuring men and women's responses to humorous radio content.

Difficulty in pinpointing these differences could be attributable to the constant changing of societal roles, perceptions of gender and even comedic fashions. For example, considering gender roles, in the 1970s much comedy consumed in the UK reinforced traditional family roles (for example, *The Good Life*, 1975-78; *Some Mothers Do 'Ave 'Em*, 1973-78; *The Fall and Rise of Reginald Perrin*, 1976-79) and the stand-up show which ran throughout the decade, *The Comedians,* drove home the 'my mother in law' type of joke. In the 1980s politically correct, 'alternative' comedy was fashionable, with programmes such as *The Young Ones*, 1982-84; *Comic Strip Presents*, 1982-2011 and *Saturday Night Live* 1985-87, capturing the imagination of the younger audiences in a backlash against traditional stereotypes. *Men Behaving Badly* (1992-1998) was

---

[186] As his analysis is on a 6-point scale, it is not directly comparable with recent appreciation data used for analysis in this thesis which is collected on a 10-point scale.

later an instigator for a reaction against the PC alternative viewpoint through 'new laddism', which ultimately resulted in a style of 'metabigotry' in the 2000s. This manifests in superficially 'offensive' humour, understood as being expressed with the sentiment of 'postmodern irony'. That is, it tends to be a character's ignorance or insensitivity that is the basis of the humour, for example, David Brent in *The Office* (2001-03) or Jimmy Carr's on-stage persona – humour that might have been considered sexist in earlier decades, now acceptable again. Perry et al (1997, p389) identify that there appears to be a moving target in this area, stating that 'differences may not be constant and may be changing in directions which are still unpredictable.'

A further explanation for the inconsistency of findings in this area is that the type of humour used in experiments can be a factor. Sanner et al (2001, p4-5) discusses how some studies have found that there are significant variations of response when the humour is:

sexual – preferred by men
aggressive – preferred by men
'sick' – preferred by men
nonsense/absurd – preferred by women
anecdotal – preferred by women

But, even at this lower level of analysis, there is inconsistency, for example, Madden and Weinberger (1982) reference a study that found no gender difference for nonsensical humour, conflicting with findings from earlier studies.

Even when respondents are expected to be particularly objective, skews of appreciation across the genders can still be found. Morinan (2010a, 2010b) analysed the 629 ratings of the 2010 Edinburgh Fringe Festival's '*Chortle*' and *Three Weeks* comedy reviewers where each show is given a star rating from 1-5 (zeros are only given if an act does not show up or behaves unacceptably toward their audience). He found that women reviewers gave scores higher than those given by male reviewers. Within this star rating, it is a difference of 0.24, statistically significant according to Morinan. Where the element of 'people consume what they like' is present in normal appreciation ratings, reviewers are allocated shows rather than choosing them. Thus, even with consumption bias[187] stripped from the data, eliminating the mix effect,[188] women still gave higher ratings for comedy than men in this study.[189]

While the gender of the respondent is of interest, the gender of the performer is too. Further analysis (ibid) found that there was significant bias towards male performers by the male reviewers, which was not found among female reviewers. 'The most plausible explanation for this is that male reviewers (in general) have a prejudice against female performers' (ibid). On Radio 4, one only has to listen to any panel show to hear that women are under-represented relative to their proportion in the population. However, this is likely to be attributable to the fact that there have always been fewer female than male comics (Galton and Simpson, 2012). This is not necessarily because women are less funny, but perhaps because female performers tend to be drawn to acting rather than stand-up (Mitchell, 2010). Also, Greer (2009) claims that it's harder for female comics to break through because, reflecting societal pressure while growing up, they 'have not developed the arts of fooling, clowning, badinage, repartee, burlesque and innuendo into a semi-continuous performance as so many men have'.

The gender of the *subject* of a joke can also be a factor in the appreciation of its humour. Jackson and Jackson (1997) found that for men, the target of a joke, be they male or female, drew no distinction in its appreciation. Women, however, were more amused when the butt of the joke was male rather than female.

---

[187] See from p150.
[188] See from p105.
[189] Not peer reviewed.

<u>Further Differences</u>

Gender is just one way in which we can segment humour responses. Madden and Weinberger (1982, p14) suggest ethnicity to be a possible element of humour enjoyment, while Saroglou and Jaspard (2001) suggest that religious fundamentalism can be a factor. Culture may also be a differential factor; de Mooij (1998) suggests that humorous content in the UK versus the US differs with British humour tending more towards the use of puns and satire and aiming towards being more 'entertaining and clever' (p223) and with more ambiguity than the US. Weinberger and Spotts (1989, p39) found that humour is used more frequently in UK advertising, perhaps related to the 'soft sell' model more typically found. Further independent variables could include age, regionality, social class and attitude. For example, Carr and Greeves (2006, p205 p209-210 p215) discuss how some groups have different attitudes to comedy.

Particularly pertinent to radio, the voice of the performer may have a key role as they may be representative of a specific group. For example, when discussing the apparent polarised response to **Count Arthur Strong's Radio Show**, Roger Bolton expressed a view from the listeners: 'Some people have written to us to suggest that actually, what's going on here is that people with northern accents are not found funny by people in the south' (*Feedback*, 2012b). Caroline Aherne (Bradbury and McGrath, 1998, p168) goes as far as to say that certain regions actually find different things funny, and claims that those in the North East of England have a particularly dry sense of humour. This is not a universal view. For example Ian Pattison, discussing *Rab C. Nesbitt* (ibid, p131), suggests that concessions *don't* need to be made to ensure that audiences can understand regional colloquialisms as he feels that people will understand and appreciate the comedy as they get used to it. Johnny Speight claims:

> They're the same in Macclesfield as they are in bloody London. A very funny Neil Simon play set in New York will work in London, because people are living in roughly the same conditions, in a block of flats, all the same things happen… and are recognisable as sufferings of people. (ibid, p23)

If regionality is a factor it may be exacerbated if we consider bias in representation, as Radio 4 has long been accused of London-centricity. Since Reith's days the BBC has tended to centralise its production (Hendy, 2007, p120), which can affect the representation of non-London areas. A memo to Tony Whitby during the 1970s highlighted this:

> It is notorious that most producers and programme executives are more familiar with the Costa Brava than with Birmingham and Manchester, let alone Cardiff and Glasgow. Few of them have been north of Barnet. (Hendy, 2007, p120)

…something that may still be in evidence far more recently:

> [Radio Comedy] *wouldn't know an equal opportunities policy if it hit it on the head. ...It's appalling. They* [BBC Comedy] *have no sense of equal opportunities whatsoever and that bothers me because I kind of rely on people coming through*. (Raphael, 2012a)

It is recognised that London is a 'massive attractor of comedy talent' (BBC Executive, 2010, p51), and while this means that the majority of programmes are written and recorded in the capital, it doesn't mean that the setting needs to be there too (ibid). However, BBC research conducted in 2009 found that audiences actually did not feel that BBC network radio as a whole seemed London-centric (BBC Audience Research, 2009a, p13).

Another consideration is the age of the listener. It has long been recognised that younger audiences are more comfortable with newer, more challenging types of comedy, e.g., Took, 1981, p162; Briggs, 1995, p805; Crook, 1999, p129 – see from .

All these factors and many more, may contribute to how a listener perceives radio comedy. As humour studies vary so much in their scope and measurement it is not surprising that conflicting data has been generated. Experimental conditions rarely replicate normal situations and use of humorous material and the material itself can be questionable (Martin and Gray, 1996, p222-223). Comedy is so specific it is hard to make generalised statements that apply to all comic stimuli on the basis of very specific ones. The advantage of using the Pulse data for analysis is, in theory, that all the incidences of radio listening that are being rated are real-life occurrences rather than from an experimental design which, for comedy research in particular, would likely be lacking in 'external validity' (ibid, p222).

4.3.4 Memory

An important consideration for appreciation ratings is that, in giving AI scores for BBC Radio comedy shows, the panel members are not rating the *actual* experience but the *memory* of that experience.[190] As they are responding retrospectively, typically hours or days later, they are giving their view of the recollection of that experience rather than how they really felt during their listening. Kahneman (2011, p381) illustrates this point:

> Confusing experience with the memory of it is a compelling cognitive illusion – and it is the substitution that makes us believe a past experience can be ruined. The experiencing self does not have a voice. The remembering self is sometimes wrong, but it is the one that keeps score and governs what we learn from living, and it is the one that makes decisions.

Kahneman (ibid, p380) presents research that found that the rating of a whole event is weighted on certain elements of the experience: how the respondent feels at the end of an event and at its 'peak' will be of greatest influence. Thus, when recalling a comedy show, a respondent would recall how they evaluated the whole broadcast at the time, and also if there were specific parts that were particularly good or particularly bad. One element can have a profound effect upon the appreciation of a whole programme, particularly if that element is bad:

> The psychologist Paul Rozin, an expert on disgust, observed that a single cockroach will completely wreck the appeal of a bowl of cherries, but a cherry will do nothing at all for a bowl of cockroaches. As he points out, the negative trumps the positive in many ways. (ibid, p302)

Given the polarising nature of the genre, this element may be of particular pertinence to comedy specifically where audiences might be likely to find some elements distasteful. Take a comedy programme made up largely of average jokes. Having one that is particularly good might not have a significant impact upon a listener's evaluation of the whole programme, but a joke that the respondent found actually offensive could be more likely to colour the whole broadcast for them.

On the most basic level, regardless of any external or internal factors that affect the appreciation scores given by people, it is recognised that there is little consistency of response for any individual.

> When consumers are re-interviewed about their attitudinal beliefs towards different brands, such as whether it "tastes nice" or is "good value for money", only about half give the same response as before. This great variability of consumers' expressed attitude does not seem to reflect any systematic erosion of their liking of the brand, but merely a degree of as-if-random or stochastic variation. (Castleberry et al, 1994, p153)

They (ibid) propose that this kind of study is almost never done by broadcasters because of 'possible conditioning, sample attrition and higher costs.' Common sense would suggest that asking someone to rate a programme, then asking them to do so at a later date when their memory of that programme has lessened, might indeed give different results. Alternatively, to avoid the effect of degraded memory, the respondent

---

[190] Measurement at the point of the experience would need to be carried out through other methods such as dials (Silvey, 1974, p131) or 'coincidentals' (Kent, 1994a, p12).

could re-listen to the programme which again might result in a changed score as the experience could be affected by having previously heard the programme.

This volatility could be seen in AI scoring, particularly where there is a time lag between hearing the programme and rating it. If the show were exceptionally good or exceptionally bad it might be easy to remember it, whereas, if it were more middle-of-the-road, is the respondent likely to bother to rack their brains to justify whether it was worthy of a 4, 5, or a 6?[191]

The gap between memory and experience can be narrowed however through prompts, which help the respondent remember how they felt when they were consuming the media, even to the extent of re-experiencing their physiological state (Kahneman, 2011, p392-393). On a practical level it could involve the listener being reminded of where they were, what they were doing at the time or who they were with when they heard a programme; this is a method which has been shown to improve memory of events (Twyman, 1994, p90, p95). These techniques are not employed by the BBC's Pulse survey.

4.3.5 Agreement

McQuail's model of expectancy (1997, p74-75) includes the concept of the audience member evaluating content through their belief system. This relates to a respondent giving a programme a higher appreciation score if they agree with the content of the programme. This has long been a consideration. In 1935, Cantril and Allport (p100) reported upon research which found that when radio listeners were asked if they would prefer to listen to speakers whose arguments they agree with, 56% responded with a yes, with 28% of the total saying that they would actively switch the programme off if the sentiments were contrary to their beliefs.[192] This phenomenon is seen too in later years. Hendy (2007) discussed how on the *It's Your Line* phone-in Radio 4 show in 1970, Germaine Greer was one of the guests. She was seen by the BBC managers to have given an exceptional performance, speaking on mainly sexual subjects. After the broadcast it was found that the AI (RI at the time) was a mere 55 compared to the 65-70 that the programme usually achieved. '"It doesn't mean to say they hated the programme", the producer protested: just that "so many listeners hated her and her views"' (Hendy, 2007, p71).

Oakley (2009) explains that research was carried out in this area at the BBC in the mid-1990s as documentary makers wanted to gain further understanding into what drove appreciation in this genre. He found that people did indeed register their disapproval to controversial concepts through their appreciation ratings regardless of how entertaining the programme was, 'almost as if registering a political verdict'.

These instances relate to factual programmes but there is reason to believe that 'agreement' can influence appreciation for entertainment shows too. In the early days of the BBC's audience research department, it was discovered that for drama programmes any differences found 'showed a suspiciously close correlation with the differences in the way the themes were regarded – the more the play's theme was "liked", the more highly was its production and performance regarded' (Silvey, 1974, p60).

If we are to consider Danaher and Lawrie's (1998) proposal that programme consumption indicates appreciation,[193] then we might infer that those that were likely to switch off the programme because they disagreed with the speaker had a lower appreciation for the show.

---

[191] See from p147.

[192] The openness to alternative viewpoints was skewed towards younger men (under 30) but away from older women who were least likely to continue to listen to concepts which challenged their understanding of the world (Cantril and Allport, 1935).

[193] See from p91.

Much comedy, satire in particular, can be seen to have a standpoint or perspective and can include elements of 'humour, irony, exaggeration, or ridicule to expose people's stupidity or vices, particularly in the context of contemporary politics and other topical issues… Holding up human vices or follies to ridicule or scorn' and exposing 'vice or folly' (Provenza and Dion, 2010, pxx). One person's understanding of what vice or folly are may not be another's. For example, the BBC is often accused of political bias in its comedy, apparently illustrated by broadcasting the views of left-wing comics such as Marcus Brigstocke, Jeremy Hardy, Mark Steel, Mark Watson, Sandy Toksvig and Hugh Dennis (West 2012). Caroline Raphael (Radio 4 Commissioning Editor for Radio 4 comedy) thinks that verity needs to be present if people are to appreciate humour: 'I think things are only funny if they are truthful' (Clinton, 2012). But one person's truth may not be another's, so this means that, regardless of the absolute 'goodness' of a joke or a programme, if a listener were to disagree with the 'truth' behind the comedic sentiment, their appreciation of the programme could be compromised. For example:

> ***The News Quiz***   13-Apr-12   Male 63 (respondent 1097015)
> Appreciation score = 1
>
> 'I usually avoid this piece of left wing propaganda. Caught it yesterday and it was worse than ever. The smug hypocrites at the BBC use our money to pay for this partial rubbish.'
>
> ***Andrew Lawrence***   22-Nov-12   Male 74 (respondent 1571269)
> Appreciation score = 1
>
> 'Misanthropy masquerading as humour. This very dull, boring 'comedian' is completely unfunny and epitomises the BBC's very left-wing bias. This wasn't humour – it was propaganda!'

Zillman and Cantor (1976, p100-101) propose that the question of agreement is not black and white and that it is a continuous scale depending on the *level* of agreement. They propose what they call a 'continuum of affective disposition' (p100) to attempt to explain how respondents might respond to humour where there is a specific viewpoint of a 'target' being 'disparaged' by another. They summarise that there is a more positive response towards humour from a 'disparaging agent' with which we are increasingly positively disposed and from humour against a 'disparaged agent' against which we are increasingly negatively disposed. That is, if you hear an insulting joke on Radio 4 you'll think if funnier if you agree with the person telling it and don't like the person it's about. For example, this joke about the Levinson enquiry from ***Tonight***, (Radio 4, 23:02, 17/05/2012, s02e01) is most funny if you are a fan of Zaltzman (you might already be familiar with his type and viewpoint of comedy) *and* if you have a poor view of Murdoch's behaviour regarding phone hacking, the joke implying that he considers himself to be all powerful and above the law:

> ANDY ZALTZMAN: It is true that God and Murdoch do share many things in common. Firstly, they're not believed by the vast majority of the public any more, and also they keep employing members of their own family in influential positions. And then they send their sons to be sacrificed for other people's sins.

However, Gruner (1976) discusses a study from 1966 which showed that, for some satirical content, the subjects *are* able to appreciate the humour 'independently' of their 'attitudes' (p300). Gruner also highlights his research of the following year, which indicated that while respondents were more likely to appreciate the humour of satirical pieces if it chimed with their beliefs, it was only if the message behind the content were made explicit. If it remained ambiguous, there was no difference depending on the preconceptions of the individuals (p299).

4.3.6 Comedy Sub-genres

Research into television appreciation shows that the type of programme, or genre,[194] is a factor influencing appreciation of a show. Variation is such that it is accepted practice that different types should not be compared with each other: 'There are big variations by genre so you need to compare within genre' (BBC Audience Research, 2009b). On the widest scale, some researchers have segmented programmes into broad sections such as 'demanding' versus 'relaxing' (Barwise and Ehrenberg, 1987), or 'demanding' versus 'entertaining' (Menneer, 1987, p247), finding that programmes within these segments are comparable in terms of appreciation. Generally it is found that entertainment programmes have lower appreciation; Barwise and Ehrenberg (1987, p64, p67) found that when comparing the mean aggregate AIs (based on a 6 point Likert scale), the demanding programmes elicited a mean of 78 compared to just 70 for the relaxing programmes (p67).

Menneer (1987, p255-256) summarised television AIs at a genre level, finding that Light Entertainment (based on BARB genre designation) did have a slightly lower mean score than the total. He did, however, comment that variations across genres 'are smaller than is sometimes supposed' (p255), finding a 7-point spread, with feature films averaging lowest and sport most highly. Barwise (2013) agreed with this in saying '*I don't think there are big differences between genres*', and that while measuring distributions of appreciation at sub-genre level could *possibly* prove fruitful, he predicted that the variation would be minimal.

Carrie (1997, p26) found, contrary to Menneer's analysis, that entertainment actually had a higher aggregate score than the mean for other types of television programme. He too found variation between specific genres but these were smaller than is often supposed. For example, sports, information, light drama and some light entertainment programmes (including situation comedy and quiz shows) generally achieve higher than average programme appreciation scores'. He found that the mean appreciation score differed for these groups, for example, 'Situation Comedy' had a higher mean score than 'Other Comedy' (Carrie, 1997, p131):

Figure 13 – Television AIs by genre (Carrie, 1997, p131)

Programme type	n (no of progs)	Mean AI	Range of scores
ENTERTAINMENT	1254	73	37
Light entertainment	427	73	34
Variety	6	67	10
Situation comedy	136	73	31
Other comedy	34	69	24
Chat shows	38	71	25
Quiz shows / panel games	164	75	30
Cartoons and animation	18	69	11
Family shows	16	70	10
Contemporary Music	25	70	22
All programmes	3015	72	51

Variation at sub-genre level is not surprising. On Radio 4, it might be expected that certain types of comedies will give the audience different expectations and deliver different types of product. A comedy drama will require a certain amount of engagement to follow the plot and understand the characters, while sketch shows and comedy panel games may need less attention from the listeners.[195]

---

[194] See p37.

[195] See p39.

Understanding these differences is important for programmers as comparisons are made between programmes for decision-making purposes.[196] If sub-genres are an independent variable in Radio 4 comedy, that needs to be taken into account when considering appreciation scores.

4.4 Limitations of AIs

Broadcast research teams restrict data dissemination, particularly appreciation analysis. Figures taken at face value can be misinterpreted and misused when not fully understood: 'The qualitative evaluation of media outputs by their audiences of "audience appreciation" as it is usually called, is inherently difficult to measure' (Kent, 1994, p16) and 'complex to interpret' (Gunter and Wober, 1992, p65). There are indeed a number of limitations to the Pulse survey, and in turn to BBC radio AI scores, that can constrain their usefulness. Many points relating directly to appreciation are expanded upon in below. In some respects however, such as for serving niche audiences, it can be argued that appreciation is a far more important measure than audience size. Collins (2010a) argues that quantity of audience members and their appreciation are best understood when used *together*, as such, both of these aspects are valuable to broadcasters.

4.4.1 Restricted Dissemination

As a rule, AIs have been subject to relatively extensive use within the BBC but are seldom published in the wider community, remaining mostly unavailable to press or academics. For the few countries where appreciation is measured, this is also the case:

> While audience size values are extensively published (at least the top ten, twenty or however many are convenient), this is only rarely done with appreciation measures. A crucial distinction to make is that, while audience size measures are essentially about programmes, appreciation indices are about people. It is this which makes them more complex to interpret and by that token, more valuable. (Gunter and Wober, 1992, p65)

Carrie (1997, p18-19) theorises that, despite interest from the press regarding the publication of appreciation ratings, broadcasters are reluctant to complicate the well-established common currency of audience size ratings. Indeed, despite the BBC's decision, in 2011, to release quarterly aggregate AI scores at station/channel level, there are no plans for systematic release of scores for individual programmes or series (Foster, 2011). Instead, this sometimes occurs on an ad hoc basis, often when a broadcaster needs to defend a programme against criticism, when AI's may be cited as 'objective evidence' that audiences like the programme (Neilan, 2012).[197]

Within the BBC, Silvey (1974, p35) argued that AI data should be easily available and well communicated.

> Though it was understood from the first that audience research findings would not be freely available to the press and public, it was equally important to preserve the principle that they should be freely available within the BBC. Hence I always tried to insist that any member of the staff had the right to see any audience research report.

Today, within the BBC, AIs for programmes with sufficient sample sizes are available on the BBC Gateway intranet system (accessible to all staff). Below is the screen that can be seen on the BBC intranet, in this case for *The News Quiz* as viewed on 24th March 2013.

---

[196] See from p88.

[197] See from p88 for an example.
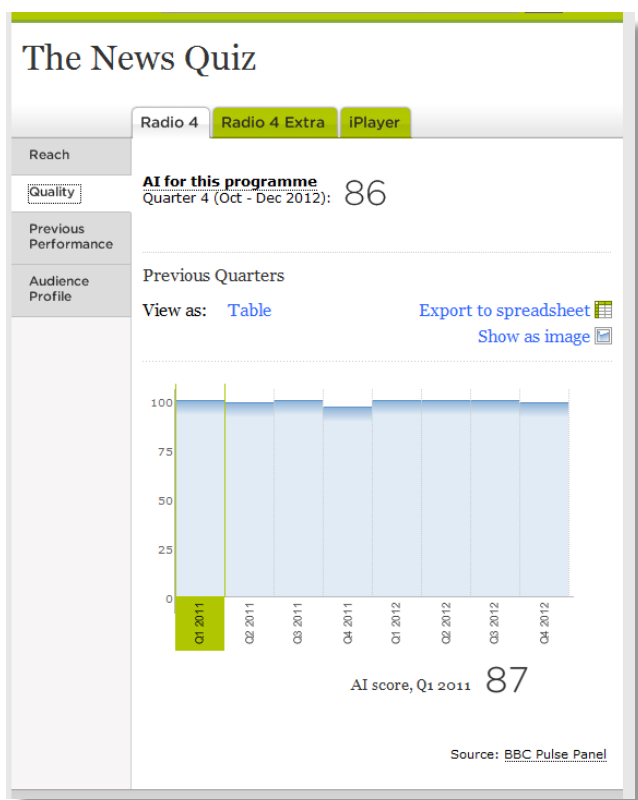
The figure above shows that the data for the most recent TX quarter is almost 3 months out of date at this stage, the 5 weeks of origination in January of 2013 not being shown in these figures. Hovering on each quarter when accessing the screen above allows the user to see the aggregate score for previous quarters:

Date	AI
Q4 2012	86
Q3 2012	87
Q2 2012	87
Q1 2012	87
Q4 2011	84
Q3 2011	87
Q2 2011	86
Q1 2011	87

BBC Gateway is relatively simple to use but is not suited to comparing AI data from more than one radio programme at a time. Lower level data - such as whether the programme AI score varies by demographic, AI for each episode or how many episodes or respondents constitute the score - has to be obtained from another web based system called Asteroid; this route is unavailable to most staff members.

Another system, AdvantEdge, allows more complex analysis and is available to members of the BBC audience research department. The Radio and Music Research Department's *BBC Radio Quality and Distinctiveness Report* is a weekly analysis of Pulse data covering all of the BBC network stations. It includes the aggregate AIs by network and shows trends over the recent quarter. There is appreciation and 'original and different' information for programmes across the different radio stations, with some programmes/slots being presented over the recent quarter to highlight the trends, for example, Comedy – Fri 18:30/Sat 12:30. The information for each station has a brief summary too, for example:[198]

---

[198] Programme names have been replaced due to sensitivity of data.

Radio 4: Programme 1 and Programme 2 both scored 87. Book of the Week: Programme 3 scored an impressive 85. Programme 4 scored 82 following the previous week's score of 81. Programme 5 scored 76 for the 3rd week on the trot. Programme 6 scored 78 – up on last week's 73.

Researchers are also able to create a limited number of ad hoc reports for those who request them.

Differential Access to AI Data

Lewis (1991, p24) claims that the relationship between BBC production and research has always been 'sketchy, to say the least'. Schlesinger (2011), talking of his time as Head of Radio Comedy, said:

> *To be honest I'm not sure we were allowed to see very much. We did get a monthly debrief from the research manager who used to talk about bigger trends. I always felt you could ring him up and ask for AIs. But I think it was largely around AIs rather than 'how many people listened to this series'. At that point I think Pulse was in its very early days and I'm not sure how much detail you could get.*

There is some suggestion that restricting data is a deliberate BBC policy. For example, the research company GfK found that their recent analysis of the distribution of AI scores, including radio comedy, did not reach anyone in Radio 4 programming, i.e., those that might use AIs in making decisions about the programmes. There is evidence of a culture within BBC research that data is restricted, as people who may not be 'research savvy' may misinterpret information if taken at face value:

> *For me, it's an interesting issue as to who has access to what at the BBC because, in theory, everyone has access to some of the reporting tools. I should imagine that given the amount of time pressure that everybody is under, they turn to the researchers and go 'Give me the latest on such and such'. And they don't have time to even think 'Maybe it would be interesting to look at something like this'... Our role is to make that information available within tools that we feel are accessible to the user. So we have in the past had easy dashboards and at the moment we have the web reporter as the main reporting tool. It's a bit clunky but it tells you the toplines. The research teams need to be quite careful not to make all the research available to everyone in such a way that they can misinterpret it and jiggle about with the numbers until they get the answer that they want. For us, all we can do is make available the tools and then the analysis and the insight is sporadic on our side because we're not quite sure, you know, there's 101 things we look at in any day. And therefore we really have a relationship with the BBC at the moment whereby we just say 'here it is, do with it what you will'. (North, 2010)*

While Gateway is generally accessible, within the Radio Comedy department, at the time of enquiry, only the Head of Radio Comedy and the two executive producers had access to Asteroid (Canny, 2011). And even this may reflect the particular desire for this data from these individuals, rather than a policy of sharing data. Raphael, as Radio 4's Comedy Commissioning Editor, gained access to Asteroid because it was more efficient than constantly requesting AI data from the research department. Thus she accesses the AI scores both for her own use and to pass them on to programme producers as part of her process of giving feedback (Raphael, 2012a). Even the existence of the Asteroid system is not widely known at producer level (Canny, 2011)

Similarly the distribution of the Quality and Distinctiveness report mentioned earlier is generally limited to more senior network staff. Additional people can be added to the distribution list, but its existence is not broadcast so few know to ask to be added. It in April 2014 it went to around 44 people in BBC Radio, plus the marketing team within R&M (BBC's Radio and Music division). However, in autumn 2014, the distribution was further limited so members of a specific radio network could only see the analysis for their own networks – previously all stations were visible on the report.

Freelance and independent producers, have historically had limited or even no access to audience data. Jon Naismith contrasts his experience as an in-house comedy producer in the early nineties (when AIs

were delivered to the pigeonholes of individuals and were the subject of discussion as a rule) with his current experience as a freelance and indie producer for Radio 4 comedy. Despite being the producer of many programmes for Radio 4 comedy, including the high-profile shows, *Clue* and *The Unbelievable Truth* he says that now: '*I don't have any audience research on my programmes and I'd like a lot more*' (Naismith, 2010). Instead, he is restricted to non-representative research (such as reviews or word of mouth) to gauge listener reactions to these long-running shows. The impact of this differential access is illustrated by the decision to decommission *You'll Have Had Your Tea* (2002-2007). As the show's producer, Naismith was told the decision was taken specifically because the AIs were poor (despite what he felt were good reviews). During a meeting with Paul Schlesinger, then Head of Radio Comedy about the issue, Naismith was unable to challenge the decision because he was told that AIs were not available (ibid).

This situation appears to be changing (if the data is viable) at least in the case of independent radio comedy producers. Talking at Radio 4's Spring 2011 Commissioning Round Launch Raphael told comedy suppliers:

> *I do send out AIs whenever I can. The absolute truth is you have to get a certain sample size* [50 or above responses] *and some of the shows, particularly late night, don't get that sample size so there isn't anything statistically valid I can send out to you.* (Raphael, 2011)

At the Spring 2012 Radio 4 Commissioning Round launch, it was clarified that, while AI's may not be available for other genres, AIs for programmes covered by the Comedy Commissioning Editor were well communicated if the data was available (Raphael, 2012a).

But even Radio Comedy, freelance producers do not automatically receive audience data. Bill Dare creator and producer of (amongst a variety of radio comedy programmes) *Les Kelly's Britain*, said he had received no information from the BBC about the reception of the show: '*I don't know what's happening at the moment because I'm freelance, and I don't know whether anyone's forwarding anything*' (Dare, 2012). Dare was unsure as to whether he had direct access to BBC Gateway data himself. Producers do, however, have the option to contact the BBC research department to request information on an ad hoc basis, should they have the desire to do so or the knowledge that it is an option open to them.

Whether talent is (or should be) provided with appreciation scores for the programmes in which they are involved is generally the choice of the producer, as that would be the only route available should a performer request that information. A producer might not want them to be aware of bad AIs, particularly at the beginning of a series, as there is a general understanding that comedy is often slow to take off.[199] Whether performers/writers are even interested in receiving that kind of information is questionable. Alan Simpson, speaking of appreciation scores, claimed that '*They were important in as much as we knew they gave us confidence. You thought, oh good, they enjoyed it. It gave you a good feeling. If the AIs were terrible, you would have lost confidence*' (Galton and Simpson, 2012). The decision to provide this information, however, was the producer's (Duncan Wood, for television's *Hancock's Half Hour*), and they never saw the reports themselves (ibid). Although none of the small sample of Radio 4 writers and performers interviewed for this thesis claimed access to appreciation data, this appears to be because they did not see a great value in such data. Sean Lock (*15 Storeys High*), David Mitchell (*That Mitchell and Webb Sound*, *The Unbelievable Truth*) and Nick Doody (*Bigipedia*) were all unconcerned that they were not given AI data and Doody was unaware AIs for radio existed. Indeed, the writers and performers seemed to distrust audience research, reinforcing Silvey's (1974) argument that the figures were seen as stating the obvious,

---

[199] See from p130.

Audience research for comedy, what are you going to find out other than the obvious that some people liked it and other people didn't? Everyone's listening to it with a different set of expectations in a different context. I think I have fairly good judgement as to what's good. (Doody, 2011)

Nobody even tells you it's being done, but you suspect. It's the nasty, ugly side of the business which they obviously keep from you. Why would they want to share that with you? Urgh, it's a horrible underhand business. If people don't have the confidence to trust talent… to me it's just lack of confidence. (Lock, 2011)

They kept that a dark secret. They were frightened we'd ask for more money. (Baxter, 2014)

And others see such data as inhibiting creativity. Black (2011) claims that if AIs became a communicated measure in the US, the broadcasting of comedy programmes would become a '*beauty pageant*' where comedians would consciously aim to try to increase their score rather than make the best comedy that they could.

To a certain extent, radio comedy performers are particularly well positioned to judge the response to their material, as many record their shows in front of a studio audience and can thus gauge reaction. Baxter (2014) claims that without having any official confirmation of a show's success, he knew that it was enjoyed by the reaction of the live audience reaction. Furthermore reception can also be ascertained by the reaction from the general public:

Everybody was using my catchphrase all of the time: 'If you want me thingummy, ring me.' Everybody was saying that about telephones.

However, how an audience behaves at a live recording may not always be indicative of how a programme is received by the Radio 4 listeners.[200] Thus, in theory, if performers are never made aware of any objective audience research data, they could merely tailor their performances to the relatively small number of those present in a studio audience, rather than the many listeners at home.

Despite limitations, the current decision makers in the Radio Comedy Department say they are happy with the amount of information that they receive from research. The Department has a '*really nice relationship*' with their research manager and are provided with bespoke desk research[201] five or six times a year (Canny, 2011).

4.4.2 The Respondents

i The Panel

The Pulse survey uses a panel of 20,000 people and, as with any panel, there is a constant need to maintain its size as people drop out. Finding the initial sample and engaging replacements after attrition is not straightforward, as there is no sampling frame for internet users (Vehovar and Manfrede, 2008, p181). GfK use a variety of ways to recruit respondents, enlisting help from online sample providers such as *Toluna* or using their own bespoke online messages targeting the harder-to-reach groups (North, 2010). Face-to-face, telephone and postal recruitment is used: '*Every which way possible, you name it, we have it. And trying to keep the sample in fairly good shape, that's a constant battle*' (ibid).

Online recruitment requires the identity of respondents to be checked as, on the internet in particular, people can easily pretend to be who they are not. GfK check all IP addresses to ensure, for

---

[200] See from p60.

[201] Desk research is a term used for 'secondary research', i.e., collation and presentation of existing research rather than conducting the primary research from scratch.

example, that there is not someone with multiple accounts just trying to get more entries to the prize draw that they run to encourage responses (ibid).

A further consideration is that those who are involved in a panel survey may not, by implication of their very participation, be typical listeners (Webster, 2001, p925). Lock (2011) describes his view of those who are prepared to undertake audience research in colourful terms:

> *The type of dipsticks that turn up for audience research, the kind of mouth-breathing window-lickers that turn up for a bit of pizza and a tenner are not indicative of the audience you're trying to attract anyway. And say, for example, you've made a witty, clever show, the audience for that aren't going to turn up.*

ii Demographics

There is evidence that different demographic groups tend systematically to give differing average AI scores,[202] and this could be considered to be a limitation in regards to obtaining an absolute measure of appreciation. For example, older, lower socioeconomic status women have been found to be the highest raters.[203] This implies that the AI score for a programme could be influenced purely by the composition of its audience. If a comedy programme attracts an audience skewed towards older, DE[204] women, it may intrinsically garner a higher AI than another programme of the same quality.

iii Sample Sizes

While around 4,000 respondents of the Pulse sample may give AI scores for radio shows per day (North, 2010), for an individual programme the sample size is often too small to be of use. AIs are highlighted as being potentially invalid if there are fewer than 50 responses.[205] For the 23:02 comedy slot in particular, the sample size seldom breaches this boundary as there are relatively few people listening at that time in the evening and thus fewer of the sample to respond. Often, a reasonable sample can be garnered if the programmes run as a series as there could be a large enough sample when considering the four or six episodes as a total. This, however, means that no trend can be seen across the series.

In some respects, this late-night slot is particularly important in being able to judge audience reaction, as it is here that more experimental comedy is likely to be found. It is always possible to get a reasonable sample size for an episode of *JAM*, as it TXs where it will attract a big audience and hence a big sample size,[206] but a series that has been running for 45 years might not be the programme where the Commissioning Editor has most need of seeing the AI. Rather, a brand new show going out at 23:02, placed there to mitigate the risk should it prove unpopular, would be of great interest but would be unlikely to provide sufficient sample size to provide a usable AI score.

iv Respondent Interpretation

As introduced from p104, the appreciation question can be interpreted in a number of ways and this could be viewed as a limitation of the process. Respondents are asked to *rate* the programmes that they have heard. The respondent is thus able to interpret the criteria with which they are rating the show any way they wish.

In the list below are all the radio programmes that you listened to yesterday. Could you please rate each of these programmes with a mark out of 10, where 10 is the highest score?

---

[202] See from p110.

[203] See p105.

[204] A, B, C1, C2, D and E are NRS social grades used to denote UK class groups based mainly on the occupation of a household's chief income earner.

[205] In the BBC's Asteroid system.

[206] See p218 – JAM received an average of 112 appreciation responses for each of its 18:30 broadcasts during 2012.

Their 'rating' could be influenced by how funny they find the show, how good they thought the main performer to be, the overall entertainment value, the production quality, or any combination of anything else that the respondent considers to be of value (Ang, 1991, p145 and see from p127). It could be that the rating is disproportionately influenced by one aspect of the whole of the broadcast (Kahneman, 2011).[207] A respondent's memory of an occurrence is not necessarily true to their experience of it at the time. While they could be making conscious judgements about the programme, they could also be rating the show due to feelings based on subliminal factors.

> If two listeners approve of a programme, one because she "enjoyed" it, the other because it was "interesting", are they expressing different reactions or merely the same reaction in two different ways? And if the latter, is that reaction felt in equal measure by both? (Crisell, 1994, p200)

The key point is that all respondents may be using different evaluation criteria as it is a *claimed* appreciation rather than an objective or behavioural measure, such as how many times the programme made them laugh. As discussed earlier,[208] there are many cognitive effects that may have an impact upon how a respondent scores a programme, such as the 'Mere Exposure Effect' (Robert Zajonc – Kahneman, 2011, p66-67), which posits that people like things they are more familiar with, or the 'Halo Effect' (Moser and Kalton, 1971, p359), which proposes that perceived aspects of an entity can extend to the perception of a wider collection to which that entity belongs. None of these reactions is measured, however, as part of the Pulse survey. So, while psychological and physiological reactions could affect the respondents, generally researchers do not have any data to analyse this aspect.

As there are many interpretive aspects to appreciation ratings, their use as a basic measure is arguably limited: 'The AI score is a useful but crude measure of audience appreciation of a programme' (Sharot, 1994, p84).[209]

v Platform of Measurement

The way in which a respondent records an AI score could have an effect upon a programme's rating. The Pulse is an online survey and many aspects of how the survey is presented can alter responses (Best and Krueger, 2008, p231; Vehovar and Manfrede, 2008, p180, p183). This can be something as simple as the font used: for example an easily readable font can improve completion rates (Best and Krueger, 2008, p224), whereas one that is difficult to read can make respondents consider the questions more carefully (Kahneman, 2011, p65). In 2005 the Pulse survey moved from paper to online responses and for a two week period both systems were used. There was a difference found in the scoring (not attributable to demographic differences) but it was relatively small, with a 2% variance at topline (Van Meurs, 2008, slide 31). Over 90% of the mean AIs were within 10% of each other across the two methodologies. However, the online measure did result in a wider distribution of scores, interpreted as a greater sensitivity in responses (Van Meurs, 2008, slide 31).

In June 2009 North (2010) oversaw a GfK pilot survey which allowed respondents to enter appreciation scores from their mobile phones. It was found that all the mean AI scores for the chosen programmes were lower than the now established 'normal' online method. Reasons behind this were thought to include the poor usability of the survey via the phones, which impacted upon the actual rating. With ever-improving smartphone technology, it seems that this problem may decrease in time. It was found that 16-24s

---

[207] See from p111.

[208] See from p95.

[209] See from p132 for expansion of this aspect as a limitation of appreciation measurement.

were most positive about this form of survey implementation: a possible part solution to issues of memory[210] as responses could be more easily recorded closer to the moment of the experience.

vi Response Error

There are technical checks that can be done which can indicate that a respondent is not filling in the survey with any care or thought. Analysis can highlight where surveys are filled in very quickly, whether they are giving the same score for all the programmes within a day's listening – 'straight lining' – or whether they are giving the same scores across many days – 'flat lining' (Van Meurs, 2009b, slides 6-34). Where this kind of behaviour is seen, North (2010) claims that it prompts a communication to go to the respondent indicating to them that their responses are being looked at and that they should be more diligent in their scoring.

> *But at the same time, there's too much going on at any one time for us to drill down into any individual response. So generally the best way to tackle these things is by chopping out the statistical outliers, so, the people who do the survey fastest, it's a comparative to everyone else who did the survey. They're the people who you kind of think, well, if anybody's going through this survey and speeding, it's those guys, 'cos they're doing it faster than anyone else. And similarly, if there's someone who is constantly giving the programme 10 out of 10, are they really making any evaluation or is it just that they do love everything that they ever watch? And you start to think, well this doesn't sound really very likely. It sounds like they're not responding exactly to what we're asking. They're doing something else here. Therefore, we have to take a view as to whether or not this is legitimate within the scope of what we're trying to do, which is get an honest response to every programme that they've viewed or listened to.* (North, 2010)

The use of incentives is seen to improve survey response rates. A meta-analysis in 2006 indicated that use of 'material incentives' increases both the chance of a person responding and their retention within the survey (Fricker, 2008, p209). The exact nature of the ideal incentive is an area of discussion as there are a number of variables, such as the point at which the respondent should be incentivised and whether certain types of incentive are more suitable for panel surveys versus one-off surveys (ibid). There is a balancing act between the additional cost of the incentive, the benefit of improved response rates and retention on a panel (De Vaus, 2002, p136). GfK's incentive for Pulse panel members is the opportunity to be entered into a monthly £1,000 prize draw with the additional chance of winning some smaller prizes of £5 and £2 vouchers, provided they complete the survey on at least ten days in the month (ibid). The total expenditure to incentivise the panel is around £20,000 per month, around £1 per panel member per month. North (ibid) says that for younger respondents GfK were considering trialling a system whereby they are 'paid' each time that they filled in a survey, say 5p, that they could cash it in once they reached a certain threshold: for example, £2.50. This approach (used by YouGov in the UK) is hoped to improve retention, as people are less likely to give up after just a few surveys, knowing that they are losing money that they have 'earned'.

Although there are definite advantages to having a panel rather than recruiting a whole sample from scratch for every survey, there is a risk that people can be subject to '*panel conditioning*'. This can result in them responding differently from the population that they are supposed to represent (Fricker, 2008, p204), '*a serious threat to valid inference*' (Sturgis et al, 2009, p113). This has been a concern since the BBC began researching its audiences and remained so with its post-war surveys. However, Silvey's research (1974, p129) found that there was no appreciable difference in reactions to programmes dependent on the duration of panel membership. Today, GfK place no restriction on how long someone might be on the panel. In the UK, since few people stay with the Pulse panel for a long period of time, recruitment is more of a concern than panel conditioning. Though such conditioning has not been observed on the Pulse panel, North (2010)

---

[210] See from

explained how some long-term panel members can cause issues when evaluating responses, using their position to impart their beliefs rather than to evaluate programmes as is the requirement:

> *There's been some people who have been extremely vociferous, particularly about a single issue. They've been very unhappy about the number of repeats* [for example]. *They feel that whether the show is good or not, the fact that it's a repeat makes it bad. So they'll get zero…*[sic], *you know, one out of ten. And it's quite problematic because, they're like single-issue politics panellists. We have to respect that, because that's actually a view that is perhaps an extreme version of what many people feel and therefore we have to keep them on but at the same time they're behaving in a way that makes them outlying. And our quality control systems are picking them up all the time going 'odd person'. So, what we don't want is a vanilla panel, which just says 8 out of 10 for everything from everybody.*
> (North, 2010).

While changes from respondents over time can be attributable to panel conditioning, they can also reflect actual changes in attitudes. For example, trends may merely be reflecting changing outlooks of panel members as they mature or, indeed, changing trends in the general population (Sturgis et al, 2009, p114). Distinguishing between conditioning and trends is not easy.

Many general aspects within a survey and its application can affect the answers given by the respondents. Moser and Kalton (1971, p271) state that an important factor is that the respondent actually wants to be involved with the process. If they do not, they can be more likely to lie or give inaccurate answers. Taking account of this can help explain non-responses: if only certain types of people continue to be involved, this can result in 'non-response bias', a phenomenon that can be hard to quantify (Fricker, 2008, p198-200). Lack of effort in answering the question is a 'fundamental problem' in surveys (Bertrand and Mullainathan, 2001, p67).

Bertrand and Mullainathan (ibid, p67-68) summarise a number of issues related to cognitive processes in filling in surveys. Question order (relating to priming – Kahneman, 2011, p101-102), question wording, scales, and respondent effort are all variables that can alter the answers of respondents. Also length of the survey (shorter being better for retention and quality of responses (Best and Krueger, 2008, p223)), the anonymity of the respondent (Walliman, 2005, p282), and even something as seemingly innocuous as the font used (Kahneman, 2011, p65).

There are also theories relating to consistency of responses. It has been observed that there can be *little uniformity* for people rating the same thing if they experience it at a different time (Castleberry et al, 1994, p153). In contrast, another viewpoint regards an element of self-regulation leading to *over-uniformity*: 'People attempt to provide answers consistent with the ones they have already given in the survey' (Bertrand and Mullainathan, 2001, p67). This could mean that people are constantly comparing the ratings against ones they have given for previous shows, giving comparative ratings rather than absolute scores. If this phenomenon occurs, there is no guarantee that each respondent will respond in the same fashion as others which could lead to patterns of response for individuals:

> Response styles (RSs) are a respondent's tendency to respond to survey questions in certain ways regardless of the content, and they contribute to systematic error. (Van Vaerenbergh and Thomas, 2013, p195)

With all appreciation scores, respondents are asked to give a verbatim response alongside the score. Raphael, as Commissioning Editor for Radio 4 comedy, has most reason to study the figures and written responses. In doing this, she has found evidence of a very specific systematic error:

> *Only today I was looking at the Pulse verbatims for the 'News Quiz USA' and some people, and I've noticed this before, they tick the dislike box and then they give it a 10. Because they see 10 as being, 'I really dislike this' and suddenly you've got a much higher AI. And I've seen that happen quite often and I think that needs to be checked.* (Raphael, 2012a)

If it is accepted that humour is a genre that provokes extreme responses (see), this phenomenon may be a particular issue for Radio 4 comedy.

vii Context

As discussed earlier,[211] an important factor that could affect comedy AI scores is the context of the listening. For example, some studies have shown that comedy is more appreciated by listeners when in groups than alone (for example, Lieberman et al, 2009, p498). However, the current AI measurement does not record how the respondent was listening: whether in groups, if using headphones, or even whether listening to the radio as a primary or secondary activity.

4.4.3 Attention, Engagement and Motive

While the Pulse survey asks the supplementary question 'And how much of each programme did you listen to yesterday?' with answers ranging from '1 – Listened to hardly any', to '10 – Listened to all of it', the AI scores are presented without reference to this aspect. Those who give a score could be giving 100% attention or next to 0% if they are fully engaged in another activity and don't even register any of the content of the programme. In the second scenario, if they, as part of the survey, say that they have 'heard' the programme, they will be asked to rate it; their resultant score could then be based on criteria other than the content of the programme, such as their expectations of the show rather than the actuality of it. The survey does also ask of the respondents, 'And how much effort did you make to listen to each of these programmes?' but this is not exactly the same thing as attention or engagement, and will be a 'claimed' response rather than an objective measure.

Appreciation of a programme tells us nothing of *why* a programme is liked (Kent, 1994, p16): 'Feelings are a poor guide to the programme's effects or influence' (Crisell, 1994, p205). For example, a listener might 'like' a broadcast of ***I'm Sorry I Haven't A Clue*** as much as an episode of *Money Box* about mortgages, but if you're about to buy a house your reasons for liking them may be very different. Currently, research attempting to understand the 'why' is only approached rarely in BBC radio, and only on an ad hoc basis.

4.4.4 AIs Are Not 'Absolute'

Programmes are only given appreciation ratings by those who listened to them. This seems an obvious statement, but the scores are likely to be relatively high as people will generally listen to programmes that they like (consumption bias)[212] so programmes are not being rated by those who are not choosing to listen. This means that the appreciation ratings for programmes are not 'absolute' and, in this light, according to McLean (2009) are an 'arbitrary' measure using a 'blunt instrument'.

---

[211] See from

[212] See

AI is interesting and sometimes quite tricky because you are only asking people who have seen [or heard] *that programme so it's not the most representative way of measuring quality or 'goodness'. But realistically, given we have limited resources, there doesn't seem to be a way of comparing programmes like for like with the same sample because you have to make people watch* [or listen to] *a programme to be able to rate it. And it doesn't really make sense.* (Collins, 2010a)

As Radio appreciation scores are only supplied by those who heard the programme, the composition of the sample can have a big impact upon the resultant AI. If the trend over a series is looked at, an increase in the AI score might be seen, purely due to people who didn't like the first show, and who gave it a low score, not bothering to listen to further episodes. Even if the people who continued to listen did not score any higher for further episodes, the resultant AI would increase as it is a mean score from all the respondents. In this scenario, improving AIs do not indicate improving content nor even improving appreciation for that content; it is merely a reduction in the proportion of 'non-appreciators'. Collins (2010a) theorises that lacking audience size figures with which to interpret the AIs, scores from earlier episodes in the series are more indicative of the 'goodness' of a programme. That is, towards the end of a series, those still listening are doing so because they like it and will give a high rating, whereas, at the beginning of the series, the score given by the (in theory) 'larger' audience is closer to being 'absolute' as it may consist of a wider listenership. This theory has merit but is based on an assumption that if listeners don't like a show, they won't listen. This may be untrue of much of the Radio 4 audience, particularly habitual listeners who may well just leave the radio on for hours at a time, not bothering to turn off or over just because one programme is not a favourite; this is considered as listener passivity or inertia (Crisell, 1994, p221; McQuail 1997, p82). In this scenario, improving AI scores throughout a series *could* be an indicator of improving appreciation; if the behavioural measure of listening figures is static through inertia, the AIs become the primary indicator (Crisell, 1994, p221). Lacking any supporting audience size information at programme level, it is difficult to identify which scenario is likely for Radio 4 comedy.

AIs have been used at the BBC since the 1940s and are well established in the psyche of the organisation (see p70). Due to the nature of the figures through which the AIs are expressed, there is a tendency for those talking about AIs to use terminology that indicates them to be an absolute rating.[213] For example:

- Respondents rate programmes on a scale 1-10.
    - Users tend to express this as 'marks out of ten' but this is not correct as there is no opportunity for respondents to select zero as a score. This incorrect terminology is adopted by industry experts and academics alike; for example used by North, 2010 and Kent, 2002, p253. This error implies a level of veracity that is not there, as there is an implicit meaning that the data is based on a ratio scale if there is a zero point.
- AIs for a programme are the mean of the responses multiplied by 10.
    - Again, while the lowest score that a programme can achieve is 10 rather than zero, AIs tend to be expressed as marks 'out of one hundred'. This is how it is expressed even on the BBC Gateway research intranet (BBC Gateway Audience Portal, 2011) and, consequently, in the press (Neilan, 2012). As AIs are so often described as being 'out of' one hundred, they are implicitly seen as being a percentage even though this is not the case; percentages again imply that data is based on a ratio scale.

---

[213] See p139 and p286 for a discussion of this issue.

In this respect, appreciation scores and the calculated AIs are expressed as being a little more robust than they really are. They are merely a comparative figure which could just as easily be based on marks from 1-5 or 0-20. Collins (2010a) agreed that within the BBC AIs are often viewed as percentages and that technically they are not. However he went on to say that it doesn't really matter as, '*we're more interested in comparing one programme to another*' than getting to an absolute base.

> *You've got this currency; it works as a means of communication. With something so established, a lot of people will forget how it's actually calculated, in the same way when working in share or reach or ratings.* (Rumble, 2010)

This means that while appreciation ratings may be statistically reliable as they have a strong consistency over time (Gunter, 2012, p240), they are not necessarily subject to strict validity.

4.4.5 Comparison

If AIs are not absolutes and are only valid as comparative figures, it is also recognised that these comparisons are, in turn, extremely limited.

> Interpretation of AIs should always be done within a relevant and appropriate context. In other words, rather than comparing AIs across networks or types of programme, we should compare with similar programmes or note trends in the same programme over time. (BBC Audience Research, 2009b)

> Owing to the quite considerable difference in output from one radio station to the next, and the fact that they are aimed at different audiences, it is not advisable to compare AI scores between radio stations. (BBC, 2012b, p7)

If this is the case, then the huge range of programme types heard across Radio 4 alone implies that little comparison can be used across different genres. For example, even within Radio 4 programmes a radio comedy show would not necessarily be expected to achieve a score within a range that would be expected by a current affairs show. For the respondent, the score that they allocate to programmes of different genres is thought to be based on different 'utility' criteria, so a comedy may be given a high score because it made them laugh, but this would not be the case for current affairs (Collins, 2010b; Rumble, 2010). Even within a genre, there is cause to be cautious in using comparisons. Sub-genres of comedy may attract different expectations of what might be considered a good score. Raphael (2012a) is careful to take this into consideration when evaluating performance. For example, when looking at a brand new comedy: '*I compare them to other things in that slot. I do not compare them to **The Unbelievable Truth** or **Clue** or **JAM**, so I find other things to compare them with. So, it's not an absolute.*' When using aggregate AIs to give an indication of station performance, the resultant figure may be at best an indication of the mix of programme genres constituting their content, at worst the mean of apples and oranges (or, in the case of comedy, perhaps bananas?).

4.4.6 Relationship with Quality

James Holden, Director of Marketing and Audiences for the BBC, posits that *all* audience research has quality of content as the most important measure: '*...any kind of research and any strategy, it comes down to delivering high quality content to a variety of audiences*' (Holden, 2010). Quality of content has long been a requirement of the BBC, with Reith's original manifesto calling for 'the maintenance of high standards' (Shingler and Wieringa, 1998, p16). Today the BBC still considers the quality of its programmes as a significant measure of its delivery: 'Licence fee payers rightly expect their investment in the BBC to offer distinctive, high-quality programmes, and at the same time deliver real and sustainable value for them and

their families' (BBC, 2011b – pF3). Quality evaluation, at both corporate and programme level, is primarily through appreciation ratings (AIs).

Programme appreciation is often accepted simply as a gauge of programme quality but it can be argued that this may not be the case.

i Appreciation Correlates with Quality

The BBC's use of AIs in determining quality of radio broadcasting was strengthened when Tim Davie (2010)[214] announced that an aggregate of the AI scores for all radio programmes was to be a new topline measure of quality. He explained that: 'There's been lots of bright people who've worked out that the correlation between that [total AI] and quality is pretty good'. One of those bright people, Collins (2010a) explained that indeed, '*Quality is very strongly correlated with appreciation to the point where <u>AIs are a meta score for quality</u>.*' However, Davie appeared not to be keen on reporting total AI as a measure and he introduced the concept with an admitted '*heavy heart*', implying that the total AI score would be just one figure of sum of the parts which measure the success of a station, rather than the total arbiter of quality (Davie, 2010).

Total AI score has since become an increasingly visible figure as it has now been included in the BBC Audience Information pack that is available outside the BBC (BBC, 2013b).[215] This has been published since Q1 2011. For example, for Q4 2012, all the figures quoted for Radio 4 specifically (age 15+) were as follows:

BBC Radio Reach
Average Weekly Reach %	20.5%
Average Weekly Millions %	10.8%
Time Spent (per user) hh:mm	12:04

BBC Radio Quality Measures
AI out of 100[216]	80.7
Distinctiveness (agree %)	74.4%

While the aggregate AI is published for each of the BBC's television and radio stations, the measure is arguably only useful to identify trends rather than for comparison with other stations because:

> Owing to the quite considerable difference in output from one radio station to the next, and the fact that they are aimed at different audiences, it is not advisable to compare AI scores between radio stations. (BBC, 2012b, p7)

This guidance seems slightly perverse in the light that it is describing published data that includes an aggregate AI for 'All BBC Radio' (BBC, 2012a, p8), the implication being that while you can't compare an apple and an orange, you can take their average. Furthermore, it may be the case that an aggregate of the components of a channel might not be an absolute indicator of the channel as a whole. A study by Gunter et al. published in 1992 (Gunter, 2000, p143) found that programme aggregates were higher than 'global' ratings at channel level. This can be seen to a certain extent in current BBC measurement whereby an overall favourability rating for the BBC as a whole ('general impression'), based on a 10-point scale, just like programme appreciation, is much lower at 6.9[217] than *any* of the station aggregates (BBC, 2013b).[218]

ii Appreciation May Not correlate With Quality

The BBC needs to know if it is producing 'good' programmes, particularly over comparable time periods, as the requirement to understand quality is important for all types of business. A higher quality of product

---

214 Then Director of BBC Audio and Music.

215 See p126 and p260.

216 See from p286 for a discussion regarding the terminology used here.

217 Comparable to an AI of 69 as AIs are multiplied by 10.

218 See p126 and p260.

and/or services has been found to give a competitive advantage to companies; quality can be the single most important factor in affecting performance as there is a 'strong correlation between quality and profitability' (Lynch, 2006, p119, p509, p434). At certain points in production, quality can sometimes be measured objectively; a pair of jeans could be evaluated for how accurately they have been produced against the planned design, for example. However, a consumer buying the product will have many more subjective criteria that may be specific to them: the price, style, colour, design and brand could all have greater or lesser appeal to them and affect how they might define the quality. If the jeans are a fabulously stylish, perhaps the loose button may not be an issue. A purchaser can have a high *appreciation* of a product that may not be of the greatest quality. But comparably, can radio programmes have an intrinsic quality at all? There are elements that might be identified, such as technical quality of the audio or performance of the contributors. Did they make mistakes or stumble over their words? Even such specifics are still ultimately heard subjectively by listeners through their individual 'filters':

> Quality is a much over-used word in programme making. Is it only something about which people say "I know it when I see it or hear it, but I wouldn't like to say what it is"? If so, it must be difficult to justify the judges' decisions at an award ceremony. Of course there will be a subjective element – a programme will appeal to an individual when it causes a personal resonance because of experience, preference or expectation. (McLeish, 1999, p294)

In Ofcom's 2007 PSB review, an attempt was made to identify what was *actually* interpreted by audiences as quality. For news coverage, accuracy and impartiality were identified as the key measures. In drama, the script was most important. Documentary quality was assessed by the excellence of the journalism. Comedy was, unfortunately, not evaluated by Ofcom in this way (Rumble, 2010). These findings indicate that quality is based on 'utility', i.e., something is good if it is doing the job that the audience expects of it (Collins, 2010b). Thus, the ideal output of news is that it is giving you the accurate and impartial facts. If it were not to deliver this, it would be of poor quality. In relation to comedy, it is possible to assume that its main aim is to entertain and amuse. Mark Damazer (2010) says of Radio 4 comedy: '[its] *primary justification… is to make Radio 4 listeners laugh'*, Thus a comedy programme that is not humorous is not providing what it was designed for and is therefore of low quality (Collins, 2010a). However, funniness might not be the only aspect of a comedy's quality. Data from Benson and Perry's 2006 research[219] relating to humour in radio indicates that the correlation between the ratings of 'high quality' and 'humorous' were positive but not perfect, at 0.71. Thus, this may indicate that the utility of being funny might not be the only driver of a radio comedy's quality.

In 1941 as Robert Silvey was attempting to design the BBC's appreciation research, he pondered the relationship between quality and appreciation:

> The panel member was asked to use this opportunity to indicate the degree of enjoyment he had derived from listening (though it wasn't put in precisely these words). He was only to award ten if he enjoyed the broadcast so much that he could not imagine enjoying it more, and zero only if he detested it. It was listeners' subjective feelings we were trying to measure, not their objective assessments of the programme's merits. (Most panel members found this distinction incomprehensible – the programmes they liked were "good" and those they disliked were ipso facto "bad"). (Silvey, 1974, p115-116)

His hypothesis is that, for many, evaluation of quality is purely based on appreciation rather than being able to distinguish between the two, therefore indicating that respondents are not rating the quality of a programme at all.

---

[219] Unpublished detail, n=69. Both variables from responses on a 6-point numeric scale.

Recent research within the BBC has proved that quality and appreciation *are* directly linked: according to the BBC's published report on its audience research, '"high quality" is found to be one of the leading factors in determining an AI score for radio programmes' (BBC, 2012b, p7). This sentiment considers that quality is a factor of appreciation, linked but not necessarily the same. Since broadcasting legislation in 1990, an increasing need for quantifying quality was driven as applicants for terrestrial broadcast licences were expected to pass 'quality thresholds' (Gunter and Wober, 1992, p4). It is indicative of the difficulty in defining programme quality that the legislation 'provided no operational definition of quality' which could provide the threshold that was supposed to have been passed (ibid).

McQuail (1997, p57-58) discusses how the perceived quality of media is part of the whole experience, not *just* the programme or artefact. He proposes that there are many dimensions – some dependent on the specifics of the medium (p58) – that contribute to the quality of the experience including:

- The user's affinity to the medium – for example, one might not enjoy a radio comedy programme if one is actively averse to speech radio.

- The anticipated experience – if a listener has certain expectations that are not met, the quality may be viewed as low; it is a comparative measure. For example, one might not expect the same slickness of audio from a comedy podcast taken from a live radio broadcast on a minor radio station versus a polished comedy drama on Radio 4.

- Attention and involvement – some types of media demand attention meaning a high quality experience might only be achieved with high involvement. For example, reading a book needs attention whereas, at the other end of the spectrum, music radio could be utilised *particularly* for its low demands on attention. A radio comedy can fall anywhere along this line depending on the type of programme and how the listener wishes to use it, for example, for entertainment or just background noise.

- Appreciation – 'Quality can also be equated with high ratings of satisfaction' (McQuail, 1997, p57).

Ang (1991, p145) summarises the issue, writing that appreciation scores are unclear as they indicate merely a 'general satisfaction' which is just a lumping together of many varieties of appreciation. Regardless, amongst senior managers at the BBC, the correlation of appreciation has been deemed proven (Davie, 2010) and the AI score remains the primary measure of quality (BBC, 2012b).

Total AI for a station is now recognised as an indicator of general aggregate quality (BBC, 2011b). However, for Radio 4 in particular, this aggregate score is the sum of different genres: apples, oranges and bananas.[220] If, as an example, the mean AI of all the comedy shows together is 70 and that of current affairs is 80, does it really indicate anything to give an average of 75? Furthermore, the mean is weighted based on relative audience sizes so the mean could be 78, say, if the audience size for current affairs is higher.

> *What you find is that the overall AI score for a particular radio station or channel is largely determined by the bigger programmes because there are many more responses. So, Today probably does impact on the Radio 4 all* [aggregate] *AIs.* (Collins, 2010a)

Currently, there is not a specific target for the Radio 4 aggregate AI. In its measurement and dissemination, it becomes something that interested parties will look at as a trend. If, say, it began to drop, there could be a drive for Radio 4 to improve its aggregate score. To do so there could be an initiative to improve the quality of the programmes, but it could also be achieved merely by changing the *mix* of the programme genres. If

---

[220] See from p88.

comedy were deemed to have a lower mean AI than other genres then fewer comedy programmes could result in growth for the aggregate score. However, if the topline appreciation was managed in this way, and it was engineered that there were fewer poorly-rated programmes, would that mean that more of the audience would be satisfied more of the time? No, not necessarily. If the diet offers apples and oranges all the time perhaps the appetite relishes variety and could welcome the odd banana. In fact, Radio 4's service licence stipulates at least 180 hours of original comedy each year (BBC Executive, 2010, p35) so the *mix* of programmes has to be fixed to some extent.

4.4.7 Programme Lifecycle

In the genre of radio comedy, it has long been widely accepted that programmes need time to bed-in and be 'given a chance' (Briggs, 1979, p715). New comedy may be difficult to establish an audience for, and one could assume that for a mature audience, such as Radio 4 listeners, it might be particularly difficult: '*People don't like new comedy. Especially on Radio 4*' (Dare, 2012). The settling period may be the time needed both for the audience to become familiar with the characters and for those involved with the production to refine the content. The classic examples can be found in television where *Only Fools And Horses* did not gain popularity until its third series even though its characters were fully formed from the very beginning, and *Blackadder*, where its first series was far less successful compared to the following series which were completely different in style (Galton and Simpson, 2011b). This can be seen in radio too as, for example, the first episode of **JAM** in 1967 received an RI (as AIs were then called) of just 57 against a typical score for the slot of 68,[221] and feedback from a listener summarised as: 'the basic idea of the game, it was felt, was "dangerously thin" and likely to pall after a few hearings'. This aspect is considered when evaluating Radio 4 comedy AIs: '*Often the newest things, the most adventurous things do take time, so I can assure you we will not be judging on the AIs alone*' (Raphael, 2011). This would suggest that, contrary to Collins' suggestion that *early* AIs could be most indicative as indicators of appreciation,[222] for the comedy genre early AIs may be actually of very little use. 'New comedies tend to get higher AIs as audiences get used to the programme' (BBC Audience Research, 2009b). For example:

> Ratings proved stubbornly low for **Weekending** [*sic*][223] during its first year. There were also some crucifying reviews in the press. But those who did listen to it over the following eighteen months recorded a measure of appreciation that crept up slowly from a truly terrible RI of 38 to a much more satisfactory 60. By the mid-1970s, when it had become something of a minor cult among younger Radio Four listeners, and especially among University students, the audience reaction figure had risen to as high as 77. (Hendy, 2007, p77)

> *What really affects me is coming to something absolutely new. Not making allowance for taking time to get to like it. It took me three weeks to get to know* **Hitchhiker's**. (Reynolds, 2011)

For the appreciation scores of a specific episode to be most useful, there needs to be an understanding of where the programme sits in its lifecycle and the score needs to be considered in that context.

4.4.8 Programme or Series Specific

Respondents are asked to give an appreciation score for specific programmes. Due to poor sample sizes, series often have to be viewed across the aggregate scores for a whole series rather than one episode. This will give a mean across a number of episodes that is seen to represent the series as a whole. Is this

---

[221] See p134.

[222] See from p124.

[223] Actually known as **Week Ending**.

representative of how the listeners might consider the full series? It might be assumed yes. However, this may not be the case, as indicated by the following example. When ***The Spam Fritter Man*** from 1978 garnered particularly low AIs throughout its run, an audience research report was compiled which asked the panel for their response to the whole series rather than any individual programme. The audience researchers found that while the mean aggregate AI for the all the individual programmes added together averaged out at a poor 36, when asked to rate the series as a *whole*, the 749 respondents' score was calculated as just 28. This disparity indicates that a listener can perceive a number of individual programmes from a series differently from the series as a whole.

Figure 15 – BBC Audience Research Report – summary of ***Spamfritter Man***, 17/07/78 – BBC Caversham Archives
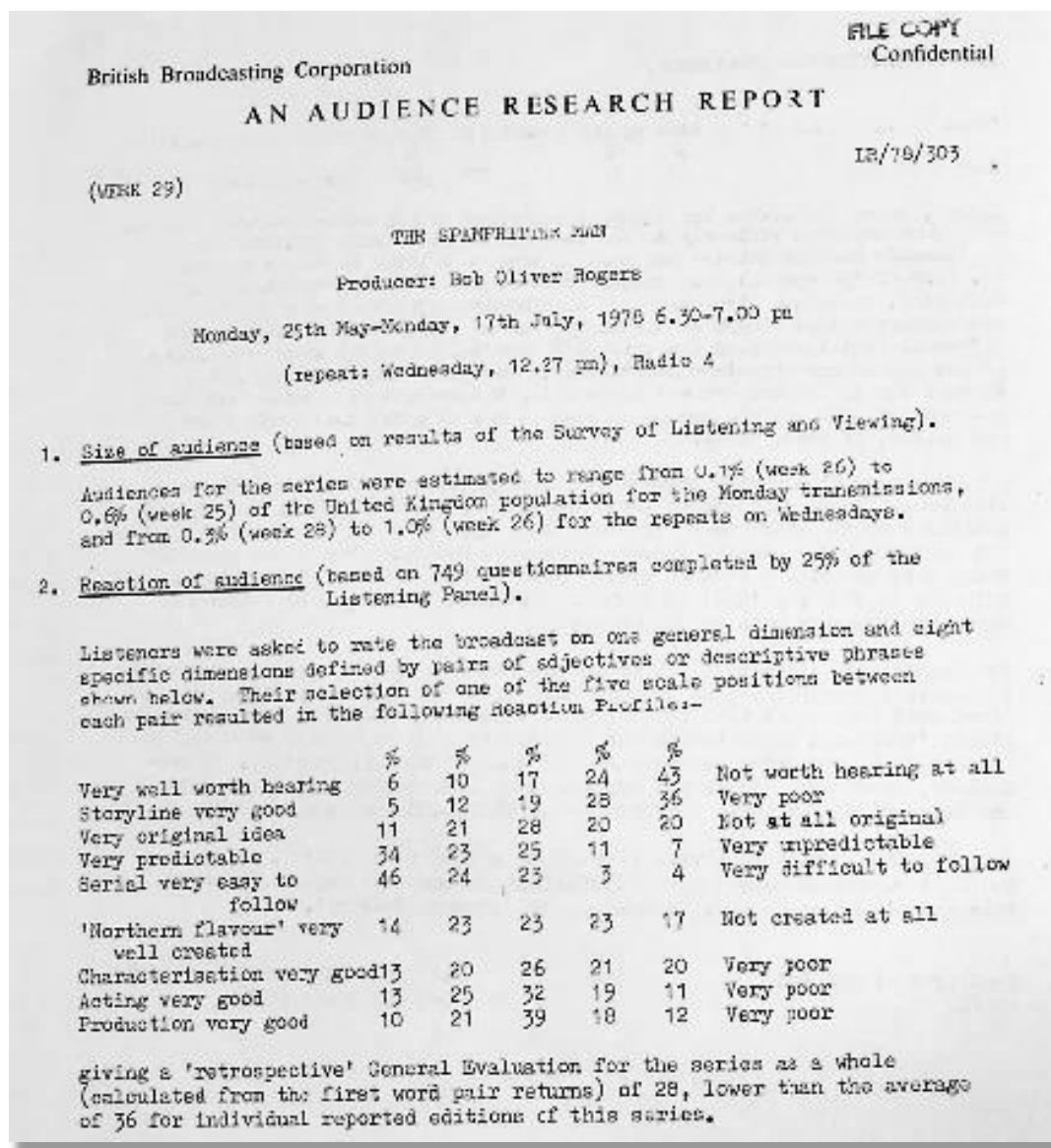
4.4.9 A Supplementary Measure

Data based on quantitative evaluations, such as AIs, is often seen as merely a 'supplement' to quantitative measures' (Kent, 1994, p16). For example, Menneer (1987, p241) describes AIs as 'a crucial and necessary "complementary currency of achievement" to the numbers'. But the issue for radio is that there are no overnights with which to temper the AI scores. Programme level AIs have nothing to be 'added to' as RAJAR information is only supplied at a quarterly timeslot level. But AIs are still used, for example, by the Radio Comedy Commissioning Editor to evaluate performance as, '*We have so little to go on*' *(*Raphael,

2012a). But, with no audience size figures: '*AIs on their own don't tell you a great deal… You shouldn't really be reporting AIs on their own*' (Collins, 2010a).

4.4.10 Elements of Appreciation

Expanding upon the consideration of respondent interpretation (see from p120) and whether appreciation and quality are the same (see from p126), the programme appreciation question is ambiguous and can be interpreted by the respondent a number of ways. Is a listener to radio comedy giving a programme a high rating because they find it funny, because they thought it high quality or because it was merely a good background noise to the primary activity in which they were otherwise engaged? Silvey (1974, p115-116) had long theorised that panel members could not distinguish between the elements that went into programme evaluation, a view supported by Ang (1991, p145) who writes: 'What exactly is measured here is not particularly clear: many varieties of "appreciation" are lumped together into a one-dimensional scale of something like "general satisfaction".' Gunter and Wober (1992, p3), albeit discussing television evaluation, postulate that appreciation may be immeasurable from a single question: 'Is it so complex that it defies measurement on even a wide array of scales?' Ang (1991, p145) highlights that this consideration, however, is not easy to account for in practice:

> The researchers themselves are quite aware of this shortcoming. For example, the Dutch did attempt to develop a multidimensional measurement instrument, breaking down "appreciation" into "informational value", "entertainment value" and "effort required by the viewer". However, this experiment did not lead to changes in the regular measurement practice because of high costs and stated difficulties in interpreting the findings.

Where individual elements such as quality and enjoyment have been compared, findings can be inconsistent. Gunter (2000, p142-144) cites studies by Wober (1990) and Savage (1992), the first of which showed quality evaluations to be lower, the second to show the opposite. He also cites a study by Johnson (1992), which found that when using a variety of dimensions (such as 'quality', 'liking', 'missability') a variety of scale types (Likert-style and numerical) and a number of scale points (5, 7 and 10-point) 'there was no indication that they represented qualitatively-different types of audience reaction'.

4.4.11 Verbatim Responses

While appreciation ratings may give an indication of programme appreciation, they do nothing to tell of the intent and listening situation of the respondents. However, appreciation ratings are part of the Pulse survey, sitting alongside other questions[224] that include open-ended queries regarding what the respondent particularly liked or disliked about a programme. Due to the volume of responses, the BBC researchers do not analyse these verbatims. 'The thing we're grappling with next is giving some summary illustration of the main themes from the comments. Rough and ready word clouds don't quite nail it' (Chant, 2009). If any progress has been made in this area, it has not yet been disseminated.

Verbatims are employed by exception as a useful source of qualitative information. For Radio 4 comedy they are used by the Commissioning Editor (Raphael, 2012a) as '*mechanical analytics*'. If an AI is exceptionally high or low respectively, verbatims could further inform the understanding of why.

> *We say 'Do you have any comments to make about this programme?' That's quite useful because people don't have to fill this in but if they really dislike or like something they tend to fill it in more. So if something gets a really low score or a really high score, you get more detail about it. If something gets an average score, a, you're less interested and b, you're not really too worried about that really.* (Collins 2010a)

---

[224] See from p85.

Whether or not they're useful in their current form is debatable. Many parties involved in radio comedy production appear, either by chance or by design, to be unaware of their existence:

> *I don't think I even realised that they did it. I probably would be quite curious with the verbatims. They'd probably just be interesting but they might help a little bit.* (Dare, 2012)

> *Might be fun but I imagine it would be really depressing.* (Doody, 2011)

> *It's depressing. In comedy it's particularly depressing. I wouldn't choose to expose producers to reading them by and large because it's so knee-jerk and so vehemently expressed quite often. You get those polar extremes and you know those things exist but whether you need to be reminded in print that that's the case is another issue.* (Canny, 2011)

While BBC researchers are as yet unable to disseminate useful analysis to verbatim responses, there may be another way that could be utilised. If it is true that people write more words if they particularly like or dislike a programme, word count could be an additional measure of strength of feeling about a programme. This was mooted by North (2010):

> *Yeah and that was a fantastic show both in terms of size but also the AI. You could see that from every measure. In the verbatims. Everything. People were writing a lot more about it. You can just look at from the amount of words that people write about a programme as a good indicator of how engaged they are with it.*

4.4.12 Distribution

While earlier measurement related the scoring of the programmes to actual attitudinal statements – for example, A+ meant 'I wouldn't have missed this programme for anything', 'I can't remember when I enjoyed (liked) a programme so much' or 'One of the most interesting (amusing, moving, impressive) programmes I have ever heard' (Silvey, 1974, p116-117)[225] - AIs are now measured based on a 1-10 scale. Thus, when using a Likert-style scale previously, it could be said that x percent of the sample thought that a programme was very well liked, whereas now it would be stated that x percent gave it a score of 9 or 10. Relating the latter to an attitudinal statement is difficult.

As AIs are presented as a mean score on BBC Gateway and Asteroid web reporter, that is the format in which users, other than researchers wishing to investigate further, will see the data. However, the data is actually based on a series of scores from 1-10 and, in theory, the distribution of these ratings could vary greatly even for two programmes with the same AI score.

In the old style Audience Research Reports, AIs were presented alongside the score distribution, allowing the user to see if the scores were tightly spaced, widely spaced or even polarised. For example, this is how the producer would have seen the information for the first episode of *JAM* (AIs were at this point called the Reaction Index):

---

[225] See from p70 for the full list of responses.

Figure 16 – BBC Audience Research Report – first episode of **Just a Minute**, 22/12/67 – BBC Caversham Archives


```
British Broadcasting Corporation                        Confidential

                AN AUDIENCE RESEARCH REPORT
                                                         6
                                                    LR/67/1638
(Week 51)
                        JUST A MINUTE

                        A  Panel Game

                   Produced by  David Hatch

      Friday, 22nd December, 1967. 7.30-8.00 pm, Radio 4


1.  Size of audience (based on results of the Survey of Listening and Viewing).

    It is estimated that the audience for this broadcast was 0.8% of the population
    of the United Kingdom.

2.  Reaction of audience (based on questionnaires completed by a sample of the
                          audience.  This sample, 246 in number, is the 17% of the
                          Listening Panel who heard all or most of the broadcast).

    The reactions of this sample of the audience were distributed as follows:-

            A+     A      B      C      C-
            %      %      %      %      %
            12     27     40     18     3

    giving a REACTION INDEX of 57.  The average for Panel Games for the first nine
    months of the year was 68.
```

This level of detail allows the producer to gauge how many people might have extreme reactions to the show. For example, here, only *3%* of the sample gave it the lowest mark of 'C-'. Compared with other first episodes this is actually very low – for example, 21% of the sample rated the first episode of **Week Ending** as a C-. Thus, despite its relatively low AI of 57, the programmer considering the performance of **JAM** might conclude that continuing with the show could be low risk as so few found it awful.

Since it seems generally accepted that comedy is a polarising genre,[226] looking at just the mean AI scores rather than any distribution will obscure the true nature of the audience appreciation. For example, two new shows may receive an AI score of 70, but these averages could be driven respectively by either all ten respondents giving the show a rating of 7 *or* half of them rating it 10 and the other half giving it 4. Superficially, the AI is the same, but looking at the distribution gives a *very* different story. The potential polarisation of these scores is not currently considered by the BBC, though it used to be:

> If the audience's reactions were sharply polarised; if one half were ecstatic and the other half bored to distraction, the arithmetic average of the marks would describe none of them. We kept a sharp watch for such cases but in fact they were very rare; the curves of the distribution of marks seldom had two humps (or, as the statistician would say, were seldom bi-modal [*sic*]. (Silvey, 1974, p116-117)

While in Silvey's time atypical distributions were seldom found, this was based on a 5-point Likert style scale. The 10-point scale now used might drive different results.

The distribution of the appreciation scores is an area that invites further discussion – expanded upon in the following chapter.

---

226 See from p48.

4.5 Summary of BBC AIs, Key Points Relating to their Measurement, Use and Limitations

Measurement

- BBC programme appreciation ratings given on a 10-point scale (1-10) and, for radio, are based on recall of a programme that the respondent has claimed to have heard for 5 or more minutes.

Use

- AIs are the primary source of objective programme evaluation used to aid commissioning decisions.

- BBC AIs can give appreciation responses for programme level detail as well as aggregate scores for genre or network for example.

- AI scores tend to be used mainly when they support existing understanding.

- There may be potential to utilise AI responses in order to estimate audience size.

- There may be potential to utilise verbatim responses in a quantitative capacity to enhance appreciation analysis.

- AI scores are made available to BBC employees but solely as a mean score, with only those in research roles having extended access to the data. Independent and freelance comedy producers do not appear to have any access to the AIs.

Limitations

- AIs are used by the BBC to indicate quality but there appear to be many factors that affect appreciation over and above solely the quality of a programme. Appreciation may reflect many dependent and independent variables related to the listening experience.

- Appreciation responses to radio comedy, as a polarising genre, may be poorly represented by a mean score.

4.6 Further Evidence Supporting the Suitability of Using Radio 4 Comedy to Analyse Appreciation Responses.

- AI's are used in Radio 4 comedy commissioning decision-making.

- BBC radio relies in particular on AI's in assessing audience preferences because of the lack of overnight listening figures.

- If the hypothesis that comedy is a particularly polarising genre (see chapter 2) is true, there may be evidence to suggest that the supposed objectivity of AIs in aiding comedy programming decision-making - it being a particularly subjective genre in need of objective information - may be compromised.

Chapter 05 – Distribution of Appreciation Scores

There's a risk that this show will only be funny to those of us that are here.
At worst it alienates a million or so.

(Mark Watson, *Mark Watson's Live Address to the Nation*, BBC Radio 4, series 1, episode 1)

This chapter theorises how BBC Pulse radio comedy programme appreciation scores might be distributed by the respondents. Firstly, we discuss that this is an area that is currently not considered within the BBC even though academic theory might suggest it an area worthy of consideration. As the appreciation data is collected on a 1-10 scale, we discuss types of scale and the levels of analysis suitable for each. In light of these scale types, we go on to deliberate whether attitudinal evaluations can truly be measured on a linear numeric scale, and we look at some examples of earlier commercially-based research, going on to propose motivations beneath distribution shapes. Then, in light of the proposal that attitudes might not be collected on an interval scale, consideration is given to the use of the mean as a representation of a typical rating, finding that it may be inadequate as a sole representation of the data. These discussions lead on to the presentation of the paper's hypothesis and three research questions.

5.1 Background

The BBC's broadcasting research team was once deemed to be one of the largest of its type in the world (Briggs, 1995, p21), and as such was able to scrutinise data at programme level and even look into the distribution of the appreciation scores (Silvey, 1974, p116-117), which was included in programme reports for many years.[227] Nowadays however, restrictions on the research teams means much less time can be spent in this area: '*There's not a lot of statistical analysis at this level going on*' (Collins, 2010a). AIs are presented to the BBC users almost exclusively as a *mean* score of these ratings regardless of the distribution of the scores, the AI being a figure that has been the main proxy for BBC programme quality for decades.[228] However, the mean is a figure that, according to conservative statistical views, can only be utilised with data that is distributed in a 'normal' or unimodal 'hump-back' spread (Ehrenberg, 1986). If programmes incite divisive responses, a single summary score might be an inappropriate representation. With a widespread view that comedy as a genre is particularly polarising,[229] might responses to humorous radio shows actually be obscured by expressing them as a mean? And when evaluation of a programme's 'goodness' rests significantly upon its AI score – as it does in radio where there are no overnight audience size figures[230] – the need to interpret them correctly has particular importance. Gunter and Wober (1992, p102-103) propose that looking into the distribution of appreciation scores may give insight into programme performance as the 'range of variation in appreciation scores can itself vary considerably across different programme types', and 'the most significant markers are often provided by scores at the extreme ends of the satisfaction-dissatisfaction'.

---

[227] E.g., see p134.

[228] See from p70.

[229] See from p48.

[230] See p81.

5.2 The Scale

Survey scales are designed to do the following (Low, 1988, p69):

- provide a set number of answers to a question (except for continuous scales).
- focus the respondent to provide relevant answers.
- provide uniformity of answers for analysis.

BBC Radio and television appreciation is currently measured on a 10-point scale of 1-10. In the very earliest panel appreciation research in the late 1930s, Silvey first set up the scale to run on an eleven point scale 0-10 on the basis that people would be familiar with such a scale, 'perhaps because, remembering their school days, people tended to start off with ten and then knock off marks for things they didn't like' (Silvey, 1974, p116). This system was updated and for many years measured with Likert-style questions rather than a numerical scale, using a 5-point A+, A, B, C, C- scale.[231]

Does the scale of measurement really make much of a difference to the outcome? Many different types of rating scale are used to evaluate media, for example IMDB[232] imdb.com uses a 1-10 star rating for people to rate films they have seen, while YouTube simply relies on a dichotomous system of thumbs-up or thumbs-down for a rating system. For comedy in particular, Chortle.co.uk has a 0-5 star rating for its reviewers (although zero is very seldom used). Even in academic research, particular scales are often chosen with no real reference to how or why they are selected: for example, Benson and Perry (2006), where respondents were asked to rate radio humour on a scale of 0-10.

As the current BBC system uses a 1-10 scale, it is this that shall be the primary focus for discussion.

5.2.1 Categories of Rating Scale

Numerical ratings scales are usually categorised into four main levels: nominal, ordinal, interval and ratio (Treiblmaier and Filzmoser, 2009, p3). These categorisations are important as they dictate what intensity of analysis can be applied to the resultant data.

i Nominal

Nominal categorisation is merely the assignment of segments that have no order, ranking or interval between them (De Vaus, 2002, p362). There is no sense of there being a 'continuum', and while numbers can be allocated to each of the categories, any calculations taken from these numbers would not be meaningful (Moser and Kalton, 1971, p352). For example, researchers could allocate sub-genres to radio comedy programmes using the list found on the BBC iPlayer for radio comedy (17/7/12):

1 Character
2 Music
3 Satire
4 Sitcoms
5 Sketch
6 Spoof
7 Standup
8 Other Comedy

In this case, the nominal allocation of numerical values to each category do not indicate any kind of order. Even being in, say, alphabetical order, would have no relation to the numbers in terms of statistical analysis.

---

[231] See p75 for a summary of the dates of the changes in AI measurement.

[232] IMDB is the online Internet Movie Data Base which allows self-selecting contributors to review and rate films and television.

Thus, if one programme were to be categorised as '3 Satire' and one as '7 Standup', the values of each category are meaningless and any numerical calculations using them would be misleading. Evaluating responses for nominal data such as this would be limited to the chi-squared test, and most typically frequency distributions with recognition of the mode, i.e., how many times each category has been selected and which is the most popular (Walliman, 2005, p103).

ii Ordinal

Categorisation of variables can be considered ordinal where it is possible to place the categories in an order but we cannot quantify a meaningful distance between them (De Vaus, 2002, p362). Likert-style scales are ordinal (Jamieson, 2004, p1217), for example, the A+, A, B, C, C- ratings used to calculate BBC radio AIs for many years.[233] In terms of their allocation, each of the marks is attributable to statements of 'goodness', ranging between 'I wouldn't have missed this programme for anything' to 'I disliked it very much' (Silvey, 1974, p116-117). Neither the marks nor the statements can be seen to have a quantifiable distance between them. We can say that A+ is better than C- but not that A+ is x times better or x percent better.

As with nominal variables we might allocate numbers to the categories but, again, they would not be open to all statistical analyses. For the calculation of AIs the BBC researchers did in fact allocate marks, as per the example below. There was an assumption that each score was equidistant from those adjacent:

Rating	Attributed Score
A+	100
A	75
B	50
C	25
C-	0

However, these values are arbitrary and without proof (if proof were possible) of their true relative values, the distances could be of any magnitude. For example, the perceived difference between how someone might view an A+ and an A- could be very small while the *distance* between the relative *goodness* of a programme rated B versus C *could* be far higher. The attributed scores applied, as above, to each of the ratings can imply a robustness that does not really exist (Bryman and Cramer, 1994, p65). Of the early BBC appreciation scores Silvey writes:

> By giving numerical weight to each position on the scale… the answers can be shown as a single figure – the number of points scored. This indeed is frequently done, although the choice of the weight given to each position on the scale is crucial. It is often based on arbitrary assumptions about the right relationship of the various answers. It may not in practice be any worse on that account, but if it is felt necessary to establish scientifically what the weights should be, this can involve very expensive and time-consuming preliminary research. (Silvey, 1974, p67-68)

In terms of what levels of analysis are viable for such scales, in addition to that which can be applied to nominal variables, they can additionally be allowed interrogations such as the Spearman's Rho, or the Mann-Whitney U-test, median and percentiles, but not anything as rigorous as means or standard deviations, i.e., only non-parametric tests (Stevens, 1946, p678; Bryman and Cramer, 1994, p65, p119; Jamieson, 2004, p1217). This is the case even if the scale points are numerical.

---

[233] See p70.

iii Interval

Whereas ordinal scales place variables in order, interval scales additionally incorporate relative distances between the categories that are *not* arbitrary (De Vaus, 2002, p360). The resultant variables can be subjected to statistical analyses that are more 'powerful' than those that ordinal and nominal can be subjected to, such as the mean, standard deviation, t-test, and F-test (Walliman, 2005, p103): in essence any parametric tests (Bryman and Cramer, 1994, p119). The scale points are continuous, i.e., it is also meaningful to consider non-integer points along the scale. Furthermore, interval (and the following ratio) scales, are likely to produce a range of scores that are distributed in a unimodal humpbacked distribution (normal or approximating normal distribution): suitable for parametric analysis (p117).

While an interval scale may allow analysis to express relative distances, it does not allow it to claim *absolute* measures as, if there is a zero point on the scale, it is just arbitrary. Classic examples are temperature scales, such as the Fahrenheit scale (Walliman, 2005, p103). The distance between each of the temperature points are the same, i.e., the *difference* between 10 degrees and 11 degrees is the same as between 20 degrees and 21 degrees, but the position of where zero is on the scale is arbitrary and could be anywhere. Zero Fahrenheit was initially based on the coldest temperature observed in Iceland and later changed to the lowest temperature obtainable with a mixture of salt and ice. This means that on such a scale, we cannot ever say that a figure is a multiple of another; 20 degrees is *not* twice as hot as 10 degrees as in this scale zero is arbitrary. In an interval scale, the choice or even existence of a zero point on the scale is of no importance.

iv Ratio

Ratio scales include the relative distances between points but also involve the absolute measure, i.e., will have a true zero value. A ratio scale can compare both differences and relative magnitude (Moser and Kalton, 1971, p353). An example of this could be audience sizes. If we take, for example, the number of listeners to radio comedy as measured by RAJAR[234] (Q4 2013, 15+):

	Programme Reach
Wed 11:30 comedy	843,000
Wed 18:30 comedy	1,210,000
Wed 23:02 comedy	615,000

This data can be said to be ratio measured on a ratio scale. The *difference* between 1,000 and 2,000 listeners versus 10,000 and 11,000 listeners is the same. Also, since there can be zero listeners, 2,000 listeners can be accepted as twice as many as 1,000. In the case of the actual figures, we can say that (with the caveat that these are estimations based on a sample) 595,000 more people listen at 18:30 than at 23:02 (distance): nearly twice as many (magnitude).

Ratio scales allow the most rigorous statistical interrogation and can be subjected to everything that nominal, ordinal and interval can be, but also enable analysis via the geometric mean, the harmonic mean, percentage variation and all other statistical determinations (Walliman, 2005, p104).

---

[234]

5.2.2 Summary of Scale Levels

It can, on occasion, be difficult to assign data to its correct scale so Walliman (2005, p104) provides a summary. If:

One object is different from another, you have a nominal scale.

One object is bigger, better or more of anything than another, you have an ordinal scale.

One object is so many units (degrees, inches) more than another, you have an interval scale.

One object is so many times as big or bright or tall or heavy as another, you have a ratio scale.

5.2.3 Scale Used for BBC Appreciation Ratings

BBC appreciation ratings are, in theory, measured on an ordinal scale as there is no evidence that the perceived distances between the scales points (1-10) are equidistant. However, BBC AI scores, in the way they are analysed, are *implicitly* treated as at least an interval scale. Respondents are asked to rate programmes from 1-10 and a mean score is given to express the summary of all the responses. We can infer from this that the *assumption* is that, for example, the difference in appreciation for programmes rated 9 and 10 is the same as the difference in appreciation for programmes rated 1 and 2. As the scale has no zero point, we *cannot* say that a programme with a rating of 10 is twice as good as one with a rating of 5 (Moser and Kalton, 1971, p353). Data measured on interval scales is robust enough to be subjected to further statistical investigations such as the mean and standard deviation.

This research investigates whether BBC Radio 4 comedy appreciation ratings really are measured on a true interval scale, as is assumed by the current method of BBC analysis.

5.3 Are Attitudes Measurable on an Interval Scale?

Moser and Kalton (1971, p353) claim that attitudes tend to be measured on what are assumed to be interval scales. In the case of BBC AIs, this would seem the supposition based on them being universally expressed as mean scores. But if this were the case, it would have to be accepted, for example, that the difference in appreciation between scores of 9 and 10 is the same as that between 5 and 6, an assumption that is not necessarily valid, and arguably not provable. Bryman and Cramer (1994, p118) claim that attitudes can only be 'basically ordinal in nature' and should be analysed as such. BBC researchers, however, appear to make the assumption that it is acceptable to treat appreciation data as interval: 'But I confirm that, rightly or wrongly, there was no debate at the time [1980s] about using a different or more appropriate summary score than the mean score' (Menneer, 2014). Academic researchers in related areas have made similar assumptions.[235] Barwise (2013) – who conducted extensive analysis on television appreciation scores – accepts in hindsight that the scale from which the scores were collected was a '*discontinuous measure',* but used the data to present mean scores in spite of this retrospective acknowledgement.

There has been discussion about the use of parametric analysis (for example, calculation of a mean) being applied to data collected on ordinal scales for many years, (for example: Stevens, 1946; Knapp, 1990; Svensson, 2001). 'Treating ordinal scales as interval scales has long been controversial and, it would seem, remains so' (Jamieson, 2004, p1218). There appear to be two schools of thought: the first that rigidly adheres to the premise that parametric tests should *only* be applied to data that is at least interval in nature, and the second that argues that there *is* merit in subjecting ordinal data to more rigorous statistical analysis. Knapp (1990, p121) calls these schools 'conservative' and 'liberal' respectively. The complexity of the issue is

---

[235] See from p286 for examples of where appreciation scores are implicitly expressed as measured on ratio scales.

illustrated by Pagano (2001) in the 6th edition of his book *Understanding Statistics in the Behavioral Sciences*. Despite the hefty tome running to over 500 pages, when discussing this aspect of measurement he states: 'The issue, however, is too complex to be treated here'.

5.3.1 Conservative and Liberal Views

i The Conservative View – The Mean Should Not Be Used With Ordinal Variables

For purists, at least, there is no doubt that the mean is not valid if calculated from ordinal scale data:

> Methodological and statistical texts are clear that for ordinal data one should employ the median or the mode as the 'measure of central tendency' because the arithmetical manipulations required to calculate the mean (and standard deviation) are inappropriate for ordinal data. (Jamieson, 2004, p1217)

Treiblmaier and Filzmoser, (2009, p3) discuss how the mean is not valid for ordinal data as the distance between the variables is not necessarily meaningful as it is essentially arbitrary.

BBC AIs were calculated using the following arbitrary weighting for each grade for decades, with equal distances being given between each category:

Figure 17 – Weighting of scale points used for AIs – Actual Appreciation Score Distribution for the First Episodes of *Hancock's Half Hour* and *Week Ending*

Category	Actual Used Weighting	Score Distribution	
		Hancock's Half Hour Episode 1	*Week Ending* Episode 1
A+	100	10	4
A	75	23	14
B	50	38	32
C	25	22	29
C-	0	7	21
Mean		**51.8**	**37.8**

Thus, as shown in figure 17, there is an implication that an A+ is twice as good as a B as the weighting is twice as much: 100 versus 50. However, the attributed scores, according to Silvey (1974, p67-68), were indeed based on '*arbitrary assumptions*'. A view could be taken that A+ and A are actually very close in 'goodness' in the minds of the respondent and perhaps the same with C and C-. Without any additional information, it could just as easily be assumed that the categories could be weighted in a new, alternative way, as follows:

Figure 18 – Weighting of scale points used for AIs and example of alternative weighing – First episodes

Category	Actual Used Weighting	Score Distribution		New (arbitrary) Weighting	New Score Distribution	
		Hancock's Half Hour Episode 1	*Week Ending* Episode 1		*Hancock's Half Hour* Episode 1	*Week Ending* Episode 1
A+	100	10	4	100	10	4
A	75	23	14	*95*	23	14
B	50	38	32	50	38	32
C	25	22	29	*5*	22	29
C-	0	7	21	0	7	21
Mean		**51.8**	**37.8**		**52.0**	**34.8**

It is no surprise that changing the weighting could change the resulting mean; it has already been discussed that AI scores are relative values rather than absolutes.[236] The bigger issue is that for each of the programmes, the mean is affected to a different degree and in a different direction. For ***Hancock's Half Hour*** the new scoring system has *increased* the mean score, but just by 0.2 points, whereas ***Week Ending*** has seen its score *decrease* by 3.0 points. This shows that the weighting can have an effect upon different programmes to differing extents.

In this example, however, in both weighting systems, ***Hancock's Half Hour*** has the higher score, so programmers could make the assumption that it is the more appreciated programme. Could the weighting have an effect upon their *relative* 'goodness'? Take the following fabricated example:

Figure 19 – Weighting of scale points used for AIs and example of alternate weighing (fabricated example, 100 responses)

Category	Actual Used Weighting	Score Distribution		New (arbitrary) Weighting	New Score Distribution	
		Programme A	Programme B		Programme A	Programme B
A+	100	140	10	100	140	10
A	75	20	150	*95*	20	150
B	50	0	40	50	0	40
C	25	0	0	*5*	0	0
C-	0	40	0	0	40	0
Mean		**77.5**	**71.3**		**79.5**	**86.3**

Using the standard weighting, Programme A has a higher mean than Programme B by over 6 points (78 vs 71), whereas in the new weighting, it actually comes out far worse (80 vs 86). Programmers would wish to know which programme is more appreciated by the respondents and, if we were to use just the actual method of weighting used by the BBC, it could be concluded that Programme 1 is 'better', but the example shows that such a conclusion is not necessarily correct.

These examples show that where weightings are arbitrarily attributed to attitudinal ratings, there can be noteworthy differences to the mean scores meaning that this is not a robust analysis. As the mean is the primary output for AIs, one can argue that AIs have not been not valid expressions of programme 'goodness' as results can be manipulated based on the weighting. For Likert-style scales in particular, without the rigorous, 'expensive and time-consuming preliminary research' (Silvey, 1974, p67-68), weightings can only ever be arbitrary, and consequently means can be volatile measures. Merely weighting the categories equidistantly due to a lack of better information is no more or less valid than any other weighting. The issue is nicely summarised by Jamieson (2004, p1218), who indicates that this issue is also applicable even to numerical scales:

> Finally, is it valid to assume that Likert scales are interval-level? I remain convinced by the argument of Kurzon Jr et al [1996], which, if I paraphrase it, says, that the average of "fair" and "good" is not "fair-and-a-half"; this is true even when one assigns integers to represent "fair" and "good"!

ii The Liberal View – But The Mean *is* Used With Ordinal Variables!

Liberal views do indeed argue that ordinal data can be subjected to more rigorous analysis than is strictly dictated by the conservative purists. In applied research there is a temptation for this to be done (Svensson, 2001, p47), as many of the most useful summary measures are those that can only be done with interval level data (Walliman, 2005, p10), with the mean being one of particular prevalence (Ehrenberg, 1986, p3). Bryman

---

[236] See from p124.

and Cramer (1994, p67) note that there appeared to be a growing trend for the liberal treatment of ordinal data, with parametric tests 'routinely applied to such variables' (p118).

Knapp (1990, p121) puts forward the view that the 'liberals' are happy to consider differences in categories as equal in order for them to be considered as interval scales, and that there are researchers who have 'shown empirically that it matters little if an ordinal scale is treated as an interval scale' (ibid). In terms of the mean in particular, Ehrenberg (1986, p7) argues that although it may not truly be statistically valid if calculated from ordinal scales, 'it generally provides a useful mental or visual focus when looking at the data', as it is a figure that analysts and laypeople both understand. This is a concession that even Stevens (1946, p679) made to his conservative view where he accepts that 'for this "illegal" statisticizing [*sic*] there can be invoked a kind of pragmatic sanction. In numerous instances it leads to fruitful results'. One view posits that the statistical analysis is valid as, 'It has been suggested that parametric tests can also be used with ordinal variables since tests apply to numbers and not to what those numbers signify' (Bryman and Cramer, 1994, p118). Thus, even for the liberal view, the caveat is that the resultant analysis needs to be treated carefully, as the output is based on numbers, *not* the facts those numbers represent. Knapp (1990, p121) presents Marcus-Roberts and Roberts' view that 'it is always appropriate to calculate means (for example) for ordinal scales, but that it is not appropriate to make certain statements about such means'. There is a risk that the wrong conclusion could be made about the data if the significance of the figures is misinterpreted (Jamieson, 2004, p1217).

Part of the issue is that it is not always completely clear whether data can be considered ordinal or interval from the offset, and even in academic research, the matter is often not even seen to be a consideration:

> Generally, it is not made clear by authors whether they are aware that some would regard this as illegitimate; no statement is made about an assumption of interval status for Likert data, and no argument made in its support. (ibid p1218)

Because there is no 'rule of thumb' (Bryman and Cramer, 1994, p67) as to when a variable is ordinal or interval or ratio, the differing approaches continue to be utilised, however inappropriately.

5.3.2 Does Normally Distributed Data Imply Interval Level Data and Vice Versa?

While it may be difficult to identify where a scale might be truly interval, there is an argument that the distribution of the data can indicate the scale level. Likert himself considered his numerical scales to be interval *because* they yielded data that approximated a normal distribution (Treiblmaier & Filzmoser, 2009, p6). This position is not unchallenged; while some agree that if a variable is normally distributed it must be on an interval scale, others have found that this is not necessarily true (Knapp, 1990, p121). However, common practice shows that, 'If the assumptions of normality are met, analysis with parametric procedure can be followed' (Allen and Seaman (2007). Conversely, while interval level data is often found to at least approximate normal distributions, there is no guarantee of this; it could be any shape, such as flat, U- or J-shaped (De Vaus, 2002, p227).

5.3.3 Are Attitudes, in Particular, Generally Considered to be Interval or Ordinal?

'Most of the scales used widely and effectively by psychologists are ordinal scales' (Stevens, 1946, p679) and into the twenty-first century, continue to be so. 'Questionnaires and rating scales are commonly used to measure qualitative variables, such as feelings, attitudes and many other behavioural and health-related variables' (Svensson, 2001, p47). But it appears that researchers do this without recognition that attitudes *may* not be interval in nature and the resultant statistical analysis may not be valid:

It can be argued that since many psychological and sociological variables such as attitudes are basically ordinal in nature, parametric tests should not be used to analyse them. (Bryman and Cramer, 1994, p118)

In the behavioural sciences, many of the scales used are frequently treated as though they were of interval scaling without clearly establishing that the scale really does possess equal intervals between adjacent units. Measurement of IQ, emotional variables such as anxiety and depression, personality variables (for example, self-sufficiency introversion, extroversion and dominance), attitudinal variables, and so forth fall into this category. (Pagano, 2001, p27)

When asked about BBC appreciation ratings specifically McConway (2010) writes: 'These things aren't really on a ratio or interval scale – all we can really say with confidence is that (for example) if something is rated as 5, it's considered better than a 4. That is, the scale is ordinal.'

Thus researchers, if even aware of the limitations of attitudinal scales, '*don't always pay any attention to it*' (McConway, 2013). Hence, in the analysis of data from the BBC Pulse survey, appreciation ratings *are* treated as if they were on an interval scale. In calculating mean scores from the appreciation data, the BBC effectively *assumes* the linear relationship seen in the figure on p147. The limitations of this kind of problem *are* recognised when challenged (albeit not widely or explicitly) in BBC research and Collins (2010b) agreed that the means should not, technically, be calculated from appreciation scores. However, Collins explained that there was a need for information about programme quality to be universally understood using simple summary figures:

If you didn't it would become very difficult to have any kind of reporting scheme that made any sense for people that don't understand stats. (Collins, 2010b)

Pagano (2001, p27) claims that, for attitudinal statements, 'Many researchers treat those variables as though they were measured on interval scales, particularly when the measuring instrument is well standardized [*sic*]'. This is relevant to AIs as they are a convention of BBC audience research, and viewed as a standard because they have been in use since the 1940s. The presentation of the data has been well-established and regular users have no reason to question the validity. Even the few who have written about BBC AIs in academic theses, such as Menneer, 1987; Barwise and Ehrenberg, 1988; and Carrie, 1997, have failed to acknowledge (until questioned at a later date) any consideration of the nature of the data being less than interval, using mean scores and even standard distribution without reference to any issue of validity.

5.3.4 Scale Points

The scale of measurement used for BBC Radio AIs has changed significantly, but not often. When these changes happened, might they have produced different distributions regardless of the quality of the programmes? Dawes (2008, p1) claims that 'surprisingly' little research has been done in this area. His own study found that while 5- and 7-point scales produced no difference in results for mean scores and distribution, a 10-point scale *did* give a difference in the mean but not the distribution. He concludes that for surveys that are attempting to measure 'customer sentiment', such as programme appreciation, the choice of numerical scale is important as a 10-point scale gives a *lower* mean than 5- and 7-point scales (ibid). He also points out that when moving from one scale to another, if they give different means, it is possible to apply a factor to make the two easily comparable – something that may have been done as the BBC changed scales but, if so, appears not to have been generally recorded and/or archived.[237] When GfK first began to administer the Pulse survey, they simply chose the 1-10 scale with little validation:

---

[237] When queried in 2012, the BBC Archivist at Caversham was unable to find any records informing this matter.

That was what we set it at the beginning. Again that was a legacy of how we were doing it before in Holland. It was, I think, the way that it was done before by IPSOS. It was what was prescribed and what was felt to be appropriate... We did do some work on 1 to 5 because we used to get the kids to do 1 to 5. Quite why, I don't know. They thought that it was slightly simpler. But you lose granularity in that. And actually what you want is granularity, particularly at the top level. Because all the action goes on between 6 and 10 really. There's a very narrow spread of scores and so what you don't really want to do is turn that into fewer. (North, 2010)

Coelho and Esteves (2007, p313-339) discuss the merits of a ten-point scale compared to a five-point scale in evaluating customer satisfaction. They highlight the following points to consider while also indicating that there is a dearth of empirical work and little consensus of opinion in this area (p316):

- The most used number of response points is seven.
- It is generally accepted that fewer points limits the discrimination between responses.
- More points may be onerous on the respondent and increase non-response rates (although they did not find this in their experiment).
- The ideal number of points can vary due to a number of factors including:
  - The nature of the data collection.
  - What is being measured.
  - Who the respondents are.
- The number of points chosen may be influenced depending on whether a neutral point is desirable.

The conclusion of their experiment to compare 5- and 10-point numerical scales summarised that the two had similar non-response rates and mean scores, and were not affected by demographic differences. However, the 5-point scale, by virtue of having an odd number of points, produced a higher proportion of neutral responses. They concluded that for their specific study – on the customer satisfaction index in Portugal – the 10-point scale was 'a better choice' than a 5-point scale (Coelho and Esteves, 2007, p335).

Rocereto et al (2011, p53) found that the type of scale was a factor in response distribution; a cultural difference was found for a Likert-style scale, while this variance was not observed for a semantic differential scale (i.e., without descriptive categories between the extremes).

i Programme Appreciation Scales

Theoretically, different scales could result in different mean scores for ratings, but is that really the case in practice? For programme appreciation in particular, meta-research has found that regardless of the scale of measurement and, even regardless of the wording used in Likert-style questions, there is a consistency in the mean scores found for programmes across the world, albeit with notable exceptions:

> This 60-to-80 range for different programmes has been found repeatedly in Britain, Canada, the USA, West Germany, and elsewhere. Studies of white South Africans found programme appreciation scores averaging about 60 when there was only a single low-budget, bilingual, government-controlled channel. Conversely, a study of Welsh speakers in Wales found average appreciation scores of 80-plus for the (mostly English-language) programmes watched over a week, although this may have reflected poor translation of the liking scale into Welsh! These studies in South Africa and Wales seem to represent the two extremes found to date. (Barwise and Ehrenberg, 1988, p50)

Carrie (1997, p61-62) notes that others have come to the same conclusion on the steady reliability of programme appreciation. The positive aspect of this is that the consistency found means that any changes to the scale will not stop comparisons between results taken from either system. This is an important factor as there can be concern by users when measurement systems are updated (North, 2010). On the downside, attempted 'improvements' to scales to try to increase 'discrimination' within the 60-80 range have not succeeded (Carrie, 1997, p61-62).

When the BBC moved from an 11-point 'marks out of 10' scale to a 6-point Likert-style scale in 1990, work was published regarding the comparison of the two scoring systems (BBC Broadcasting Research Department, 1993, p17-25). For Radio 4 in particular, even using like-for-like programmes there was an observable difference in the results (ibid, p20) with the newer system giving a mean aggregate score of 13 or 14 points lower than the previous system: 65/64 as opposed to 78 previously.[238]

ii Continuous (VAS) Scales

The current system asks respondents to rate programmes on a 1-10 scale. This allows for just 10 *discrete* categories as the scoring system offers only integers to be selected. In 2010, North said that AIs could be considered for measurement on an analogue scale:

> *Now maybe we could have – because web technology is such that you could do this – you can have a slider. And then you could have a slider bar that's 1 to 100. And then you'd be able to score much lower than 10. You'd be able to score 1* [*sic* – a slider could register zero too]. *And we're looking at the impact of those sorts of measures at the moment.* (North, 2010)

A continuous scale question for radio programme appreciation ratings might be similar to this:

Figure 20 – Example of a Continuous Scale Applied to Appreciation Ratings
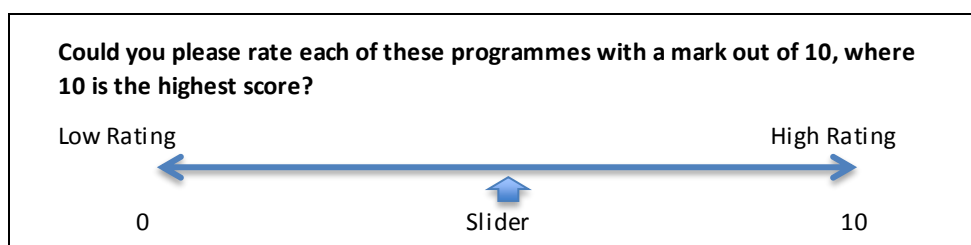
Continuous or Visual Analogue Scales (VAS) were first described as early as the 1920s, but were not widely utilised with pen and paper surveys as collating the data was time-consuming involving hand-measuring a point marked along a line (Reips and Funke, 2008, p699). Now, with the ease of use of computer-based surveys, the use of VAS scales is much more viable (Studer, 2011, p2). A general rule of thumb is that for scale responses it is better to have more categories than fewer, as categories can always be collapsed if necessary; if the data is collected on too few categories, it is not possible to go back and add granularity at a later date (Allen and Seaman, 2007). A continuous scale offers, in theory, an infinite number of levels and as such can be analysed after the data collection is complete in whatever categorisation group is most useful.

Reips and Funke (2008, p700) looked at scale responses on online surveys (the BBC Pulse Survey has been online since 2005) and compared radio button 4, 5, 7, 8 and 9-point scales with VAS. They found that the two types of measurement did garner different distribution results when the same questions were asked. They argue that VAS offers data that *approximates interval* level while discrete integer scales are only *ordinal* in nature (ibid, p699). They conclude that, whenever possible, a continuous scale is preferable to a discrete scale as the interval nature 'allows for applying advanced robust statistical analyses' (Treiblmaier and Filzmoser, 2009, p1). Overall, continuous scales have been found to have a number of advantages (Reips and Funke, 2008, p700-704; Treiblmaier and Filzmoser, 2009, p2):

---

[238] See from p75 – In 1992 the scale was reworked to be a 1-10 scale. A cynical observer might imagine that this may have been done because the 1990 6-point version gave a particularly low figure. Whilst, it appears to be the case that appreciation scores are not absolute, it may still have been a difficult task to justify the drop in scores, thus the reversion to a system resulting in higher means.

- Approximation of interval level data allows for robust statistical analysis.
- High level of detail recorded, allowing bespoke post-survey categorisation.
- Fine differences can be detected.
- Can reduce 'noise' in surveys.

Some studies have found, however, that VAS can have a detrimental effect on response time and completion rate, but Studer (2011, p7) theorises that this could be attributable to respondents being less familiar with this kind of measurement and may be eliminated as they become ubiquitous.

Although there seems to be evidence that VAS gives data based on an interval level scale, much of this kind of research has been considered for medical-based subjects (for example, Reips and Funke, 2008, p700) rather than media appreciation. Ideally, there needs to be validation before VAS can be adopted for programme ratings. Where, for example, GfK found that there was a big difference to the respondent between a 9 and a 10 (North, 2010), might a similar situation be found for extreme scores on an analogue scale? If one drags the marker to the highest point possible – 100% of the scale – does it hold a particular significance that anything below it might not come close to? Reips and Funke (2008, p699) significantly use the word 'approximate' when expressing how a VAS relates to an interval scale. Just as numerical scales instinctively *appear* to have more 'intervalness' than a Likert-style scale, continuous scales may *appear* to have yet more again, although this has not been proven to be the case for appreciation.

5.3.5 Relationship Between Scale Points

While it is clear that Likert-style scale weighting can be open to interpretation, such as the A+ to C-categories, is the problem solved when using the 1-10 numerical scale on which BBC AIs are now based? Likert-style generally takes its weighting from an arbitrary measure, often based on equal distances between categories, but for a 1-10 scale, is it not the case that the intervals between each of these categories is equal by definition? While it is indisputable that *numerically* the difference between 5 and 6 *is* the same as between 9 and 10, is it the case for the attitudes that these numbers are representing? For example, the Fahrenheit scale is an interval scale, as the distances between each of the integers is equal, but can we really say that of programme appreciation scores? Jamieson (2004, p1218) thinks not (see from p141), as all that has been done is to assign integers to represent 'fair' and 'good'. However, based on the fact that the mean *is* calculated from BBC appreciation scores, it is inherently assumed to be based on interval-level data. The following chart shows how the scores are *implicitly* understood to represent appreciation:

Figure 21 – Appreciation scores – assumed relationship to appreciation
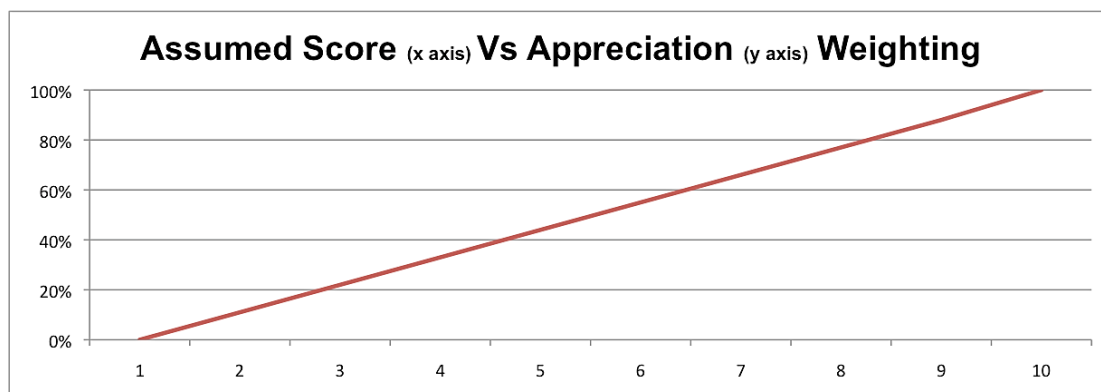
In this scenario above, as a programme is *x* more appreciated, the score increases in direct proportion to that appreciation. If this assumption is correct then statistical analysis up to that allowed by interval scales – including the mean – would be allowable. But is this assumption correct? Would a respondent hearing a BBC

radio comedy show have an understanding, or even an unacknowledged instinct, that the 'distance' between 5 and 6 is the same as between 9 and 10? One could argue that as the dimension is a simple 1-10 scale, people will easily understand that the distance between the integers is intended to be equal and will implicitly base their reaction on that understanding. Can this be the case, however, with something as subjective as appreciation? Even if we take something that is truly measurable, such as passage of time, which is measurable even up to ratio level (i.e., something could last 0 minutes or 10 minutes), human observation is such that there is variation in how time is perceived. With time, we could plot 'scores' allocated by respondents against measured times and check if they correlate in an attempt to validate perception (means and variances of time perception are discussed in Grondin, 2010, p526). But with a variable such as appreciation, this type of validation is not possible as it is fundamentally subjective and there is no objective 'quality' base against which it can be correlated.[239] So can we just assume that, in the 1-10 scale, appreciation is directly proportional to the scale rating as in the figure above as we cannot easily prove otherwise? GfK, who collect the appreciation data for the BBC, had considered this very point, particularly for the higher scores (unpublished data). North (2010) described how their findings indicated that there is *not* necessarily a directly proportional linear relationship:

> *We've done a lot of work on what an AI score means and the difference between 9 and 10 for example… So what we did there was to take into account that a 10 out of 10 is something that was much more significantly 'better' than an 8 out of 10 or 9 out of 10. So what we found was the likelihood to consider a programme better than another programme went up exponentially so that programme that you yourself had rated 10 out of 10 was considered always better than a programme that was a 9. Whereas a programme that you had scored a 6 and a programme that you'd scored a 5, sometimes you just thought the 5 was actually a bit better programme than the 6. So, we referred back to what you'd actually said, and said 'so which programme do you think is better?' we found, almost in every case, anything that someone had rated 10 was considered better than something rated 9.* (North, 2010)

With this knowledge perhaps the relationship can be redrawn as the chart below, with the difference between the central scores being of low relative differences while the extremes indicate relatively wider ranges:

Figure 22 – Appreciation scores – possible relationship to appreciation, number 1
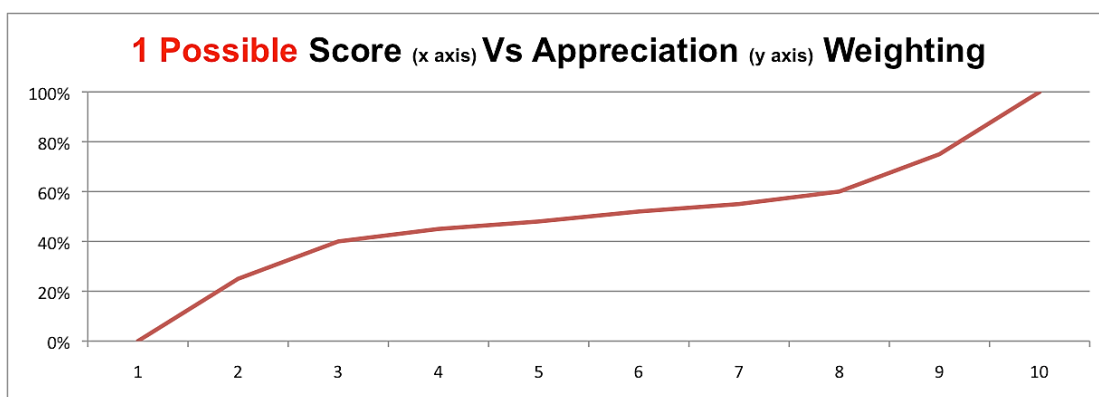
However, without validation, the relationship could in theory be any shape, such as:

---

[239] Even behavioural aspects such as audience size (which cannot accurately be measured for radio at programme level) may not provide validation as just because someone likes a radio comedy programme, it does not mean that they will always listen to it and conversely, just because they listen, does not mean that they like it.

2 Possible Score (x axis) Vs Appreciation (y axis) **Weighting**

The relationship could even vary for different people, at different times, under different circumstances and for different types of media. There may be innumerable factors involved in appreciation[240] that could impact upon this relationship.

> *There's some sort of bizarre cognitive model of what somebody might think… a kind of internal continuous scale of how good they think the thing is and it's nice, unimodal and symmetrical in their head. And what they do when somebody comes along and says 'give me a number on a 10-point scale' is that they have some internal transformation between that scale and the 10-points. The problem is, we don't know how that transformation works… What we don't know is if it's a linear scale… You can't tell what's going on.* (McConway, 2013)

Collins (2010b) claims that while it is possible to use qualitative techniques to understand the true relationships between numerical scale points, it is technically very difficult and problematic for the layperson to comprehend. This would effectively be a test for 'validity' which can generally be hard to verify, particularly for subjective evaluations (Moser and Kalton, 1971, p355-p356, Studer, 2011, p8)

This issue has been identified in the past when the Canadian Broadcasting Corporation, having developed routine audience reaction measurement since the mid-sixties, utilised non-equal weighting in the calculation of its EI (Enjoyment Index.) Using a 5-point Likert-style scale, the points were allocated thus: 100 = very much, 60 = quite a bit, 40 = all right/not bad, 20 = not too much, and 0 = not enjoyed at all. They had effectively given extra weighting to scores at the very highest end of the scale.

5.4 Further Factors Affecting the Distribution of Appreciation Ratings

While it remains arguable whether appreciation scores are truly *allocated* based on an interval level scale, there is another consideration that dictates whether data can be subjected to parametric level statistical analysis such as the mean. This consideration is related to how the scores are distributed. Even if one assumes an interval or ratio level scale for allocation of appreciation scores, there is no guarantee that the resultant data will provide a distribution that is well represented by a measure of central tendency (Jamieson, 2004, p1218). The issue is that parametric statistics are, for the most part, only valid if the data is distributed in a certain way, ideally normally distributed (Bryman and Cramer, 1994, p93).

The normal distribution is a specific shape of symmetrical, hump-backed spread that can be observed in the real world through measures such as height or IQ (Ehrenberg, 1986, p53). Bryman and Cramer (1994, p93) claim that the term 'normal' is actually misleading, however, as perfectly normal spreads are rarely found, even though researchers often assume normal distributions based on 'eyeballing' the shape

---

[240] See from p95.

of the data (Field, 2000, p38). This assumption is often made as the most common and widely understood statistical techniques 'presume' a normal distribution (Bryman and Cramer, 1994, p93).

While attitude ratings are not seen to be provably interval in nature, might it be the case that the BBC's appreciation score distribution is unimodal and hump-backed in nature and thus suitable for parametric analysis anyway? Hu et al (2009, p144) indicate that the law of 'large numbers' might imply that we would expect a normal distribution for quality ratings of media output, i.e., focused around a centre with few extreme scores. This theory would support finding a unimodal distribution. However, in their study of online film ratings for IMDB from US users, they found that the scores actually tended toward the extremes:

Figure 24– IMDB Ratings Distribution (Koh et al, 2010, p7)

Fig. 1. IMDB movies with average rating = 5 (92 out of 1000 movie items).
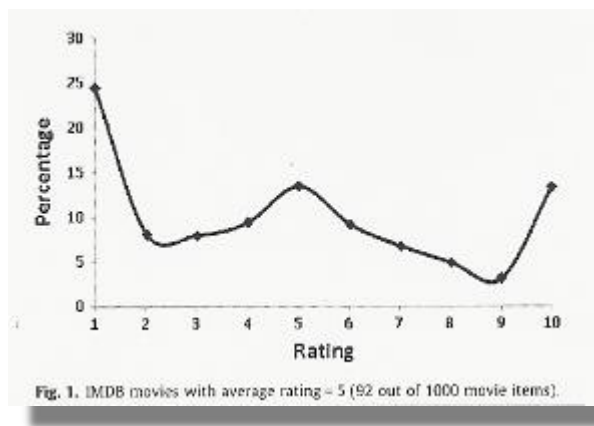
While this distribution illustrates a polarised view with a skew towards the negative, Hu et al (2009, p144) claim that evidence suggests that 'product review systems' for services such as Amazon, tend to garner ratings that are 'overwhelmingly positive':

Figure 25– Amazon Ratings Distribution (Hu et al, 2009, p145)
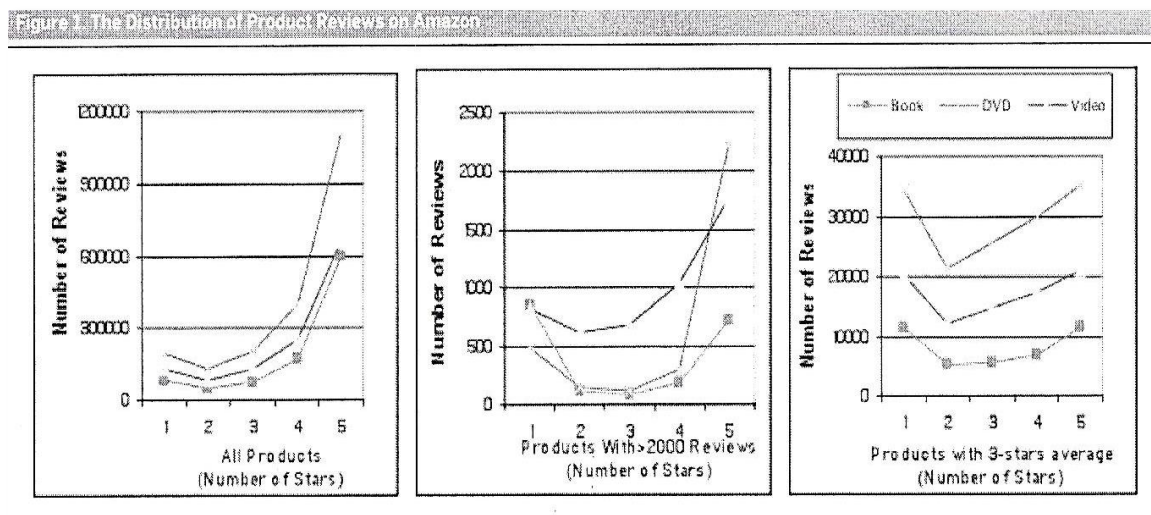
Furthermore, they state that for this 'J-shaped' type distribution 'the average is a poor summary of product quality' (ibid). They claim that product quality can only be illustrated by ratings if the full distribution is displayed. This view is supported by Sun (2012, p696) who cites a number of websites that are now 'making the distribution of ratings salient to consumers by putting up bar charts that demonstrate the percentage of reviews that are associated with each level of ratings'. For example, Amazon and IMDB also provide demographic information to subscribers:

Figure 26 - Illustration of Amazon Ratings Presentation –accessed 31/08/2013
(e.g., Hancock's Half Hour: The Very Best Episodes: v. 1 – Radio Collection)

Figure 27 - Illustration of IMDB Ratings Presentation –accessed 31/08/2013
(e.g., Hancock's Half Hour DVD – Volume 1)
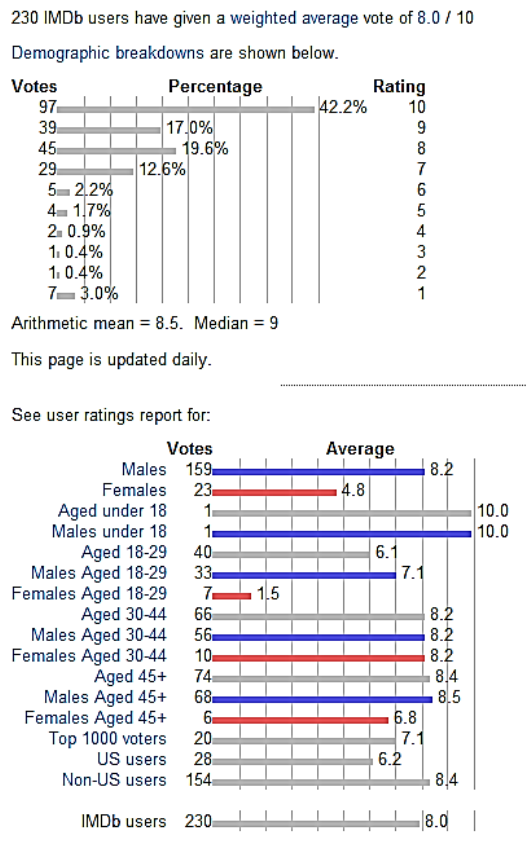
In terms of appreciation ratings, the general assumption regarding the shape of score distribution appears to be that the data approximates normal distribution, because the BBC offers just the mean as a representative figure for a programme. There is no attempt to describe the distribution, the implication being that the data does not vary from a normal distribution. However, there are a number of aspects of subjective evaluation that may explain why appreciation ratings may not be distributed in a normal or unimodal humpbacked fashion:[241]

5.4.1 General Effects on Attitude Distribution

i Under-reporting bias

For a distribution with peaks at the extremes,[242] Hu et al (2009, p145) and Koh et al (2010, p2-3) both posit an 'under-reporting bias'. This means that while there might be a full range of views on a reviewed item, it is only respondents that have extreme views who are driven to express them. This concept was discussed by

---

[241] See from p95 for general points on appreciation.

[242] See p150 for example.

Silvey (1974, p31) in regard to BBC unsolicited responses, which he illustrated with the example of the 1955 broadcast of Orwell's 1984 on television. The transmission inspired 2,375 letters, two-thirds of which were of protest and the remaining third in high praise. Silvey found that within the official panel of respondents who were representative of the population as a whole, a similar proportion to the letter writers had negative and positive responses. But there was a large proportion whose views were merely moderate:

> What was significant was that, taken together, these two [extreme] groups were only a minority. A majority of the panel members failed to react strongly either way but their standpoint had not been represented in the post-bag at all. (ibid, p31)

Koh et al (2010) consider under-reporting through the prism of Hofstede's 1980 work on the cultural dimension. They looked specifically at individualism and collectivism and found that when comparing movie ratings between the US and China, the US respondents were more likely to give extreme views while the Chinese reviewers' responses were distributed in a unimodal bell-shaped curve:

> Perhaps Americans are more honest and willing to post extreme views because they are less influenced by the mean. Another explanation is that Americans might try to be different by giving extreme ratings, since simply giving an average rating does not show that they are individuals. Alternatively, Americans may be less willing even to rate unless extremely motivated by very strong attitudes, positive or negative, towards the film. The Chinese, on the other hand, are demonstrably less likely to give extreme ratings, perhaps because they are more influenced by the consensus and the average sentiment of the reviews already posted or alternatively perhaps because they are not accustomed to express extreme emotion. (ibid, p7)

Koh et al found evidence of less under-reporting bias from the Chinese versus the American respondents (ibid, p9).[243] De Mooij (1998, p76 – citing Hofstede) tells us that western countries tend to be individualistic rather than collectivist and that within Europe, England is the most individualistic. In this respect we might expect to see more extreme views as an expression of individuality (assuming that findings on England are applicable to the UK as a whole).

In theory, under-reporting bias should not be present in AIs as the Pulse panel is designed to be representative of the population rather than self-selective, as people who write letters or rate products online might be. However, Pulse respondents are not expected to rate every programme that they see or hear, so we cannot rule out the suggestion that some people might be more driven to give responses to programmes that they have stronger views about. If this were to occur, we might expect to see an unrepresentatively high proportion of scores for Radio 4 comedies at the lowest and highest ends of the 10-point scale as respondents, moved to express their opinion, rated the programmes in question.

ii Consumption bias

For a J-shaped distribution[244] Hu et al (2009, p145) offer the theory of 'Purchasing Bias'. They state that only those more pre-disposed towards a product are likely to buy it and therefore likely to review it. 'Purchasing bias causes the positive skewness in the distribution of product reviews and inflates the average' (ibid) - also known as 'acquisition bias' (Koh et al, 2010, p2). This theory could relate to that of media consumption in that people tend to consume what they like; viewers 'quite like what they watch and watch what they quite like' (Barwise and Ehrenberg, 1987, p63). To cover purchasing and media together, we can refer to this under the umbrella term of *consumption bias*. Consumption bias indicates that there would be few ratings at the lower end of the scale as, with sufficient available variety, only shows that the audience members like would be consumed:

---

[243] NB: when discussing the limitations of their study, Koh et al do not attribute any differences found between the societies, to validity of the data which came from two different rating scales 1-10 numerical scale for the US versus 1-5 stars for China.
[244] See for example.

The idea of the marmite type programme, which some people like and some people hate, on television you basically only get those situations with forced choice. When there's free choice then the consumers who don't like something stop consuming it, or stop buying it… Were someone driving home from work, they get the comedy slot on Radio 4, they don't want to change channels for whatever reason, and so they try a comedy programme. It's possible. I'd be very, very surprised even in that situation under natural conditions if you actually got something with two humps. (Barwise, 2013)

This has been observed in programme appreciation research (generally relating to television); regardless of the type of scale or country of origin, the mean score tends toward the higher end of the scale. 'Ninety-five percent of all programmes will achieve an appreciation score between 62 and 82. This [roughly] 60 to 80 range has been consistently found in Britain, Canada, the USA, Germany and elsewhere' (Carrie, 1997, p129). He found that for UK television, 91% of ratings were positive responses and only 3% were negative (ibid, p79). His results, with n = over 500,000 individual responses, were as follows:

POINT ON SCALE SCORE – % OF RESPONSES
Extremely interesting and/or enjoyable –	20%
Very interesting and/or enjoyable –	38%
Fairly interesting and/or enjoyable –	33%
Neither one thing nor the other –	6%
Not very interesting and/or enjoyable –	2%
Not at all interesting and/or enjoyable –	1%

For Radio 4, the aggregate AI for the station is around 81,[245] which indicates that the majority of scores are also on the positive end of the scale. While radio appreciation is also on the increase, it is not on the same '*trajectory*' as television as there is, as yet, less time-shifting[246] and fragmentation[247] (Collins, 2010a).[248]

Consumption bias may have an additional tangential effect whereby the mere act of having made the choice to consume something could make one more inclined towards it. Danaher and Lawrie (1998, p55), referencing Barwise and Ehrenberg's earlier work, propose that there could be self-justification involved as, 'after all, only a fool would admit to not liking a program they had just spent one hour watching'.

Of course the Pulse survey will only measure responses to Radio 4 comedies from people who listen to Radio 4 comedies. In this respect, the resultant overall appreciation is for the consumers only rather than any absolute measure.[249] This consumption bias is an accepted element of BBC programme appreciation research.

iii Overconfidence and Popularity

Online reviews are given by contributors on the understanding that others will read and take into account their views. IMDB ratings are one example of such reviews.[250] Hu et al (2009, p145) claim that a polarised distribution might be observable if we were to consider that respondents might exaggerate their responses towards extreme ratings in the belief that other readers would be more likely to be interested in divisive reactions. It could be argued that this aspect would be inapplicable for BBC AIs as they are done privately by the respondents and not published to their peers. However, if we consider that respondents *would* be aware that BBC researchers may be using their responses, it could be theorised that they might give extreme scores

---

245 See p169.
246 Non-linear consumption of linear broadcasts.
247 Proliferation of broadcasting channels and platforms.
248 See from p104.
249 See p124.
250 See p151.

under the apprehension that this could make their ratings more noteworthy and, therefore, have greater influence on the whole.

iv Extreme Response Bias

Extreme Response Style (ERS) is a recognised tendency for respondents to either favour or avoid selecting extreme scores on rating scales *independently* of the item that they are rating (Greenleaf, 1992, p328), contributing to systematic error in the system (Van Vaerenbergh and Thomas, 2012, p195). Various studies have attempted to investigate two main areas: whether extreme responses are actually valid responses and whether the tendency varies for different demographic groups (Greenleaf, 1992, p328). Van Vaerenbergh and Thomas (2012, p198) discuss how the source of the response style can stem from either the stimulus (i.e., the survey) or the respondents themselves.

Findings have had conflicting results and there remains controversy around the subject. Various studies have been found to relate ERS to 'personality variables' such as age, education level, income, race and gender, but some studies have found no differences.[251] Van Vaerenbergh and Thomas (ibid, p196) cite a study that found that women were more likely than men to use the extreme response scores. The number of factors that can affect ERS are potentially huge, with studies on response style being conducted in areas that include number of scale points, modes of data collection, cognitive load, interviewer effects, topic involvement, language, culture, education, age, gender, income and personality (ibid, p200). ERS could also be a result of respondents attempting to manipulate results to adhere to a personal agenda (Webster, 2001, p925).

Hu et al (2009, p145) posit that extreme responses may in fact merely reflect the true nature of the items being rated and that people do in actuality have extreme tastes. However, they go on to say that this seems unlikely as it would imply that most products would tend to be '*outstanding or abysmal*'. On the basis that comedy is often thought to incite particularly extreme responses,[252] we might indeed expect to see extreme scores in our Pulse appreciation responses.

Contrary to these points, Koh et al (2010, p7) discuss how some prior studies have found that respondents have actually tended to avoid extreme scores in surveys. This behaviour is known as 'response contraction bias'.

5.4.2 Minimising Bias

Looking at the examples from IMDB and Amazon above,[253] none of them present a unimodal, hump-backed distribution. However, we must consider that they are self-selecting respondents. Hu et al (2009, p146) conducted an experiment where they asked respondents to rate a randomly-selected CD, rather than a product that the respondent might have chosen to buy. The experimental situation eliminated under-reporting bias and consumption bias. Furthermore, as it was not for a published review, potential issues with the respondents attempting to gain popularity were also eliminated. The distribution of the scores was then compared to the distribution found on Amazon:

---

[251] We have discussed how such variables may affect appreciation on a general level – see from p105.

[252] See p48.

[253] See from p150.
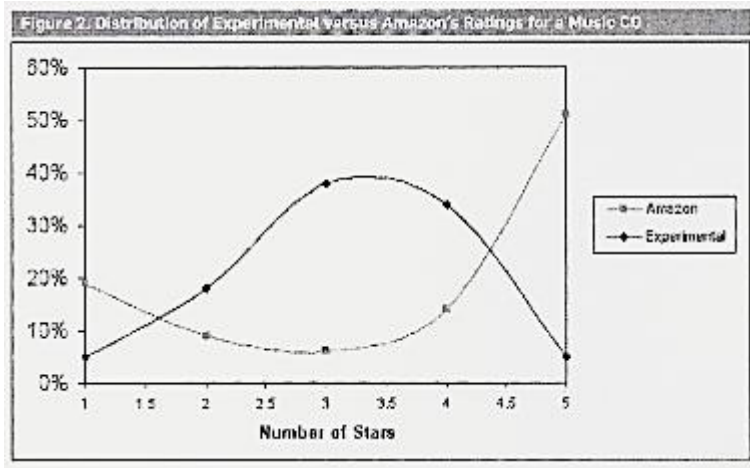
Looking at figure 28, it is apparent that they found in this experimental research that this item's reviews (on a 5-point star scale) followed a unimodal distribution, 'implying that most people have moderate tastes'. Just 3% of people gave the product 5 stars and 7% gave 1 star ratings.

Might comparable results appear if we were to conduct a similar experiment for Radio 4? Unfortunately, the scope of this research did not allow for a large, randomised group of people listening to Radio 4 across a range of genres and rating the programmes. However, as this research is ultimately looking at Radio Comedy, there were two datasets available that could be of interest.

Benson and Perry (2006) conducted research on the reactions of listeners to humorous extracts from a commercial radio show. Respondents did not choose what they listened to and their responses were not published, so the biases regarding popularity and attention were eliminated. The chart below shows the distributions of the scores given to the excerpts heard (N.B.: n=68, so the sample size is limited).

Figure 29 — Distribution of responses to 'quality' for humorous radio, n=68 (Benson and Perry 2006, unpublished data)
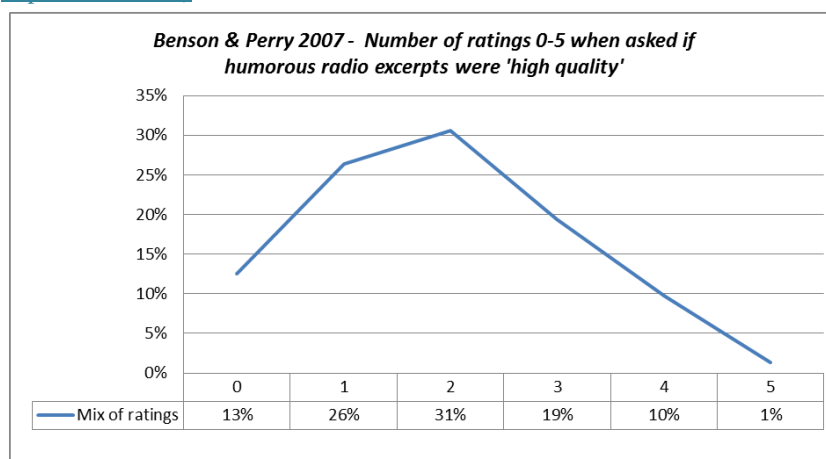
Thus, in a situation where the respondents did not select the material, figure 29 shows a unimodal distribution (albeit from a small sample size). Free from under-reporting, acquisition and popularity biases, the distribution approximates to a bell-shaped curve (albeit skewed) as found by Hu et al (2009).

Another accessible dataset were the responses from the judges of the Leicester Comedy Festival 2012 (277 responses from 11 judges across 160 different shows – the author being one of the 11 judges):
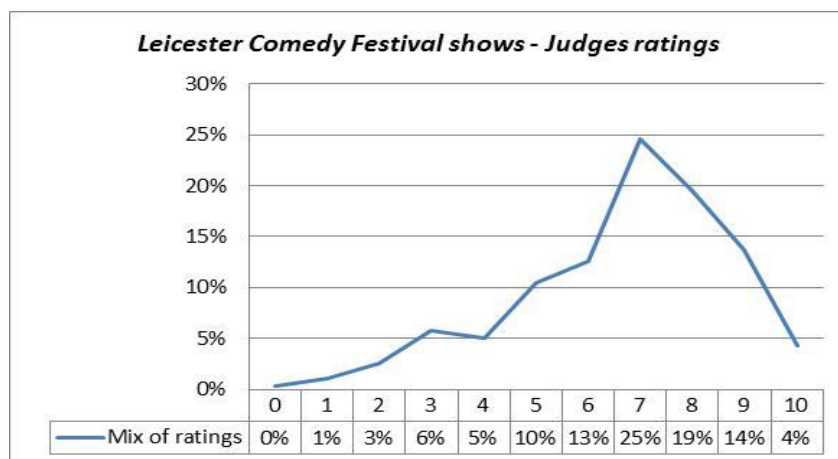
Again, with this data (figure 30), where the judges were told which show they were to attend and rate, thus
avoiding issues of self-selection, the shape of the distribution – while not by any means a perfect bell-shaped
curve – again tends to approximate a unimodal distribution; it is certainly not polarised. The role of the
judges in this exercise was specifically to be as objective as possible in their ratings. As well as this dataset
avoiding any self-selection bias, any hint towards 'overconfidence' or 'popularity' biases would have been
looked upon with a dim view from the other judges.

In relation to the theories discussed, we might expect any number of distribution patterns from those listening
to Radio 4 comedies. A unimodal, hump-back distribution might be expected if we were to consider that
respondents are attempting to be subjective, whereas, if we consider that people might only listen to the
shows that they like, or expect to like, we are more likely to see a J-shaped distribution. Furthermore, as our
respondents are UK citizens, we might not be surprised to observe a polarised response. If the analysed data
is distributed in a fashion other than that of unimodal humpbacked, theory dictates that expressing the data as
a mean would not be acceptable as there would be no proof of the data being on an interval scale.

5.5 The Mean

BBC AIs are a *mean* of the 1-10 appreciation scores given by the members of the Pulse panel for
programmes that they have heard or watched. The mean is an indicator of the data's *central tendency* or the
average, and is used to represent the data in one typical, simple, easy to understand figure (Bryman and
Cramer, 1994, p82). 'Averages are the main tool in analyzing statistical data' (Ehrenberg, 1986, p3) both for
statisticians and laypeople: 'The trouble is that busy, senior managers can only take aboard a single figure'
(Menneer, 2014).

Despite the mean's widespread use, however, it technically should only be used with interval-level
data (De Vaus, 2002, p361), and with due consideration of the distribution shape of that data.[254] Ehrenberg
(1986, p4) explains that where a mean score is offered by itself as a summary of data, as BBC AIs are, there
is an implication that the distribution of that data is 'hump-backed' and 'symmetrical'. This implication
stems from the fact that if the data does *not* adhere to this shape, then offering the mean as a descriptor is an
inadequate way to summarise the data. As discussed by Bryman and Cramer (1994, p117), one school of
thought dictates that parametric analysis should only be used when the data is specifically normally

---

[254] See p141.

distributed. However, studies that attempt to prove otherwise have found that there are tests that are 'robust' enough to be used in both conditions without 'differing greatly'. In terms of the mean in particular, it *can* be used for non-normal distributions, but with the following important caveats (Ehrenberg, 1986, p3-9):

- A mean alone is not enough to describe such data – information about the distribution must be included.
- For very polarised or very skewed data, the mean may be so atypical it could be a poor representation of a common value.
- If the data has more than one peak, giving one figure as typical is misrepresentative.
- Comparing means from two datasets is only meaningful if the shapes of the distributions are the same.

For Radio 4 comedy appreciation in particular, consider the examples below.[255] The tables show the distribution of appreciation scores for selected Radio 4 comedy programmes all TXing in the same slot (Tuesdays at 18:30) from 2009[256] (from the Pulse Survey: only programmes with verbatim responses).

Figure 31 – Appreciation response distribution frequency table, Radio 4 Tuesdays 18:30, January to April 2009 (BBC Pulse survey) – frequency of responses and percentage mix

Frequency of appreciation scores	Appreciation scores										total number of responses
	1	2	3	4	5	6	7	8	9	10	
Any Tuesday comedy	47	17	12	21	27	31	53	64	39	21	332
Down the Line all Eps	43	17	24	23	19	24	64	62	65	63	404
Down the Line Special	3	2	2	1		1	3	8	5	5	30

% distribution of appreciation scores	Appreciation scores									
	1	2	3	4	5	6	7	8	9	10
Tuesday comedy	14%	5%	4%	6%	8%	9%	16%	19%	12%	6%
Down the Line all Eps	11%	4%	6%	6%	5%	6%	16%	15%	16%	16%
Down the Line Special	10%	7%	7%	3%	0%	3%	10%	27%	17%	17%

The frequency tables above provide a summary of the responses given. While the information in the charts is not particularly complex, it is still difficult really to see patterns or ascertain which programmes are more highly appreciated. The way that this data is actually presented for programmers to use is as AIs, i.e., as a single figure: the mean score appreciation score multiplied by ten:

Figure 32 – Appreciation response summary, Radio 4 Tuesdays 18:30, January to April 2009 (BBC Pulse survey)

Mean scores	Mean Appreciation	AI
Any Tuesday comedy	5.9	59
Down the Line all Eps	6.5	65
Down the Line Special	6.8	68

Taken at face value, the ***Down the Line Special*** has a higher AI than the programmes that had recently been in that slot and all previous episodes of ***Down the Line***. However, what is seldom considered in current BBC Radio analysis is the distribution of the scores.

---

[255] This data was part of the pilot study for this research. Similar examples (not relating to appreciation scores) are used in Allen and Seaman, 2007; and Treiblmaier and Filzmoser, 2009.
[256] 'Any Tuesday comedy' are all the appreciation scores for programmes TXing on Tuesday 18:30 January to March 2009 – these were; *Heresy*, *Broken Arts* and *Cabin Pressure*.
'Down The Line all Eps' are all the scores for any originations (i.e., excluding repeats) for any episode of *Down The Line* that TXd up to April 2009.
 -'Down The Line Special' is just the scores for the one TX of the *Down The Line Credit Crunch Special*, which TXd on 21st April 2009. (NB: small sample size of just 30).

Scores can be plotted on a bar chart (1) or a line chart (2) to compare the distribution shapes visually:

Figure 33 – Appreciation response distribution bar chart, Radio 4 Tuesdays 18:30, January to April 2009 (BBC Pulse survey)

x axis = appreciation scores, y axis = % mix of responses
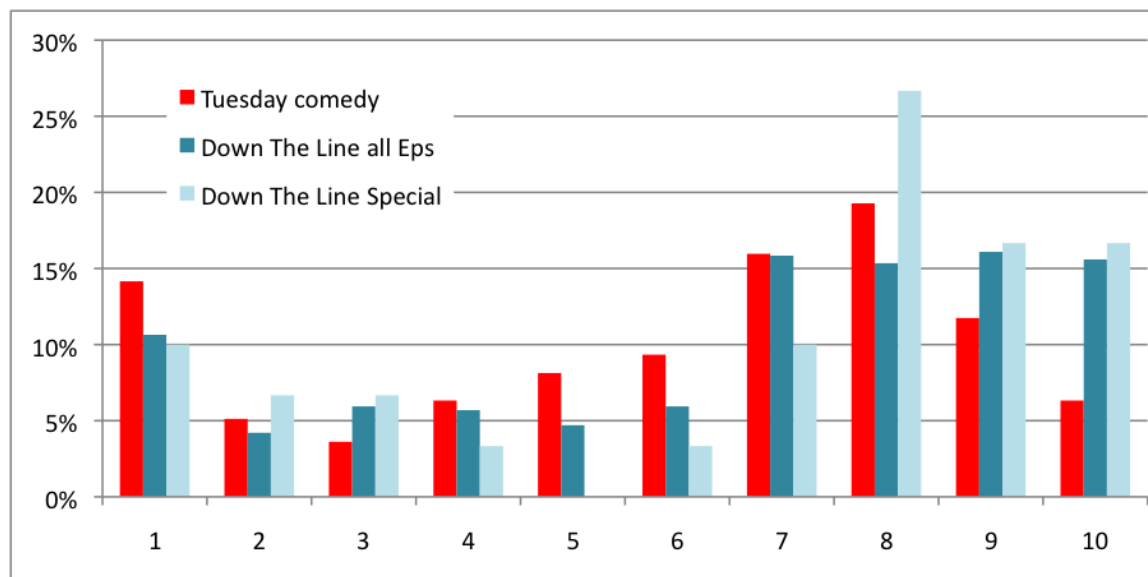
Figure 34 – Appreciation response distribution line chart, Radio 4 Tuesdays 18:30, January to April 2009 (BBC Pulse survey)

x axis = appreciation scores, y axis = % mix of responses
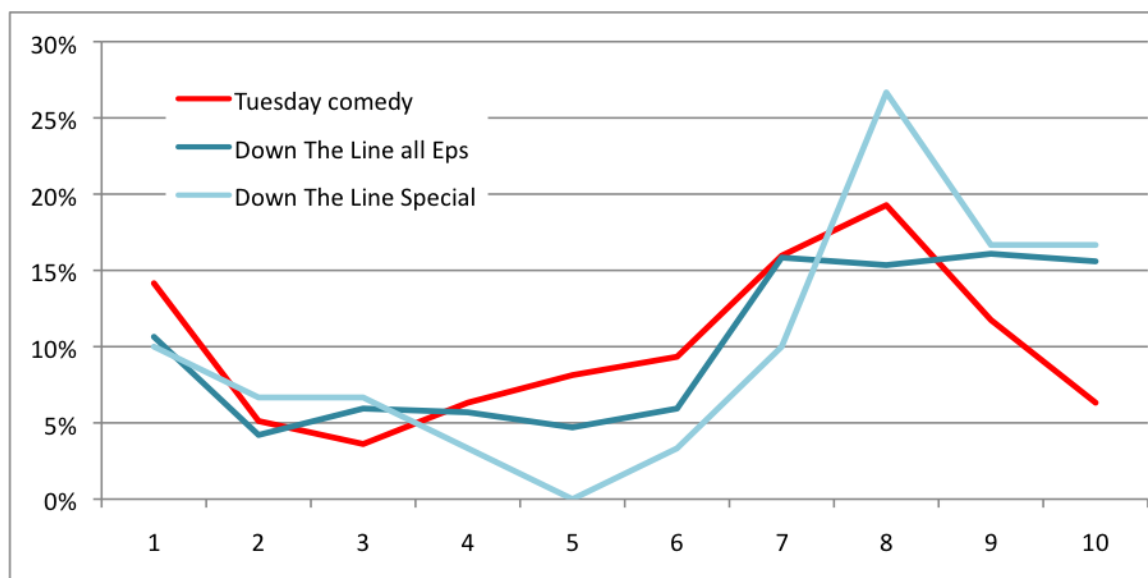
The charts above show that none of the distributions above are unimodal, humpbacked distributions, thus should not be expressed solely as a mean.[257] Ehrenberg claims that regardless of what shape is displayed, if they all took exactly the same shape, there would be some validity in comparing the means. In the example above, the line graphs are not really the same shape. Therefore, with this example, the mean scores, and in turn the AIs would not be valid for comparison.

---

[257] See from p156.

5.5.1 The Mean Score Needs Qualification

The example programmes used earlier were taken purely at random and include a dataset that is of very low sample size. Nevertheless, the figures provide evidence that appreciation scores may not be normally distributed and may in fact be skewed and/or polarised, and may also vary across episodes within a series of programmes. In theory, one could find that the AIs for different programmes were exactly the same and if read at face value – as they are by radio programmers – the assumption would be that the programmes are equally appreciated, with the consequent inference that they are of equal quality. However, different distribution shapes of the data could indicate very different attitudes about the shows. If a programme gets an AI of, for instance, 70, does it matter how the individual scores (of which the mean is a representation) are distributed? A mean of 7 could result from only 7s or an equal mix of 4s and 10s. Is that a significant difference? Appreciation could also be normally distributed. The mean could be the same in these examples, but the sentiment behind it would not be. As an example, all the following distributions have the same AI score:

Figure 35 – Illustrations of AI calculations: different distributions giving the same mean – (fabricated examples, 200 responses)

Frequency of appreciation scores	Appreciation scores										mean	AI
	1	2	3	4	5	6	7	8	9	10		
Programme 1							200				7	70
Programme 2				100						100	7	70
Programme 3				5	15	40	80	40	15	5	7	70

In these examples, the programmers would only see that each programme had an AI of 70. However, with further information about the distribution of the scores, important detail could be seen. Programme 2 might warrant further investigation; why do some people love it and some dislike it? Does it appeal to certain types of listeners? Do the younger people enjoy it but their elders just not get it? Is there content in it that some people might have found offensive? Any number of factors might be at play but there would be no indication of potential issues if the mean were the sole indicator.

Even where data *is* distributed in a unimodal, humpbacked distribution, the correct procedure is that the mean should not be presented without context, i.e., information about how 'steeply' the variable is distributed, as there could be a very narrow or very wide 'scatter' of data points that both result in the same mean. This can be expressed through a variety of measures, typically the 'standard deviation', the 'range' or 'percentiles' of the data (Bryman and Cramer, 1994, p85). The mean needs qualification even when data is normally distributed:

> When quoting an average, like "The average age of the class is 21," the scatter of the individual ages must also be described, at least implicitly. Often a simple statement will do, like "Most of the ages lie between 19 and 23." But sometimes more precision is needed. (Ehrenberg, 1986, p19)

If we assume that Radio 4 comedy appreciation scores are *not* necessarily spread unimodally, we not only need to describe the central tendency and the wideness of the scatter but also the shape of the distribution. Distributions that are not symmetrical and hump-backed cannot be summarised by any single measures for central tendency nor distribution, and may even demand tailored descriptions depending on their complexity:

> For skew data there is usually no simple routine way of summarizing the scatter by a single measure. More specific descriptions may have to be used… Skew, J-shaped, and multimodal distributions tend to be more difficult to summarize. They usually have no typical reading and may therefore require tailor-made descriptions. (Ehrenberg, 1986, p15-28)

Collins (2010a) admitted that, at that time, reflection upon the distribution of BBC Radio AIs was not a consideration for the BBC research team:

> *At the moment, we don't ever go into that. The idea around the AIs is to have one score that's quite understandable. I think that's certainly an interesting thought.*

When presented with the analysis of the data used in this research compared with the typical presentations of AIs, McConway (2013) stated: '*It makes it clear that the mean might be telling you something, but it's clearly not telling you the whole picture*'. Researchers have had a tendency to make assumptions about the spread of the scores. Barwise (2013) was asked whether he considered that the scores might not be distributed unimodally in any of his extensive research on television AIs. His response reveals the supposition made:

> *I don't think we ever looked at that. They roughly are, yes... The AIs are going to be roughly symmetric... Unimodal? Yes. I guess we did assume that.*[258]

5.6 Hypothesis and Research Questions

AI data is important in the commissioning of BBC programmes,[259] particularly in radio where there is no access to overnight ratings figures.[260] Radio AI scores take on a particular significance as they present the most established, objective measure of quality in the absence of ratings.

AI scores are collected through a survey that asks respondents to 'rate' the shows they have seen or heard against a numerical scale numbered 1-10. BBC Audience research presents the resultant data to BBC commissioners and other decision makers either as a mean for the programme or series or an aggregate mean for the whole channel. However, statistical theory suggests that mean scores alone are an inadequate representation of any distribution. This misrepresentation is more pronounced if the data is not measured on at least an interval scale: but there is no evidence that appreciation scores are measured on anything more than an ordinal scale. Distributions that are not unimodal, or are inconsistent in shape across comparable datasets, are represented especially poorly by mean scores alone (the pilot analysis indicated that the data is not unimodally distributed),[261] and should not be subjected to comparison. The academic literature, as well as received wisdom in the industry, suggests that comedy as a programme genre is particularly divisive, even actively encouraging polarised, love it or hate it, audience responses. If this is the case, then presenting appreciation scores as a mean (AIs) may be an especially inaccurate way to represent the data collected for comedy programmes. A mean score, for example, makes it impossible to distinguish between programmes that incite polarised scores from those that garner accordant appreciation. The hypothesis and research questions of this thesis are presented below:

---

[258] When Barwise was shown the actual distribution of Radio 4 appreciation scores (talking about himself in the third person) his response was: '*Wow. You've got a fabulous result there. You've got a really interesting result which is different from what Barwise thought*'. (Barwise, 2013).

[259] See from p88.

[260] See from p81.

[261] See from p158.

H1: Radio 4 comedy programme quality is poorly represented as an unqualified mean of appreciation scores

RQ1: Are appreciation ratings, as collected by the BBC Pulse survey for Radio 4 Comedy, usefully expressed as a mean score?

RQ2: Is the distribution of appreciation scores for Radio 4 comedy particularly polarised, and therefore an extreme case of the problems with using AI scores as a measure of performance?

RQ3: Are there underlying independent variables affecting appreciation ratings that can impact on comparisons of the distribution of appreciation scores across various Radio 4 comedy sub-genres?

Chapter 06 – Method

Sunday May 18th

I wrote to Radio Four.

Dear Sir or Madam
I write as a long-time listener to your amusing programme 'I'm Sorry I Haven't A Clue'.
I must have missed the broadcast when Mr Humphrey Lyttelton explained the rules of Mornington Crescent to the panellists. I wonder if you would furnish me with them.

Yours faithfully
A.A. Mole.

(Townsend, 2004, p371)

As it is important to allow replicability for further research, this chapter first explains the details of the process of obtaining and preparing the data prior to analysis, as issues were present particular to this data set, concerning accessibility and reliability. During the actual analysis, evidence of illegitimate scoring was found and it is explained how these erroneous ratings affect the data, and how they were excluded from the main analysis.

6.1 Rationale and Approach

The BBC's Pulse survey collects attitudinal data from thousands of people every week and in so doing, results in an extremely large data set: the number of responses for just Radio 4 programmes alone during 2012 exceeded 560,000. Using this data allows far larger sample sizes than would be reasonable to undertake in an experimental survey, and should aid generalisability. Furthermore, this data is collected from listeners giving responses to normal audience habits so this should eliminate any illegitimacy that may be expected from experimental measurement of listening – a particular issue that can be found in some humour experimentation.[262]

Existing literature indicates that appreciation is incomparable across different genres[263] so it was decided that the focus of this analysis should be primarily on one genre. Due to the supposed nature of comedy's divisiveness,[264] it was selected for the analysis as theory indicates that if it receives the most polarised appreciation, reflected in the appreciation scores, it could be particularly poorly represented by the mean as a summary of how much it's liked. Comedy is also an important genre in terms of its costliness in comparison to other radio genres, its potential to feed television through piloting and its value to the listeners as entertainment.[265]

While television appreciation has been the subject of existing studies, radio Pulse data was the preference for this study. Firstly because it is a medium that is comparatively overlooked, and secondly because, with lack of overnight audience size figures, appreciation ratings have a greater importance to programme evaluation compared to television.[266]

Data from 2012 was selected as, at the time, it provided the most recent, 'commercially' useful information.

---

[262] See from p110.
[263] See from p114.
[264] See from p48.
[265] See chapter 2 from p15.
[266] See from p85.

Concentrating on one year, genre and platform was also a consideration in terms of usability. With the resources available, widening the analysis would not have been viable due to the limitations of Excel. The result is that this analysis, being focused solely on Radio 4 Comedy for a specific time period, provides findings that are not necessarily generalisable across all time periods, genres or platforms. However, the data spread across all responses to Radio 4 programmes (see p177) indicates that the "issue" of a non-unimodal distribution is not limited to comedy alone.

Existing literature provides understanding into statistical theory and how that might be applied to appreciation ratings.[267] Also, research on AIs specifically gives insight as to audience attitudes that might be applied to this study.[268] However, possibly due to the sensitivity of the data, very little information appeared to be available on how BBC AIs are currently used, and by whom, and attitudes toward them. To address this, private interviews were undertaken.

6.2 Accessing the data

The data required was as follows:

- All responses to Pulse survey questions for all programmes on Radio 4 for the calendar year 2012.
- Each of the responses to include demographic information for each of the respondents.
- The data was to be in a format that was readable by Excel.

BBC Pulse survey data is collected by GfK and passed to TRP Ltd for consolidation before it is then passed to the BBC. The BBC receives the data in a summary form that does not contain response-level analysis. Although negotiations for gaining access to the raw data for this study commenced in 2010, the process proved problematic. GfK were happy to supply the author with the data, but it was supplied in an unreadable format. Approaching TRP for the information was met with a positive response albeit with a major caveat; to be able to provide it would incur a cost of around £3,000 and months of development time. As this was a prohibitive cost, another approach was needed. With the help of a BBC researcher, on re-approaching GfK, the author was able to negotiate with them to get the data in a form that was suitable for analysis. In addition to the Radio 4 data, Radio 4 Extra data was additionally included. This was received in early 2013. A few weeks later TRP, having considered this area of research to be of potential interest after the discussions with the author, had taken it upon themselves to develop the process. While the data had already been obtained for analysis for this research, TRP's development allows future opportunity to access additional data over a longer time period, i.e., it will be possible to select a limited number of programmes but for a number of years. For example, a particularly interesting case study may have been a longitudinal view of *I'm Sorry I Haven't A Clue*, which had a change of Chairman after the death of Humphrey Lyttelton, or perhaps a study tracking *Unbelievable Truth* from its first series to a Radio 4 staple panel show. On investigation, however, the data they could supply currently only went back to August 2009 so the opportunity for a meaningful longitudinal study was not available at the time of this research.

The raw data was supplied in two Excel files. One file contained all the Pulse scores for all the programmes for Radio 4 and 4 Extra throughout 2012 and the other included the respondent identification numbers and the programme TX details. These two files were amalgamated using the lookup function. The resultant file consisted of 665,147 rows of data and 37 columns, meaning that it contained over 24 million data points.

---

[267] See from p136.

[268] See from p85.

6.1.2 Preparation of the Data for Analysis

Once the raw data was obtained, a number of steps were taken in order to enrich and code the figures in preparation for analysis. Further data was extracted from BBC systems, documents and meetings and added to the raw data using the Excel lookup function. The configuration of the columns involved in the lookups must be exactly the same, so care was taken to ensure that they were compatible.[269]

- A number of variables relating to the programme transmissions were taken from Radio 4's scheduling database (Proteus). These were:

  - Episode number, i.e., where the TX sat within a series.
  - The series number of the programme, i.e., for how many series each show had been running.
  - The cost price of the programme (only available for the BBC Radio Comedy department).
  - Whether or not the programme was supplied by a BBC in-house supplier or an indie.
  - Day of the week of TX.
  - The duration of each programme (radio comedies tend to be either 28 or 14 minutes long).
  - Whether or not the programme was an origination, a narrative repeat or an ad hoc repeat.

- With the collaboration of the Commissioning Editor for Radio Comedy, all comedy programmes were segmented by the following criteria:[270]

  - Genre. (Genre was one of the fields supplied with the data but, on inspection, they did not necessarily identify what BBC programmers would consider to be radio comedy. For example, while *Clue* was denoted as 'Comedy', *The News Quiz*, *Just a Minute* and *The Unbelievable Truth* were labelled as 'Game Show', *The Now Show* was 'Entertainment', and *Tim Key's Late Night Poetry Programme* a 'Documentary'. Furthermore, in some cases the genre was inconsistent, for example, *Andrew Lawrence: How Did We End Up Like This?* was sometimes classified as a 'Comedy' but other times was denoted as a 'Game Show'. For each of these examples, they were all allocated the genre 'Comedy'.)
  - Sub-genre. A BBC Radio Comedy Executive Producer was consulted on the segmentation of radio comedy into sub-genres, and sub-genres where then attributed to each programme with the aid of the BBC Radio 4 Commissioning Editor for Comedy.
  - 'Dipability' - is it a programme that the listeners don't need to hear for the whole duration to still enjoy it?
  - Whether or not it had a studio audience.
  - Whether it had one main character/performer or was an ensemble piece.
  - The gender of the main character/performer, or gender skew if ensemble.
  - The general age (for example, 'older, 'middle', 'younger', 'assorted') of the characters/performers.
  - Whether or not the programme was talent-led.
  - Whether or not the show was topical.

- Figures relating to online listening were added for a potential further behavioural measure in the absence of programme-specific RAJAR figures.

- BBC Radio media planning and marketing provided information as to whether each of the TXs was promoted via either on-air radio or television trails. The relative weightings of each campaign were also recorded.

- The data was then sorted in the following order: respondent, date, time. This would ensure that for each respondent, the date replicated the order in which they would have rated each of the programmes.

---

[269] This aspect alone took around fifteen hours.

[270] This aspect alone took around twelve hours.

Problems Encountered with Preparing the Data

- The verbatim responses result from the two open-ended questions: one about what they like and the other about what they don't like. Responses were inconsistent. Sometimes it looks as if they were the wrong way round in the raw data. It's unclear whether the data is incorrect or whether the respondents did not read or understand the questions correctly. However, for the purposes of this analysis this aspect was not an issue as verbatim responses were considered regardless of which question they were attributed to.

- Ideally, having access to data across a longer time period would have been useful and would have allowed a longitudinal view.

- There is no unique identifier for TXs across the various data sources; variables had to be concatenated in order to create a combination field which could be used to look-up from one data source to another in order to combine in one excel sheet.

- The time needed to code, inspect and sanitise the data, let alone undertake any analysis, was very lengthy due to the volume of information.

- The sheer volume of data proved unwieldy. The resultant Excel file consisted of nearly 28 million cells of information and was 200 Mb. A file this large was found to crash constantly, regardless of the speed of the computer or the operating system on which it was used.

- Of all the rows of data, 59 did not include an appreciation score.

6.3 Sanitising the Data

The first task after coding data was to look at the distribution shape of the appreciation scores.

Figure 36 ALL RESPONDENTS, ALL RESPONSES, Weighted, All Radio 4 and 4 Extra – Total number of appreciation scores[271]
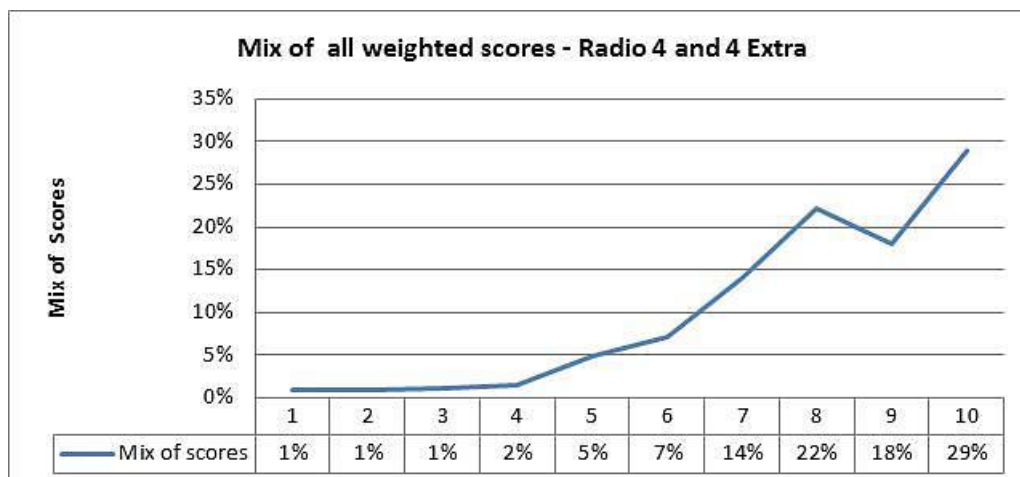
It has already been discussed that AIs are a mean score and means are arguably only statistically valid when applied to unimodally-distributed data.[272] Visual inspection of the figure above clearly shows that the data is not unimodal, let alone normally distributed.

While there was an expectation that the data could be polarised, at total level the distribution was entirely unexpected as no literature found had illustrated or even implied such a shape. Due to this unexpected result, a brief inspection of the data showed that some responses may not have been legitimate

---

[271] Weighted and unweighted distribution was very similar.

[272] See from p135.

and this observation led to the need to review the data at a line level to identify all such responses.[273] Due to the sheer number of responses, the sanitisation of the data took over fifty hours. The following action was taken:

- Where any respondent had entered fewer than 10 responses over the whole year, this data was excluded as it was felt that they might be just sampling the survey.

- Any respondent who had given the same appreciation score to *every* one of the programmes that they had claimed they heard was excluded. This is known as '*flat lining*' (North, 2010).[274] An example was respondent 10704779, found to have given every one of their 4324 responses the same appreciation score of 10.

- Van Meurs' (2009b, slide 6) defines 'straight lining' for appreciation responses as '5 or more identical answers'. There is no further qualification in the document. The assumption is that this means 5 or more in a row. There is also no mention of whether this is over a specific time period. For example if a respondent were to give the same response for 5 programmes in a row on one day compared to, say, having listened to just 5 programmes over 5 days and given each the same response. For the purposes of sanitising the data, having 5 responses in a row, regardless of whether this was over more than one day, was considered as straight-lining. Thus it was highlighted in the data where this had occurred. However, on inspection of the data, including validation against the verbatim responses where necessary, it appeared that many of the responses in such cases appeared valid. Many respondents might have a very limited listening pattern, such as just the *Today* programme, and would give it a consistent score. There was no reason to believe that these responses were not sincere. Each of the circa three and a half thousand respondent response patterns[275] were considered individually. North (2010) gave an alternative definition of 'straight lining' as giving all the same scores within a day's listening. Where this was seen consistently with a respondent, it was deemed to be invalid. For example, respondent 10650080 was found to have done this, typically giving all programmes on any day a score of either 3 or 4; generally there was never more than one score used across programmes on any given day.

The process of sanitisation involved elements of subjectivity, and verbatim responses in particular indicated where appreciation scores were not just rote responses. Respondents such as 591573 were excluded as 98% of the responses were '10' and some programmes that were given a '10' had a verbatim response of '*adequate*'. This indicates that '10' was merely a default response and had no consideration behind it . Alternately, respondents such as 592548, despite giving 99% of responses as '10' were included, as verbatim responses such as 'Fantastic absolutely fantastic' and 'I think that this is quite possibly the best programme on either television or radio at the moment; it is 30 minutes of total fun!' indicated that this person loved what they listened to and listened to what they loved, thereby legitimising their high scores. Likewise, those giving consistent low scores but positive reviews were excluded, such as respondent 10797979:

Date	Programme title	Verbatim Response	Appreciation Score
07-May-12	*The Shipping Forecast*	'It was sexy'	2
10-Apr-12	*Rumpole of the Bailey*	'It was brilliant'	1
10-Apr-12	*The Arts and How They Was Done*	'It was excellent'	1

---

[273] It is worth noting that the BBC will have used and published summary data taken from the whole without any sanitisation, including illegitimate ratings observed in the process of this research.

[274] See from p119.

[275] Any respondent giving fewer than 10 responses was not considered, and was just excluded automatically.

6.2.1 How Much of the Data was Illegitimate?

The 665,147 BBC Pulse responses for Radio 4 and 4 Extra came from 7690 different respondents.[276] A summary of the responses is shown:

Figure 37 ALL RESPONDENTS, ALL RESPONSES, Unweighted, 4 AND 4X – Legitimacy of Responses

Legitimacy of response	Number of respondents	% mix of respondents	Number of responses	% mix of responses	Mean number of responses per respondent
– Illegitimate – < 10 responses	3955	51%	11635	2%	3
– Illegitimate – Flat line	201	3%	39141	6%	195
– Illegitimate – Straight line	18	0%	3813	1%	212
– Illegitimate – Other	18	0%	9287	1%	516
Illegitimate – total	**4192**	**55%**	**63876**	**10%**	**15**
Legitimate – total	**3498**	**45%**	**601271**	**90%**	**172**
Grand total	**7690**	**100%**	**665147**	**100%**	**86**

- Overall, 55% of the respondents were considered to have given illegitimate responses, although the majority of these were from those with low response rates: thus, only 10% of responses were ultimately deemed worthy of exclusion.

- The sanitisation process identified that 51% of the respondents gave fewer than 10 responses. This seems like a high number to exclude considering that their responses are not necessarily illegitimate. However, as they are low response rates, the impact of excluding them is low: only 2% of the total responses.

- Flat lining accounts for 6% of all responses and is therefore the biggest issue.

- Straight lining is a relatively low figure as only 18 respondents were identified as doing this.

The biggest issue here appears to be from those who flat line. Although they only account for 3% of the respondents, they gave 6% of the responses. This appears to be an issue that BBC researchers are aware of. When questioning the legitimacy of the data in April 2013, the response from a member of the research team included the following (anonymous private correspondence, 2013):

> If you look at the data, most of that [flat lining] occurs in the first 6 months of 2012. GfK have taken action to remove the worst offenders and it's less prevalent now.
> …A small proportion of outliers can't actually have that much effect when there are several thousand respondents per quarter. It has had more of an effect on R3 and networks with smaller audiences. Not that we've really looked at it at a programme level yet – but it's clear that it's going to have a bigger impact on programmes with smaller audiences and it could down-weight them a bit.
> I think it's fair to say we're still working through the impact of this…

So, is it true that the issue was solved midway through 2012? And is it also true that the effect on figures is small?

---

[276] A 'respondent' is defined by a unique 'respondent ID' in the raw data.

6.2.2 Has the Issue with Flat Lining Been Resolved?

There were 201 respondents identified as flat lining. Were their responses limited to the first half of 2012 as is the view of the BBC research department?

Figure 38 ALL RESPONDENTS, ALL RESPONSES. 4 AND 4X – Legitimacy of Responses – split by half year

Legitimacy of response	Number of responses Jan-June	Number of responses July-Dec	Total Number of responses	% Mix of responses Jan-June	% Mix of responses July-Dec
Legitimate – total	**291807**	**309464**	**601271**	**89.3%**	**91.4%**
– Illegitimate – < 10 responses	5473	6162	11635	1.7%	1.8%
– Illegitimate – Flat line	22763	16378	39141	7.0%	4.8%
– Illegitimate – Straight line	1870	1943	3813	0.6%	0.6%
– Illegitimate – Other	4799	4488	9287	1.5%	1.3%
Illegitimate – total	**34905**	**28971**	**63876**	**10.7%**	**8.6%**
Grand total	**326712**	**338435**	**665147**	**100%**	**100%**

If GfK have indeed taken action to reduce the number of respondents who are flat lining, there appears to be evidence that it has had an effect, but it has not been wholly successful. Flat line responses accounted for 7.0% of all responses in the first half of the year. There is a visible drop in the second half, but it still remains noteworthy at 4.8%. Focusing on some individual respondents who were guilty of flat lining can further illuminate the issue:

Figure 39 TOP 10 FLAT LINING RESPONDENTS, 4 AND 4X – split by half year

Respondent ID	Number of responses Jan-June	July-Dec	Grand Total
10704779	4272	52	4324
1097121	1359	1290	2649
1570867	638	1142	1780
10669762	1297	295	1592
906415	710	721	1431
1107461	730	689	1419
1164438	623	615	1238
583898	662	391	1053
1187734	454	469	923
1075025	393	488	881
Total	11138	6152	17290

Looking at the top 10 flat liners, we can see that some individuals have contributed thousands of responses throughout 2012. The biggest offender, respondent 10704779, only gave responses up to October 31st, which might be the result of the attempt to reduce the 'worst offenders'. However, from the earlier chart we can see that the reduction in flat line responses is 6,385. We can attribute 4,220 of these to this one respondent. A reduction of 1,002 can also be attributed to respondent 10669762, who was, nevertheless, still giving survey responses into December 2012. In this case, we cannot attribute their reduction in responses to GfK's action. The majority of these respondents who were giving high volumes of flat responses did not significantly decrease their volume throughout 2012. This indicates that the issue of illegitimate responses was still ongoing at the end of 2012.

> 10% of appreciation scores for Radio 4 throughout 2012 can be classed as illegitimate and while there is evidence that the proportion may be decreasing, the issue does not appear to be completely resolved.

6.2.3 Do Illegitimate Responses Skew AIs?

While this research has allowed us to identify illegitimate responses, the inevitable time-consuming nature might be prohibitive for BBC researchers. Bearing this in mind, does the inclusion of these responses make a material difference to figures as they are used by the BBC?

i Illegitimate Response Effect at Station Level

The BBC uses aggregate AIs as a published measure of quality.[277] The topline figures published throughout 2012 included the responses that we have identified as illegitimate. Firstly, it was worth checking that the raw data being used for this analysis bore relation to the figures published. (NB: 59 of the responses had blank cells for the Appreciation score. These were excluded when analysing AIs):

Figure 40 ALL RESPONDENTS, ALL RESPONSES, Radio 4 – Aggregate AI: Calculated vs Published by Quarter

Quarter of year	Published AI	Calculated AI (unweighted)	Calculated AI (weighted)	Variance of Calculated AI (weighted) vs Published
Q1 2012	81.1	80.4	81.0	0.1
Q2 2012	80.7	80.5	80.7	-
Q3 2012	80.7	80.6	80.7	-
Q4 2012	80.7	80.6	80.6	0.1

Figure 40 shows that once the appreciation ratings are weighted[278] the calculated aggregate AI is almost the same as that published. There is no obvious explanation for this the slight variance but it could be attributable to the 59 responses with no scores or, as the BBC does not describe the exact dates it uses to define a 'quarter', a slight difference in timing could contribute to the variance. This minor variation is deemed acceptable.

Figure 41 ALL RESPONDENTS, ALL RESPONSES, Radio 4 – Legitimate vs Illegitimate Aggregate AI: by Quarter

Quarter of year	Calculated AI (weighted)			Variance of All ratings vs just legitimate
	All	Legitimate	Illegitimate	
Q1 2012	81.0	79.8	90.2	1.3
Q2 2012	80.7	79.5	90.6	1.2
Q3 2012	80.7	79.9	88.4	0.8
Q4 2012	80.6	79.7	89.4	0.9

Figure 41 shows the effect of the illegitimate responses upon the total for each quarter. The continued inclusion of the illegitimate responses clearly shows that the aggregate scores *are* affected, as illegitimate scores are considerably higher than legitimate scores. Including the illegitimate scores plainly results in the figures being overstated despite the view of BBC research that: 'A small proportion of outliers can't actually have that much effect when there are several thousand respondents per quarter'.[279] The variance is slightly lower in the second half of the year versus the first half, supporting the claim that there is a reduction in the 'worst offenders', but the overstatement is still observable.

---

[277] See p88.

[278] All AI scores will be weighted from this point on unless otherwise stated. Respondent weightings were included in the raw data files.

[279] See from p167.

2012 – all responses	All	Legitimate	Illegitimate	Types of Illegitimacy				Variance of All ratings vs just legitimate
				<10 responses	Flat line	Straight line	Other	
Calculated AI (weighted)	80.8	79.7	89.7	76.4	96.4	62.4	89.7	1.1
Number of responses (with AIs)	561669	509098	52571	9691	32263	2537	8080	

The figure above shows that across the whole of 2012 for Radio 4, scores identified as illegitimate raised the aggregate AI by 1.1. This boost in the score is mainly down to respondents who are flat lining, as the aggregate AI for them is 96.4 – far higher than the 79.7 of the legitimate scores. This is because the majority of flat lined scores were 10s (78% of the 32,263). Straight line scores give a far lower aggregate AI at 62.4. However, there are far fewer incidences of this, so the effect on the total is less pronounced.

Is the Inflated AI, Due to Illegitimate Scores, Noteworthy?

While the claim by BBC researchers that illegitimate scores 'can't actually have that much effect',[280] we can see that across all of 2012 it resulted in an overstatement of 1.1 point (80.8 versus 79.7). Is this noteworthy? Apparently, yes. For example a headline in Broadcast magazine (Chapman, 2014) claimed 'Appreciation for all BBC TV channels falls' and specified:

> BBC1's AI score dropped to 81.5 out of 100 [*sic*] in 2013/14, down from 82.6 the previous year, while appreciation for BBC2 deteriorated from 84.2 to 83.9 over the same period. Audiences scored BBC3 83 out of 100 - down from 2012/13's rating of 84. Finally, BBC4 saw its AI fall 0.5% to 85.1 year-on-year.

Each of these drops in AI are no greater than we have seen as the increment as a result of illegitimate scores, thus, this level of variance in aggregate AIs *is* noteworthy in that it has been noted by a media publication.

ii Illegitimate Response Effect at Programme Level

Examination of aggregate scores is done at corporate level but bears little on the decisions made for individual programmes. If it has been identified that illegitimate scores account for an increase of 1.1, can it be assumed that this will be applicable at a lower level, say for an individual radio comedy series? Might it be possible just to subtract 1.1 from a programme's AI to eradicate the effect of illegitimate scores?

---

[280] See quote (from Anonymous private correspondence, 2013) from p167.

Type of programme	Title of programme	Total number of responses			Calculated AI (weighted)			Variance of All ratings vs just legitimate
		All	Legitimate responses	Illegitimate responses	Total	Legitimate responses	Illegitimate responses	
Long-running panel show	*Clue*	3707	3431	276	88.4	88.0	92.8	0.4
	JAM	4742	4409	333	85.8	85.0	95.1	0.8
Established sitcom	*Cabin Pressure*	527	497	30	76.4	75.5	90.0	0.9
	Count Arthur	929	873	56	63.0	61.3	83.8	1.7
First series sketch show	*Kevin Eldon Will See You Now*	105	96	9	59.1	59.5	55.9	-0.4
	Meanwhile, It's Will and Greg	76	71	5	54.9	49.4	100.0	5.5

The figure above shows a number of different Radio 4 comedies selected to illustrate a range of programme types, some with many responses and some with far fewer. Comparing *Clue* and *JAM*, both long-running panel shows, both originating at 18:30, and both with thousands of responses, there is an observable difference in the effect that illegitimate scores have upon the AI. *Clue* garners an increase of 0.4 while JAM sees an extra 0.8, increasing its score from 85.0 to 85.8. Looking at all the variance figures, each programme has been affected to a different extent. The extremes in these examples are for the new sketch shows. *Meanwhile, It's Will and Greg* has been boosted by 5.5 points because all 5 of its illegitimate responses were 10s. Thus the Commissioning Editor might look at the average AI thinking that 55 is an acceptable score, whereas without the illegitimate 10s it was only really 49. *Kevin Eldon Will See You Now* has actually registered a drop of -0.4 because of the inclusion of the illegitimate responses. While volatility is more likely for these new programmes as the number of responses is relatively low, they are both above the cut-off point of *50* that is seen by the BBC to be an acceptable point for (careful) consideration for programme evaluation.

Van Meurs (2009b, slide 31) found that illegitimate ratings increased television AI ratings. For the programmes he looked at (none of which were comedy), he observed increases from 0.18 to 0.77.[281] These figures are broadly comparable with our 'long running panel show' figures and therefore illustrate that there may be a common issue for radio and television appreciation ratings.

While systematic, ongoing overstatement of appreciation might be something that could be taken into account, for example, by subtracting 1.1 from an aggregate AI score, we can see that no such easy fix is viable at programme level. Different programmes are affected to a different extent so illegitimate responses appear to be a problem when it comes to using appreciation scores to evaluate programme performance.

6.2.4 Are Respondents Who Give Illegitimate Responses Different From Those Who Don't?

Van Meurs' report *Checks on the Quality of Online Responses* (2009b, slides 6-34) looked at a number of surveys to check for incidences of flat lining.[282] For BBC television Pulse data, 7.9% of the responses were found to have 5 or more identical appreciation scores in a row, or 11 in a row for any of the other Pulse questions (slide 6). This is not wholly dissimilar to the proportion of illegitimate incidences found by this

---

[281] NB: in Van Meurs's report appreciation is expressed as means of scores rather than AIs, therefore his results have been multiplied by a factor of 10 to make them comparable.
[282] In this report he calls it straight lining as he makes no differentiation between flat and straight lining at this point.

research for the total of flat lining and straight lining at 7%.[283] He discusses how certain demographic groups are more likely give illegitimate responses, including (slide 13):

- The oldest and youngest respondents.
- Low income groups.
- In studies from New Zealand and Barcelona, women were found to be more likely to be at risk of straight lining.

Might these groups be seen to be giving illegitimate responses for Radio 4?

i Illegitimacy by Age

Figure 44 ALL RESPONDENTS, All Radio 4 Responses only – Respondent demographics and legitimacy – Age groups

Age Group	RAJAR all 2012 – weekly hours	mix	All responses	mix	weighted mix	Illegitimate responses	mix	weighted mix
15-23	n/a		5452	1%	1%	386	1%	1%
15-24	2,124	2%	14,015	2%	4%	5,273	10%	16%
25-34	5,426	4%	23,583	4%	6%	3,035	6%	8%
35-44	10,597	8%	38,549	7%	8%	5,966	11%	13%
45-54	20,878	16%	88,937	16%	13%	8,035	15%	10%
55-64	30,686	24%	188,100	33%	23%	12,304	23%	15%
65+	58,453	46%	203,033	36%	45%	17,572	33%	37%
TOTAL	128,163	100%	561,669	100%	100%	52,571	100%	100%

The youngest group that RAJAR includes in its adult figures – 15-24s – accounts for 2% of all Radio 4 listening. However, while 2% of appreciation responses come from this group, their responses have been up-weighted, i.e., contributes 4% of the appreciation scores. Unfortunately, this small group accounts for 10% of all the illegitimate responses, and due to up-weighting the result is that 16% of all the illegitimate data comes from this age group. 25-34s and 34-44s also account for more illegitimate responses than might be expected based on the number of responses that they give and the proportion of listening that they account for. Van Meurs (2009b, slide 13) found that the oldest and youngest respondents are likely to give illegitimate responses, but our data shows that this behaviour is prevalent in just younger respondents.

ii Illegitimacy by Social Grade

Figure 45 ALL RESPONDENTS, All Radio 4 Responses only – Respondent demographics and legitimacy – Social group

Social grade	RAJAR all 2012 – weekly hours	mix	All responses	mix	weighted mix	Illegitimate responses	mix	weighted mix
AB	52,951	41%	276,962	49%	40%	20,429	39%	42%
C1	44,651	35%	154,267	27%	29%	14,936	28%	29%
C2	17,522	14%	44,323	8%	12%	6,092	12%	12%
DE	13,038	10%	86,117	15%	18%	11,114	21%	17%
TOTAL	128,163	100%	561,669	100%	100%	52,571	100%	100%

Van Meurs (ibid) found that those in lower-income groups were more likely to give illegitimate responses. Pulse and RAJAR data do not include data on income but there is information on social grade, which is likely to be analogous. So, looking at the straight, unweighted mixes, DEs account for 17% of all illegitimate

283 See figure on p167.

responses but only 10% of the listening; thus they are, similarly to Van Meurs findings, most likely to give illegitimate responses.

iii Illegitimacy by Gender

Figure 46 ALL RESPONDENTS, All Radio 4 Responses only – Respondent demographics and legitimacy – Gender

Gender	RAJAR all 2012 – weekly hours	mix	All responses	mix	weighted mix	Illegitimate responses	mix	weighted mix
Men	56,524	44%	301,336	54%	51%	28,625	54%	52%
Women	71,638	56%	260,333	46%	49%	23,946	46%	48%
TOTAL	128,163	100%	561,669	100%	100%	52,571	100%	100%

Van Meurs (ibid) found that women were more likely to give illegitimate scores. As shown above, the weighted mix of illegitimate scores is 48%, which is lower than both the weighted Pulse mix and the proportion of listening. Thus there is no evidence in our data that women are more likely to give illegitimate scores than men; in fact, our figures show the opposite.

> There is evidence that some demographic groups may be more likely to give illegitimate responses than others: younger, C2 DE men are most likely to do so. This could be related to the typical Radio 4 audience being skewed towards older ABC1 women (RAJAR 2013), i.e., those who are least likely to be Radio 4 'fans' may be most likely to be careful about the survey responses they give for the station.

6.2.5 Are There Other Factors Affecting Likelihood of Illegitimate Responses?

In addition to looking at demographic differences, Van Meurs (2009b, slide 13) investigated whether the circumstances of participation in the survey affected occurrences of illegitimate responses. He found that aspects such as longer survey length and more days of participation in the survey both increased the probability of respondents giving illegitimate responses.[284] He also implied that there is a relationship between attention and effort made by the respondent in listening to the programme and the likelihood of giving illegitimate responses: 'Straight lining probability is related to participation' (ibid, slide 8). The exact nature of the relationship is not made clear. However it is possible to investigate whether anything relating to participation and effort can be seen in the data considered here. One of the Pulse questions is:

> - **And how much of each programme did you listen to yesterday? Please mark on the scale below where 1 means listened to hardly any and 10 means you listened to all of it.**
>
>   **1 – Listened to hardly any, to…**
>   **10 – Listened to all of it**

The scores given by the respondents for this question relate to an aspect of participation with the programme.

---

[284] The data available for this research does not include any information about the duration of participation.

Amount listened to	All responses	mix	weighted mix	Illegitimate responses	mix	weighted mix
1	8695	2%	1%	888	2%	1%
2	19142	3%	3%	1743	3%	3%
3	31515	6%	5%	2824	5%	5%
4	25867	5%	5%	1836	3%	4%
5	38002	7%	7%	2338	4%	5%
6	32992	6%	6%	2057	4%	4%
7	39527	7%	7%	2062	4%	4%
8	45181	8%	8%	2161	4%	4%
9	38329	7%	7%	2053	4%	3%
10	282419	50%	50%	34609	66%	67%
TOTAL	561669	100%	100%	52571	100%	100%

One might think that a respondent listening to only a portion of a programme might be less likely to be careful with their appreciation responses, but the data in the figure above does not show this. Where respondents have scored 10 for a programme – i.e., they have listened to the whole programme – they are *more* likely to have given an illegitimate score. However, these are claimed figures; just as a respondent might give a flat line response of 10 for appreciation, they could merely be giving a flat line response for 'amount listened to'.

A further measure of participation can be seen through the Pulse question relating to effort to listen:

- **And how much effort did you make to listen to each of these programmes?**

   I made an extra special effort to listen to the programme.
   I usually listen to this show as part of my usual routine.
   I did not make any particular effort to listen to the show.

Effort made to listen	All responses	mix	weighted mix	Illegitimate responses	mix	weighted mix
01 Extra effort	81381	14%	15%	13058	25%	25%
02 Usual listening	361916	64%	65%	31265	59%	60%
03 No effort	118372	21%	20%	8248	16%	15%
TOTAL	561669	100%	100%	52571	100%	100%

Just as with the claimed amount listened to, those who claim to participate more – in terms of effort – appear to be more likely to have given illegitimate responses. It seems likely that those giving illegitimate ratings responses would be giving illegitimate scores for all questions, so this relationship may not be valid to measure with this data.

There is no evidence that illegitimate responses are related to claimed lack of effort or attention paid to a programme.

6.4 Weighting Issue

Earlier figures highlight a further issue: that of weighting. The Pulse survey is supposedly 'representative of the UK population' (BBC Gateway Audience Portal, 2011), and as such should represent the behaviour of the population. However, when we compare the mix of weighted Pulse responses, for Radio 4 at least, the proportions do not match the proportion of RAJAR's listening figures exactly. For example, looking at our data split by age groups,[285] while 45+ respondents account for 86% of all listening hours, their weighting means that their responses only account for 81% of Radio 4's Pulse responses. This means that their Pulse responses are being underrepresented compared to the amount of listening that they do. In terms of social groups, we can see that DEs are only responsible for 10% of Radio 4's listening, but responses have been up-weighted to account for 18% of all the Pulse responses.[286] Although women listen to 56% of the hours heard on Radio 4[287] they account for just 46% of the appreciation scores, which while they get up-weighted, are done so only to 49%. Thus, women's ratings are underrepresented versus their proportion of listening.

This issue with weighting occurs because the weighting of respondents is done by TRP across the *whole* Pulse survey to reflect the demographic make-up of the UK; it is not done to imitate any individual BBC channel/station (Tingay, 2013). Unfortunately, the weighting errors are not simple to account for. While we could re-weight all our Pulse responses to reflect audience size as measured, we have to remember that RAJAR figures are taken from a sample and have their own limitations which restrict their absolute legitimacy.[288] With this in mind, the legitimate weighted Pulse scores will be used as is, because we cannot guarantee that re-weighting them will be 100% accurate.

> The Pulse respondents for Radio 4 do not exactly represent its listeners as measured by RAJAR. This means that the attitudes of some groups are underrepresented in the Pulse data compared to the proportion of listening that they do.

---

[285] See figures 44 to 46 from p172.

[286] See p172.

[287] See p173.

[288] See from p81.

Chapter 07 – Research Question 1 – Are Mean AI's a Valid Representation of Audience Responses to Radio Comedy?

When the BBC first began broadcasting in the 1920s there were no radio 'comedy' shows as such; programmes weren't intended to make the listener laugh. How different to the late 1990s when there was a whole BBC department devoted to making comedy shows specifically designed not to make the listeners laugh.

(Brooke-Taylor et al, 1999)

This chapter is the first of three that present the results of the analysis of BBC Pulse appreciation data described in chapter 6. Each of the three uses the data to attempt to answer one of the three main research questions with this chapter presenting an attempt to answer RQ1.

Each of the following three chapters uses only legitimate responses, unless highlighted otherwise.

RQ1: Are appreciation ratings, as collected by the BBC Pulse survey for Radio 4 Comedy, usefully expressed as a mean score?

In this chapter the analysis of the data is used to answer RQ1 in three steps. Firstly, data is analysed to show the distribution of the ratings at the topline level. The analysis establishes that this distribution indicates that the data is not strictly suitable for parametric analysis and thus that mean AI scores are not a statistically valid representation of the dataset. Next, the data is analysed at the respondent level – this demonstrates that although different people distribute their responses in different patterns, the data is not random. Instead, these shapes can be broadly described using a limited range of categories. Finally, the analysis suggests that while mean AI's may not be a useful representation of AI data, these means correlate very closely with the Net Promoter Score (NPS), a figure that is statistically viable for non-parametric data. However, while NPS is robust, it still suffers from the same issues as the mean in that polarised scores[289] can be obscured so is not a useful sole representation of the data.

7.1 How are Radio 4 Appreciation Scores Distributed at Topline?

Initial inspection showed that the distribution of the data was bimodal[290] but this included illegitimate scores. These scores have been extracted from the data set leaving just the sanitised, legitimate scores. The first step in analysing this legitimate data is to follow Bryman and Cramer's recommendation of looking at the shape of a curve to identify whether it is distributed in a unimodal fashion: 'there is much to be said for a visual inspection of the data' (1996, p96-67):

---

[289] See from p222 for demonstration of Radio 4 comedy as a polarised genre.
[290] See p165.

Figure 49 ALL RESPONDENTS, All Radio 4 Responses only – Distribution of Legitimate vs Illegitimate appreciation scores

Mix of all weighted scores - Radio 4 and 4 Extra - Legitimacy

	1	2	3	4	5	6	7	8	9	10
Legitimate scores	1%	1%	1%	2%	5%	8%	15%	24%	19%	24%
Illegitimate scores	1%	1%	2%	2%	2%	4%	5%	8%	9%	66%
All scores	1%	1%	1%	2%	5%	7%	14%	22%	18%	29%
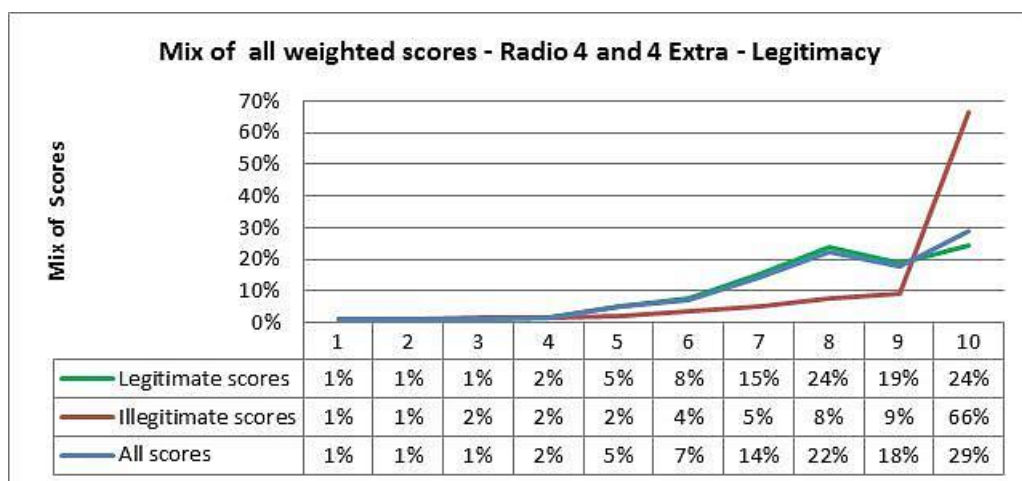
In figure 49 the green line, represents the legitimate scores, the red line the illegitimate responses. The red line is highly skewed toward 10s, demonstrating that the illegitimate scores artificially boost the proportion of 10s found across all the responses. The figure on demonstrates that illegitimate scores boost the aggregate, mean AI for Radio 4 from 79.7 to 89.7. However, the green line shows that even when the illegitimate scores are stripped out, the resultant shape is *still* bimodal. We can discount sample size as a cause of this, since the legitimate sample comprises 561,669 responses, with even the lowest scores (1s and 2s) each based on samples of over 5,000 responses.

Figure 50 shows just the distribution of the legitimate scores for all of Radio 4 and 4 Extra, and then for Radio 4 comedies alone in figure 51. A visual comparison demonstrates that comedy responses are distributed in a similar manner to all genres of programmes, although the shape for comedy has a slight peak at the lower end of the scale and the most popular score (mode) is 10 rather than 8.

Figure 50 ALL RESPONDENTS, Weighted, All Radio 4 and 4 Extra Responses – Distribution of Legitimate appreciation scores

Mix of all weighted scores - Radio 4 and 4 Extra - Legitimacy

	1	2	3	4	5	6	7	8	9	10
Legitimate scores	1%	1%	1%	2%	5%	8%	15%	24%	19%	24%
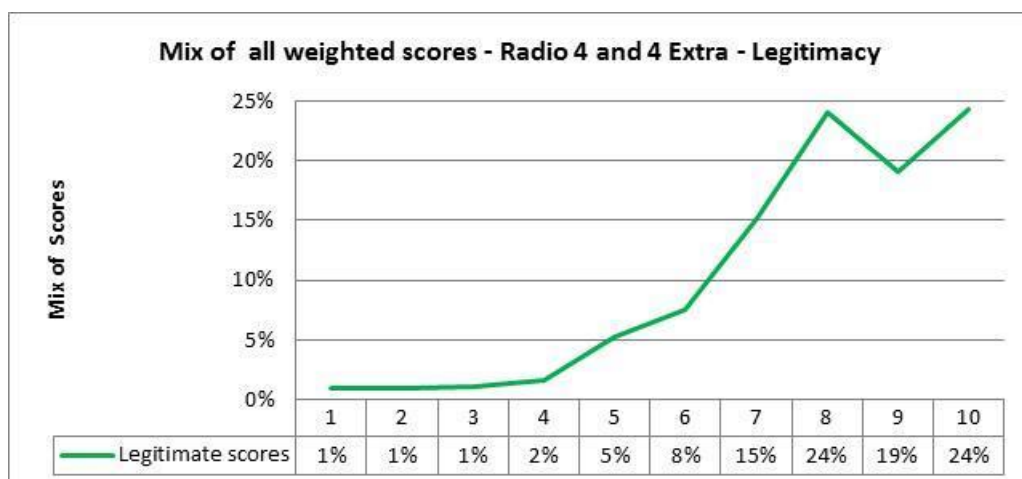
Figure 51 ALL RESPONDENTS, Weighted, Radio 4 Comedy Responses only – Distribution of Legitimate appreciation scores
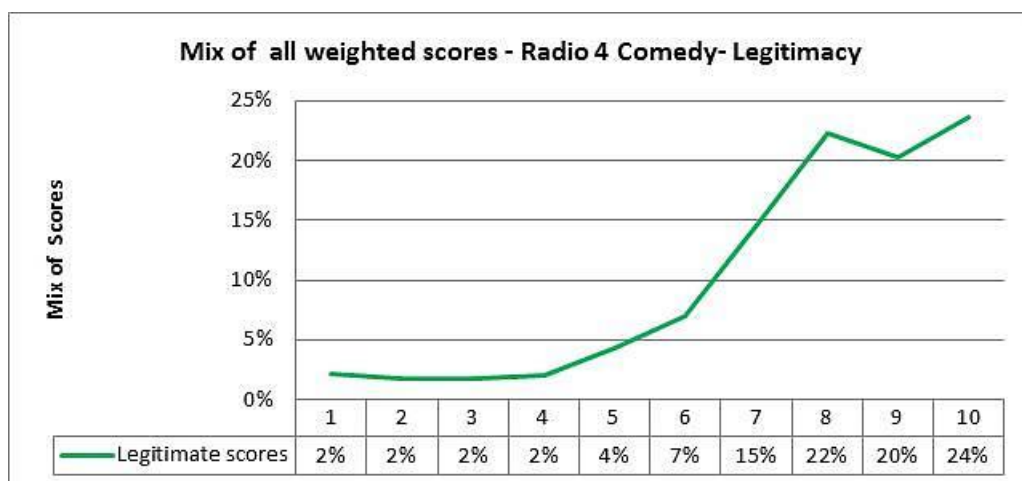
Even when illegitimate responses are eliminated, Radio 4 comedy appreciation scores do not have a unimodal, humpbacked distribution. The aggregate responses have a broadly bimodal distribution with peaks at 8 and 10. As the distribution is bimodal, it cannot be adequately represented solely by a mean score, i.e., the AI. At the top-line level, radio comedy does not appear to be a strongly polarising genre, following a pattern broadly similar to the aggregate of all Radio 4 programmes (albeit with a small peak for scores of 1).

7.2 Do Respondents Have Patterns of Response?

But why is the distribution for Radio 4 and Radio 4 comedy this particular shape?[291] The discussion of expected distribution shapes from attitudinal data[292] hypothesised that a unimodal hump-backed curve, a J-shape or even a polarised distribution might be observed. This distribution shape observed – with a double peak at the top end of the scale – did not feature in the studies discussed.

A specific feature of the AI data is that where other studies tend to look at ratings for one product but from a number of different respondents, this data comes from *many* responses from *many* people across *many* programmes. Could it be that response patterns might be different when people rate a number of products (or in this case programmes) compared to when they rate a single product? Respondents have been found to 'attempt to provide answers consistent with the ones they have already given in the survey' (Bertrand and Mullainathan, 2001, p67) and also, if lack of effort in completing the survey is present, rote responses from respondents might lead to a standard pattern which might mean that they will need to think less about each rating.[293] The Radio 4 data might reflect this process of 'panel conditioning',[294] for example, the figure on p180 shows that respondent 2586179 marks most Radio 4 shows as 8 as a default, and only uses scores from 6 to 9. This relates to the concept of RSs (response styles) discussed earlier (see from p122) whereby respondents respond to surveys in certain ways regardless of the content; this can be considered a source of systematic error (Van Vaerenbergh and Thomas, 2013, p195).

While Van Vaerenbergh and Thomas (ibid, p197) identify a number of types of RSs and indicate how they can cause bias in a 7-point scale, they do not illustrate the phenomena through distribution shapes. Upon inspection of the appreciation data for Radio 4 it is possible see response *types* through visual inspection:

---

[291] See figures immediately above.

[292] See from p143.

[293] See from p122.

[294] See from p119.

7.2.1 Observed Response Types

Carrie (1997, p89, p95) found varying response types in his television AI study. AI respondents could be segmented into 3 types of scoring: 'consistent', 'regular' and 'varied'.[295] However, if Carrie's segmentation is applied to this data, we find little commonality as his analysis was done on a 6-point scale whereas the data for this study is taken from a 10-point scale:

Figure 52 Radio 4 and 4 Extra – Distribution of appreciation based on Carrie's 1997 segmentation

Category	% of Individuals Carrie's TV research	% of Individuals Radio 4 and 4 Extra
Consistent Scorers	16%	23%
Regular Scorers	59%	32%
Varied Scorers	25%	44%

While Carrie considered patterns in terms of *spread* of the data, he did not consider the *shape* of the patterns which is the approach desired for this study.

Identifying individual responses for analysis involved two steps. Firstly, 4192 of the 7690 respondents were identified as having given illegitimate responses (including those who gave fewer than 10 responses) and these were excluded. Of the remaining 3498 respondents, a further 2114 had given fewer than 100 responses and these too were excluded from this part of the analysis because they were considered too small a sample size to illustrate response patterns. This left 1384 respondents whose response patterns were each visually inspected and categorised based on commonality of shape.[296] The summary of how the response patterns *types* were proportioned in the data is shown in figure 53.[297]

Figure 53 Radio 4 and 4 Extra – Pattern of response summary – Response Types

Legitimacy	Type	Number of respondents	Mix of those analysed
Legitimate	1 unimodal	701	51%
	2 'Silvey'[298]	371	27%
	3 bimodal	229	17%
	4 erratic	39	3%
	5 2 scores	24	2%
	6 polarised	15	1%
	7 split	5	0%
	X <100	2114	
Legitimate Total		3498	
Illegitimate	- <10 responses	3955	
	- Flat line	201	
	- Straight line	18	
	- Other	18	
Illegitimate Total		4192	
Grand Total		7690	

Number of respondents analysed = 3498-2114   **1384**

Thus while in aggregate the legitimate AI data is bimodally distributed, this is merely a product of a number of response patterns: '*Technically this is a mixture distribution*' McConway, (2013). The following sections illustrate the detail of pattern types found.

---

[295] See from p108.

[296] McConway (2013) suggests that the process of identifying and categorising the response patterns might have been automated, to some extent, by constructing an algorithm to predict responses. This has not been addressed as part of this research as, '*it would be horribly complicated*'.

[297] Details of the distribution shape of each response pattern follow from p180.

[298] See p181 for explanation of this term.

i Response type 1 – Unimodal

When individually examined, over half of the respondents (51%) distributed their appreciation responses in a hump-backed, unimodal distribution. This is the kind of shape found in datasets that eliminated consumption bias, under-reporting bias and over-confidence bias (for example, Benson and Perry, 2006; Hu et al 2009;[299] and Leicester Comedy Judges).[300] The aggregate shape of the responses from these respondents alone is as follows:

Figure 54 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 1 unimodal – 701 respondents

	1	2	3	4	5	6	7	8	9	10
1 unimodal	0%	0%	1%	1%	5%	10%	23%	33%	22%	5%
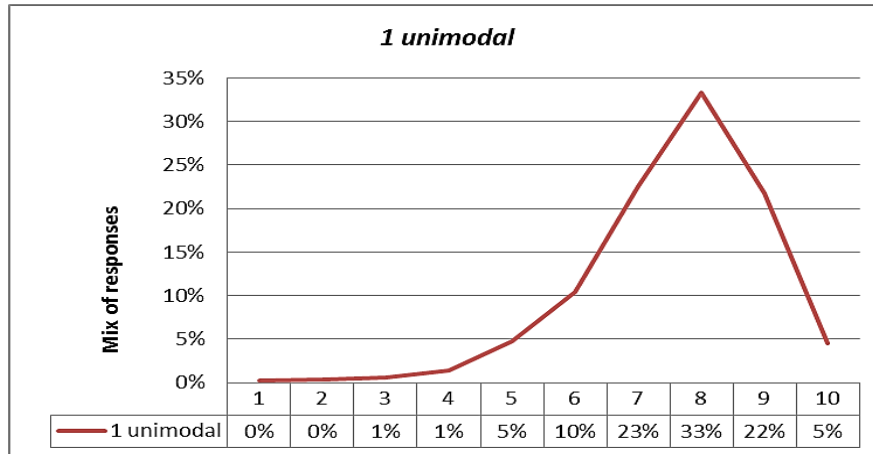
It is particularly interesting that despite this group constituting over half of the measured scores, their score distribution bears little relation to the total aggregate shape (see figure on p177), with very few scores of 10 being allocated – only 5% of the total compared to 24% for the aggregate.

While this group appear to allocate their scored in a distribution shape suitable for parametric analysis (unimodal and humpbacked),[301] even within this response type a wide variation of responses was found, for example:

Figure 55 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 1 unimodal – individuals – skews

	1	2	3	4	5	6	7	8	9	10
2586179 - Female, 69	0%	0%	0%	0%	0%	4%	10%	85%	2%	0%
10717542 - Male, 24	4%	17%	25%	22%	15%	10%	5%	2%	0%	0%
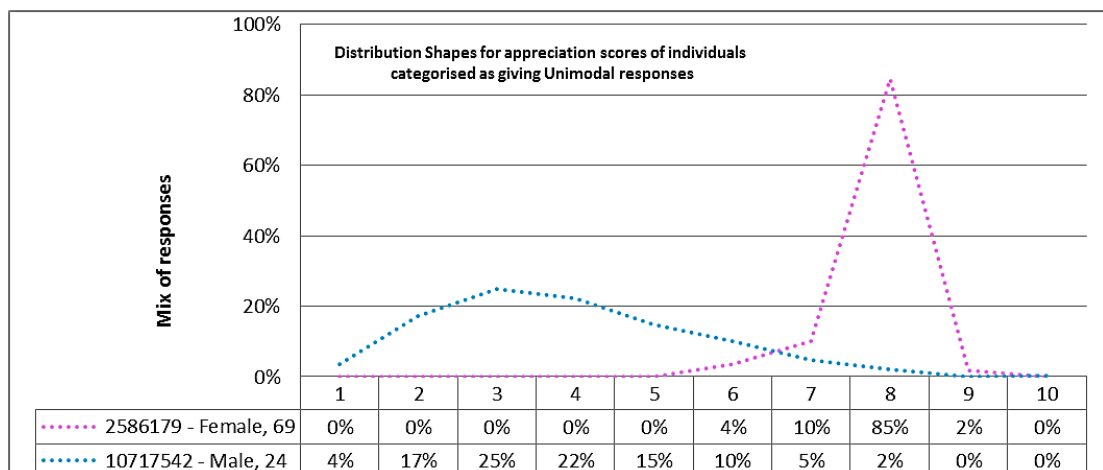
Figure 55 shows the extremes of skewness of individual response patterns that were classed as unimodal type. Respondent 2586179 has a mode at 8 and allocates 85% of programmes heard with an appreciation score, while respondent 10717542 has a mode at 3 with 83% of scores at 5 or below.

---

[299] See p155.
[300] See p156.
[301] See p156.

	1	2	3	4	5	6	7	8	9	10
⋯⋯ 1193499 - Female, 70	0%	0%	0%	1%	1%	13%	65%	19%	0%	0%
⋯⋯ 1818815 - Male, 68	0%	0%	0%	0%	0%	1%	97%	2%	0%	0%
⋯⋯ 2502511 - Male, 31	0%	3%	3%	7%	10%	20%	21%	19%	14%	3%

Figure 56 illustrates that, even distributions which are roughly symmetrical around a common mode, can be very different in terms of steepness of the curve.[302] Respondent 1818815 clearly focuses on the score of 7, using it in 97% of cases and only straying from it with 6s or 8s 3% of the time. This differs strikingly from respondent 2502511 who is far more 'promiscuous' with his scores, giving a far flatter distribution with the mode of 7 only being used 21% of the time.

Across all of the 701 respondents that have given unimodal responses, there are effectively 701 different distribution shapes as there are as many variations on skew, mode and steepness. All of these impact upon the aggregate AIs for each of the respondents. For given individuals the average AI scores range from as little as 33 (respondent 13262425, Male, 26) up to 94 (respondent 2621764, Male, 61). What we don't know from this research is whether the respondent who gives a mean of 3.3 is doing so because he thinks little of the programmes he is hearing, or because he has taken a decision that 3 or 4 is his baseline average and any programme heard will be around those scores (in this case, the respondent gave no verbatim responses to add to our understanding). While we can identify the different response patterns, using this data alone we can't say *why* they occur.

ii Response type 2 – 'Silvey'

As discussed earlier, a J-shaped curve was explained by Hu et al[303] as purchasing bias, people tend to consume what they like and like what they consume, tending to give high scores (consumption bias). Silvey (1974, p116) found, during WWII, that responses to his 0-10 radio appreciation scale also clustered around the higher ratings, He explained this phenomenon as people starting at 10 as a default and then taking points away for aspects on which they were less keen (ibid, p145).[304] Thus this pattern type identified (see figure 57) has been attributed to the appellation 'Silvey' rather than 'J-shaped' because within this group not all of the respondents have necessarily distributed their scores in a J-shape; rather, the commonality in this group is that respondents appear to have set themselves an upper limit from which is the default that they then subtract points.

---

302 Known as kurtosis.

303 See p150.

304 See from p137.

Figure 57 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 2 'Silvey' – 371 respondents
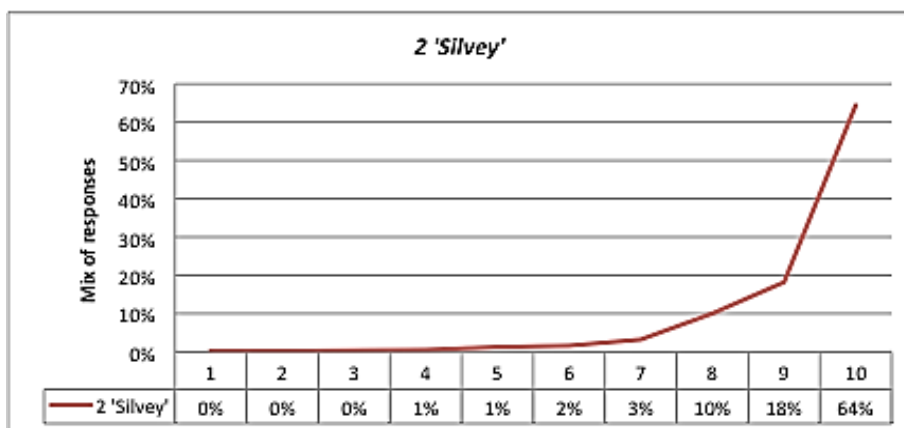
Figure 57 shows that at topline, this next group aggregates to a J-shaped distribution. It accounts for 27% of the responses analysed. As with the unimodal group, there is great variation between the respondents. For example:

Figure 58 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 2 'Silvey' – individual examples – steepness
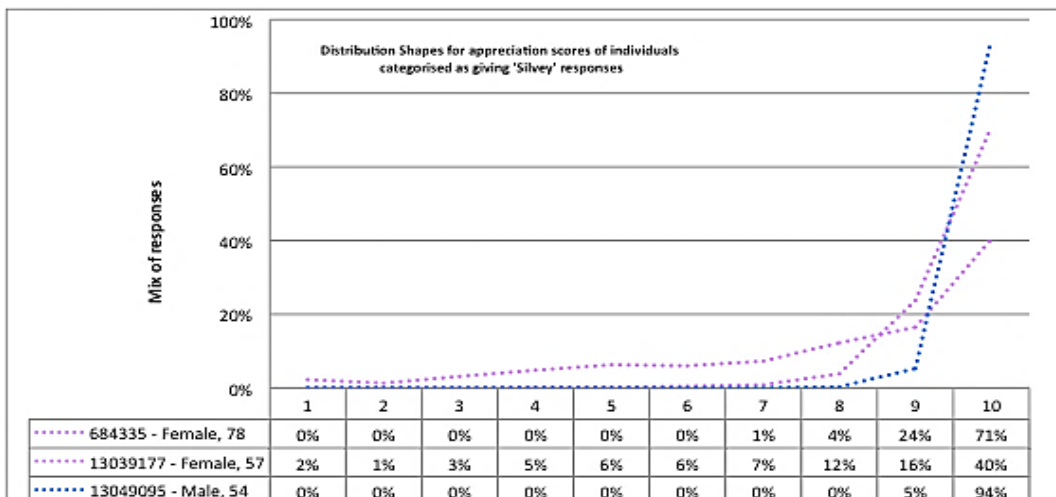
The respondents' distributions seen in the figure 58, could all be called J-shaped with an upper limit of 10 but there is a vast difference in the shape of the steepness of the 'J'.

Figure 59 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 2 'Silvey' – individual examples – non J-shaped
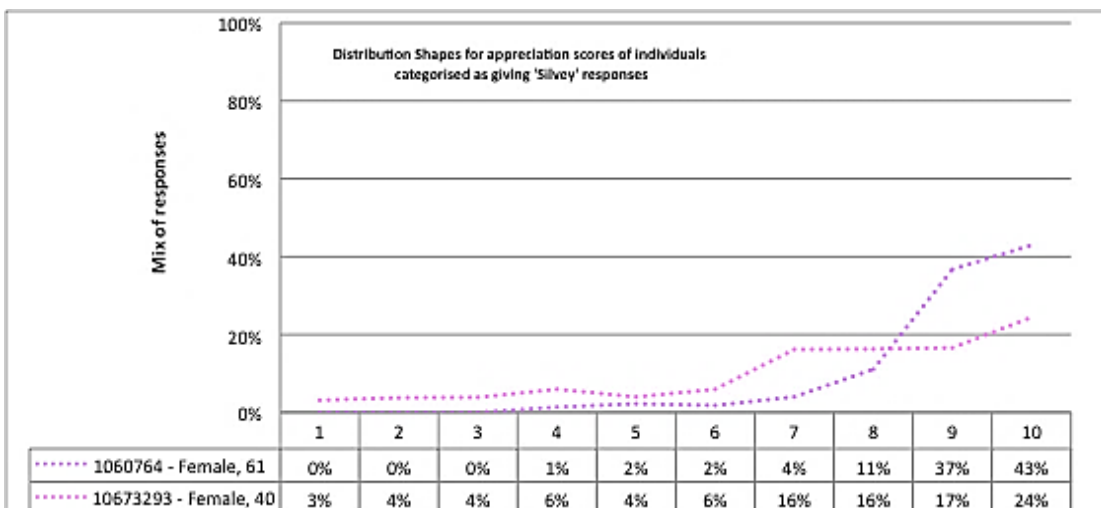
In this next figure, we can see examples of respondents who have a distribution that cannot exactly be described as J-shaped, despite having a modal score of 10 with roughly declining scores as they go down the scale.

Figure 60 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 2 'Silvey' – individual examples – mode lower than 10
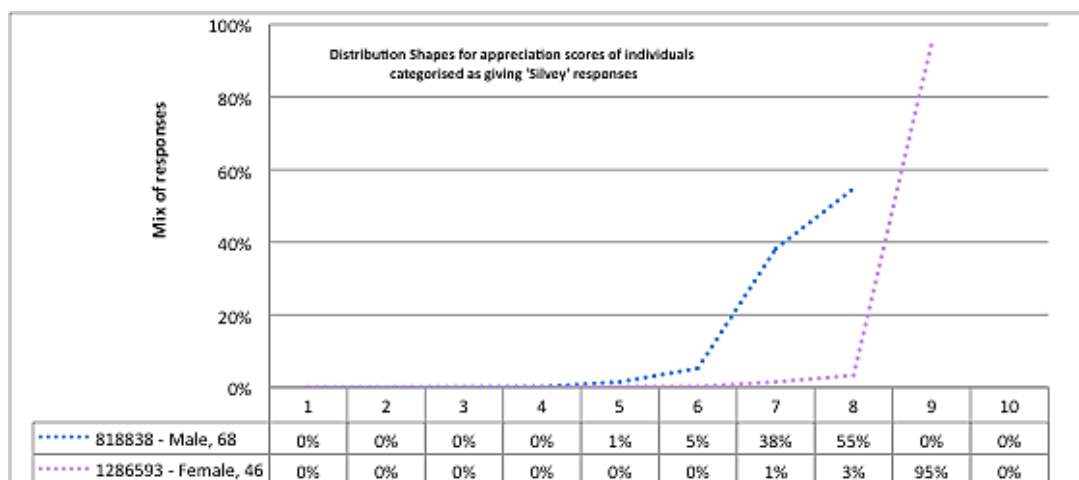
Another notable variation is illustrated in figure 60, where it can be seen that there are some individuals who do not set their default 'best' score at 10, but rather 9 (respondent 1286593) or even 8 (respondent 818838). Other respondents peaked at 7 or even 6. Again, there is variation in steepness and shape between these examples.

There is a huge range of shapes within this 'Silvey' group and the mean AI score ranges from as low as 53 (respondent 2498955), up to 99 (respondent 2651128). Respondents in this category whose mean AIs are 99 or higher are effectively marking all but the odd exception with a score of 10. These individuals account for 23% (86) of the 371 in this 'Silvey' group and 6% of the total 1384 analysed.

This type of response applies to over a quarter of the respondents. It may be that Radio 4 listeners, likely to have made their mind up that they like Radio 4 as a default position, might only give a lower score than their default high score if they hear something particularly objectionable.

While it is possible that those whose upper limit is lower than 10 may mark this way because they haven't experienced anything good enough to warrant that high a score, looking at verbatim responses indicates that they may have set themselves a 'self-prescribed' upper limit, and while programmes may be good enough to warrant a high score these respondents will not breach their imposed boundary. For example, the comments below come from respondents whose ratings never exceed 8 suggesting that 8, for them, is may be as good as it gets:

Ayres on the Air   15-Oct-12   Male (respondent 1190730)
Appreciation score = 8
'fun fun funny.'

Clare in the Community   16-Feb-12   Female (respondent 2624423)
Appreciation score = 8
'well written and well acted – very funny comedy drama.'

iii Response type 3 – Bimodal

A further response pattern observed was bimodal, i.e., distributed across a range of scores but with two distinct humps. 17% of respondents analysed fell into this group:

Figure 61 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 3 bimodal – 229 respondents

3 Bimodal

	1	2	3	4	5	6	7	8	9	10
3 binomial	2.5%	3.1%	2.8%	2.7%	8.4%	8.0%	14.2%	24.7%	14.7%	19.0%
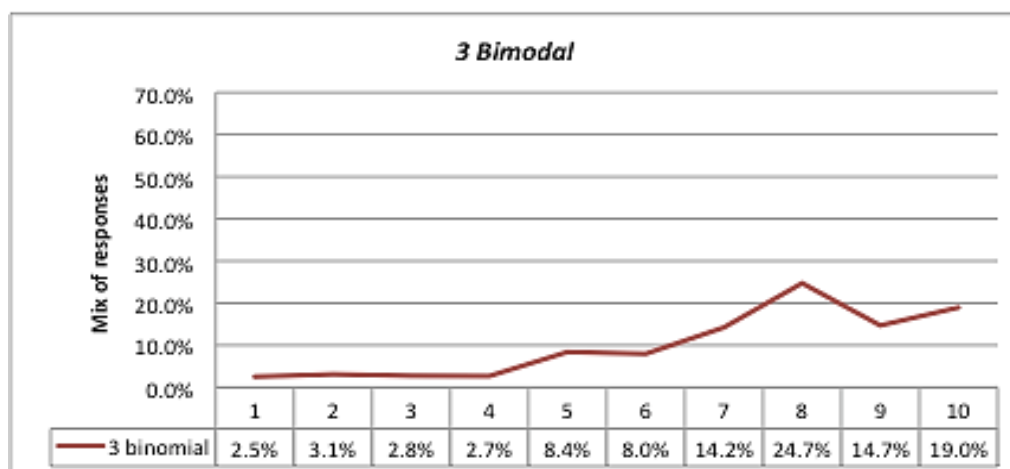
The figure above shows that the aggregate responses from these respondents has a peak at 8 but also peaks at 5 and 10. However, once again the aggregate hides a number of variations at respondent level:

Figure 62 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 3 bimodal – individuals – skews

Distribution Shapes for appreciation scores of individuals categorised as giving Bimodal responses

	1	2	3	4	5	6	7	8	9	10
1057421 - Male, 55	17%	1%	1%	2%	6%	20%	35%	14%	4%	0%
2580123 - Male, 74	2%	28%	15%	8%	9%	12%	14%	10%	2%	0%

Figure 62 shows two respondents who have very different response patterns despite the fact that both distributions are bimodal. Respondent 1057421 gives a score of 7 for over one third of the programmes rated and frequency drops away either side of this peak of 7 in a shape that we've seen in monomodal distributions. However, we can see that unlike with a typical hump-backed distribution, there is a further peak at the score of 1, which accounts for 17% of his ratings. In comparison, respondent 2580123 has a peak of distribution around the score of 2 – the lower end of the scale – then another peak at 7. Peaks have been observed at a variety of points along the scale for those in this group, for example:

Figure 63 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 3 bimodal – individuals – 'steepness'
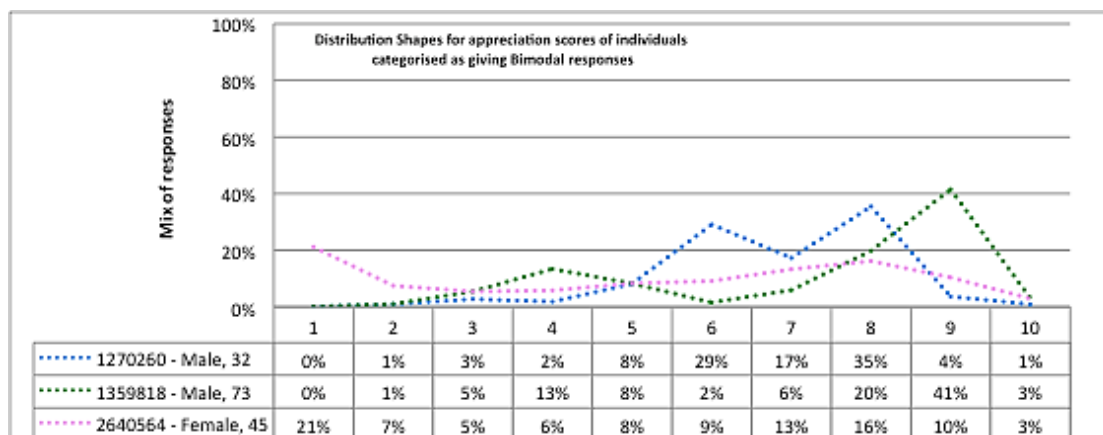
In the figure above we can see a variety of bimodal distribution shapes with peaks at 6 and 8, 4 and 9, and 1 and 8, respectively.

The wide variance in the distribution shapes within this bimodal group leaves little to help our understanding of the thought processes behind the ratings. It seems that there are scores that each respondent tends towards but not to the exclusion of other scores on the scale, a phenomenon that we will discuss when we come to '2 scores'.

These first three groups – unimodal, 'Silvey' and bimodal – describe 94% of the response patterns[305] and are therefore of particular importance. Furhter patterns observed are now presented with the caveat that each group has fewer than 50 respondents.

iv Response type 4 – Erratic

A number of respondents whose response patterns cannot be readily categorised have been segmented as 'erratic'. Of the respondents analysed, only 3% fell into this group:

Figure 64 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 4 erratic – 39 respondents


Figure 64 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 4 erratic – 39 respondents
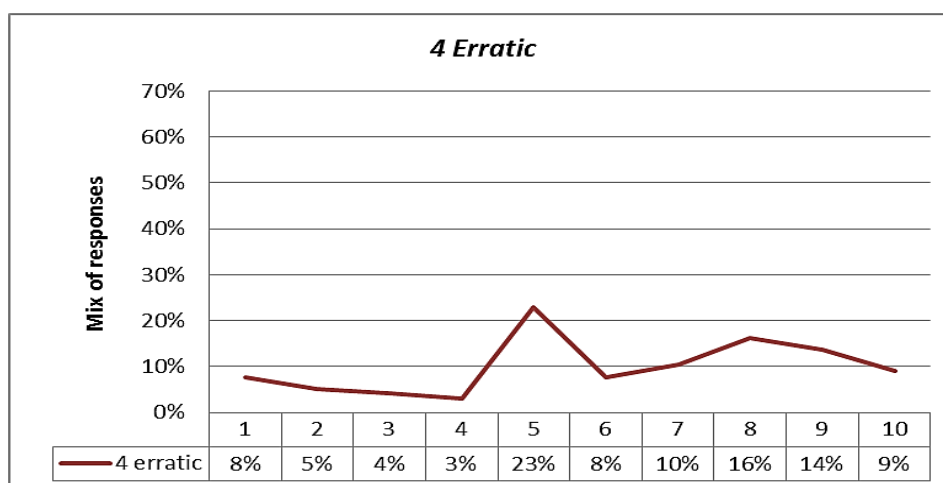
This figure shows the aggregate distribution, but to garner any understanding we must again look at individual examples:

---

[305] See p179.

Figure 65 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 4 erratic – individuals – skews

	1	2	3	4	5	6	7	8	9	10
624705 - Male, 50	10%	14%	14%	8%	12%	8%	10%	10%	11%	3%
1216036 - Female, 26	5%	2%	2%	1%	70%	3%	4%	6%	7%	1%
1317610 - Male, 54	2%	12%	39%	0%	5%	29%	0%	1%	12%	0%
1333177 - Female, 70	0%	0%	0%	0%	27%	3%	7%	19%	10%	34%
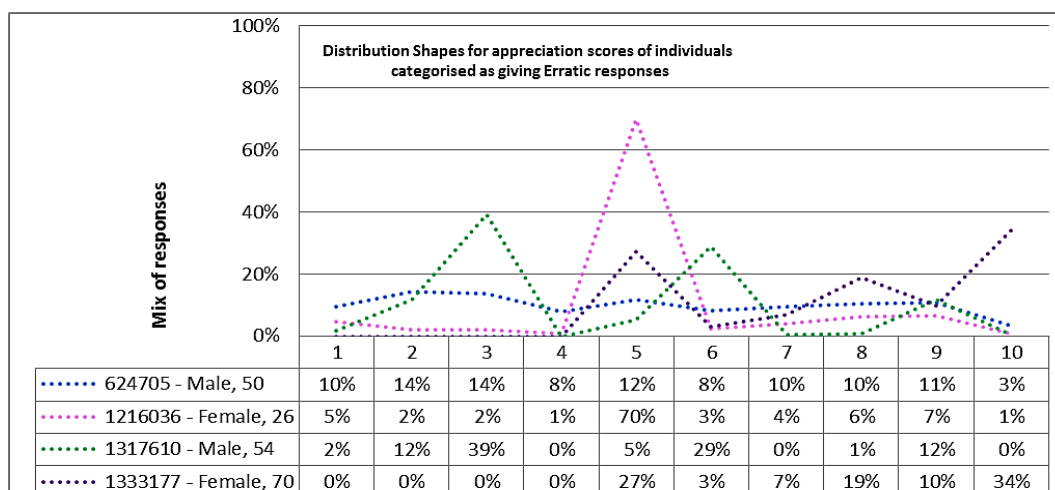
The figure above shows four respondent examples that have been categorised as 'erratic'; we can see that there is no similarity between each distribution, nor do any fit into the more straightforward categories such as unimodal or 'Silvey'. Respondent 624705, for example, has distributed his scores relatively evenly across all the scores, a pattern seen from very few respondents. In comparison, respondent 1216036 has focused on the score of 5 for 70% of her scores but there are also peaks at 9 and 1.

There is such a disparity in these responses that it is hard to see how aggregating them can result in any meaningful distribution and, it follows, the mean will not give a worthwhile measure of central tendency.

v Response type 5 – 2-Scores

While there are respondents such as 624705 who use the whole scale from which to allocate their scores (figure 65) there were a number of respondents whose scoring was limited to just 2 scores:

Figure 66 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 5 2-scores – 24 respondents

	1	2	3	4	5	6	7	8	9	10
5 2 scores	0%	0%	0%	0%	3%	9%	18%	43%	15%	12%
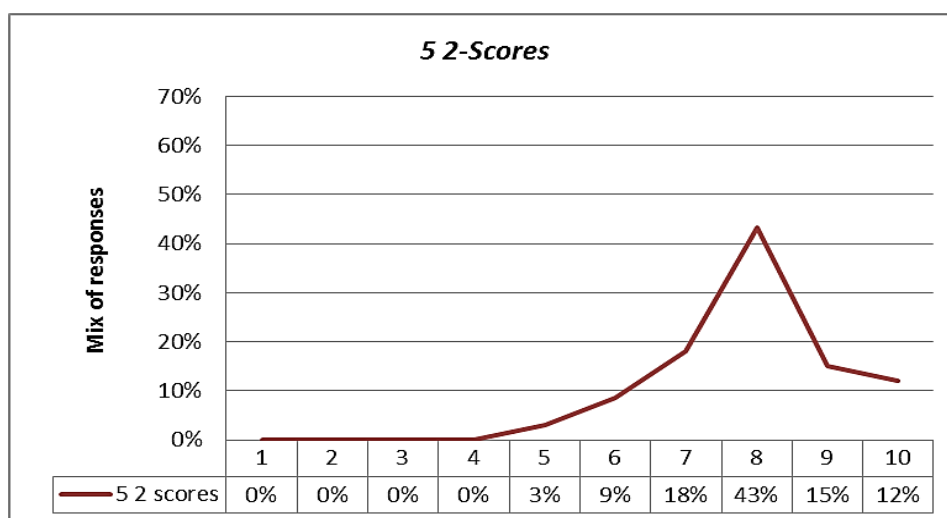
The figure above illustrates the aggregate of the responses classed as '2 scores'. While the aggregate gives us a fairly clear unimodal distribution, drilling down to respondent level shows that the 'average' is again obscuring the detail. For example, from within this group, respondent 1164029 (Male, 26) rated 305 programmes, 303 of which he gave a score of 6 and the remaining 2 a score of 7; this is a highly skewed pattern. However, in comparison, a more symmetrical pattern can be seen from respondent 13107652 (Female, 27), whose 245 responses were split 135 as 10, and 110 as 9.

None of the respondents in this category had given more than 461 responses (low compared to a respondent in another group giving over 4,000 responses), indicating that this style of scoring might not be sustainable over many responses. Perhaps the temptation to utilise a score other than the two selected could be overwhelming over a longer time period? Just 2% of respondents fell into this group.

vi Response type 6 – Polarised

A further pattern was observed, that of polarised scoring whereby respondents tended to stick to extreme scores.

Figure 67 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 6 Polarised – 15 respondents

	1	2	3	4	5	6	7	8	9	10
6 polarised	15%	1%	1%	1%	2%	1%	7%	3%	4%	68%
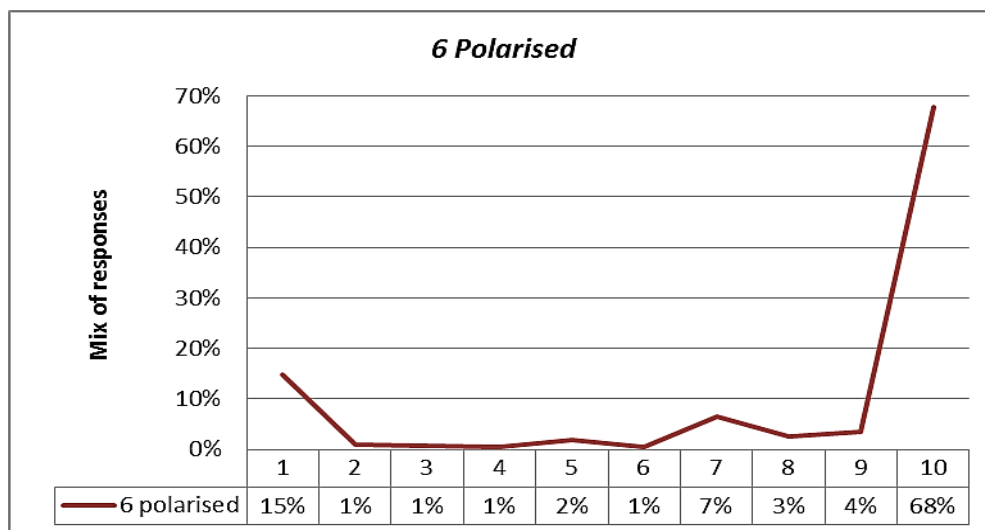
The figure above shows the aggregate scores from the respondents categorised as 'polarised'. Only 1% of analysed respondents fell into this group.

Figure 68 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 4 Erratic – individuals – skews

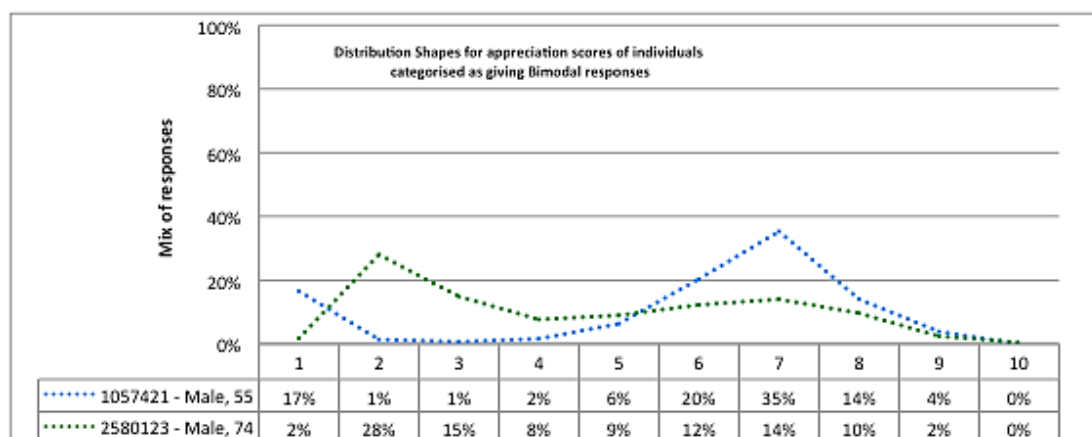Distribution Shapes for appreciation scores of individuals categorised as giving Polarised responses

	1	2	3	4	5	6	7	8	9	10
625830 - Male, 64	88%	0%	0%	0%	0%	0%	0%	0%	2%	9%
1334267 - Male, 45	0%	0%	0%	0%	12%	1%	3%	5%	6%	73%
1563530 - Female, 38	34%	2%	1%	1%	3%	1%	2%	3%	6%	48%
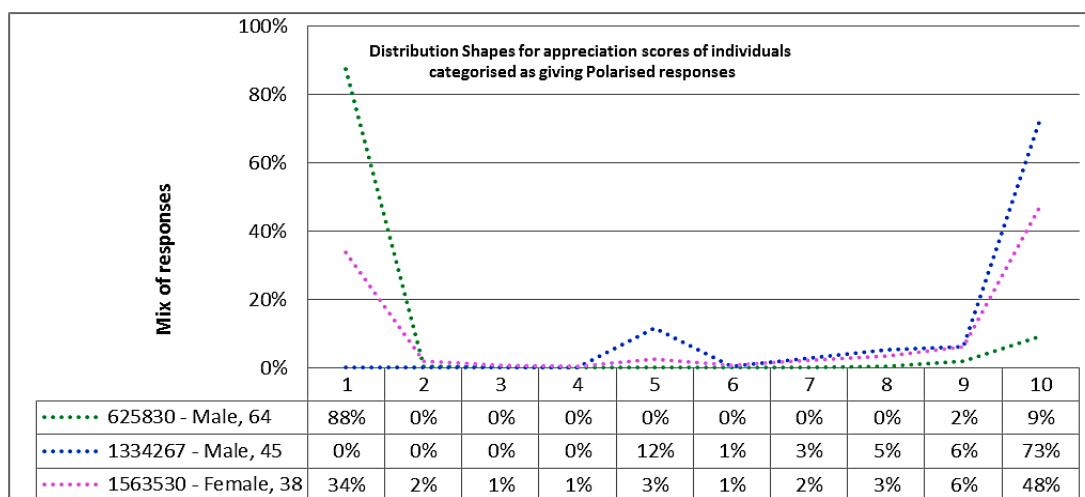
This figure illustrates some of the variations seen in this category. For example, respondent 625830 tends to mark programmes heard with a score of 1 (88% of his responses) but will on occasion give scores of 9 or 10. Alternately, respondent 1563530 is most likely to give a score of 10 (48%) but also allocates a score of 1 in many cases (34%). This category is not limited to scores of 1 and 10, as exemplified by respondent 1334267 who tends towards a score of 10 (73%) but has a lower end second peak at 5 (12%).

Extreme responses could relate to earlier discussions. As De Mooij (1998, p76: citing Hofstede) posits, those in western societies are keen to demonstrate their uniqueness and this could be seen through their

responses.[306] Furthermore, Hu et al discuss how extreme responses might be partly explained for online ratings where a respondent could be under the impression that their score might be of more interest to others if extreme.[307] This could still be applicable if we consider that the BBC Pulse respondents might believe that a BBC researcher would be more likely to take notice of them if their score is extreme, and that they would therefore have more influence than someone giving a more moderate rating. This runs counter 'consumption bias'[308] whereby we might would expect very few low scores as it is arguably unlikely that someone would consistently listen to programmes to which they give a score of 1. However, there are examples of respondents, for whatever reason, giving very low scores for Radio 4 comedies over and over again Below is some illustrative verbatim feedback to programmes rated as 1 by this respondent which suggests that these scores do indeed appear to reflect his opinion of the programmes.:

Male, 64 (Respondent 625830):

> ***The Unbelievable Truth*** 02-Jan-12
> Appreciation score = 1
> 'Yet another feeble BBC celebrity panel game – why so many ??'
>
> ***Isy Suttie: Pearl and Dave*** 11-Jan-12
> Appreciation score = 1
> 'DREADFUL'.
>
> ***My Teenage Diary*** 27-Jun-12
> Appreciation score = 1
> 'It is hard to imagine any worse programmes to put on at this time'

During 2012, this respondent claimed to have heard ***The Unbelievable Truth*** on seven separate occasions, giving it a score of 1 and very disparaging verbatim responses in every instance.

However, since only 1% of respondents fell into this group, this style of scoring is clearly not a typical pattern.

vii Response type 7 – Split

A small number of respondents gave scores which were categorised as 'split' – i.e., a range but focused on specific scores:

Figure 69 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 7 Split – 5 respondents

	1	2	3	4	5	6	7	8	9	10
7 split	2%	1%	1%	0%	19%	0%	1%	3%	1%	72%

---
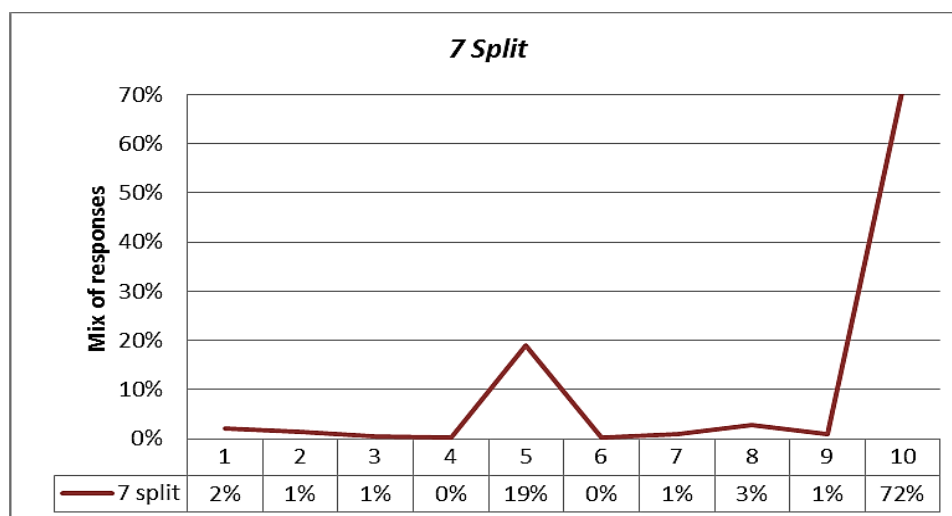
[306] See from p154.

[307] See from p153.

[308] See from p152.

As there are only 5 respondents in this group, the aggregate is not at all useful to view. A couple of individual examples can be seen below:

Figure 70 Radio 4 and 4 Extra – Distribution of appreciation scores from segmented respondents – 7 split – individuals – skews

|	1	2	3	4	5	6	7	8	9	10
10758785 - Male, 68	4%	6%	2%	1%	75%	1%	1%	4%	5%	0%
13405729 - Male, 54	8%	1%	0%	1%	4%	1%	3%	17%	0%	66%
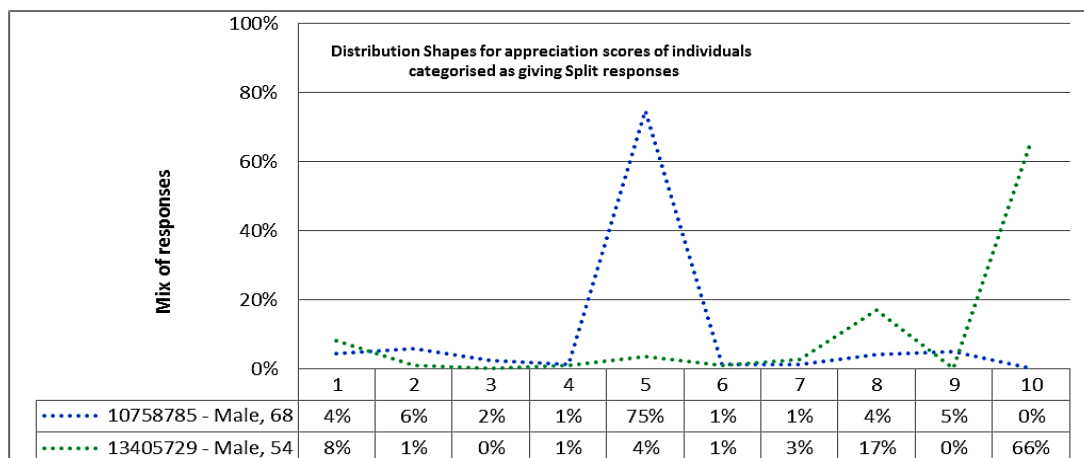
The figure above shows two examples of 'split' distributions. We can see that each has a number of peaks but there is not necessarily any other similarity between the distribution shapes. For example, respondent 10758785 has a strong peak at 5 whereas respondent 13405729 peaks at 10. There are very few of these respondents and they could arguably be included in the 'erratic' category..

Across the respondents analysed, there was a huge range of distribution patterns for Radio 4 and 4 Extra appreciation ratings. 51% of the respondents allocated their ratings in a unimodal distribution, 27% in a 'Silvey' distribution and the remainder in a range of other patterns. This analysis suggests that the bimodal distribution of the aggregate scores is a product of a 'mixture distribution'.

Theory suggests that non-alike distributions are not well-embodied by a single measure of central tendency[309] and in this respect the mean is not a good representation of appreciation responses to BBC Radio 4 comedy programmes.

7.2.2 Do Each of The Response Groups Give a Different Mean Appreciation Score?

As the key performance indicator for programme quality,[310] the resultant AI (whether or not its calculation is statistically valid) is currently important for users. We've seen earlier that aggregate appreciation varies between legitimate and illegitimate scores,[311] but is there an observable difference between each of the legitimate response types?

---

[309] See from p156 (Ehrenberg, 1986, p3-9).
[310] See from p88.
[311] See p170.

Legitimacy	Type	Number of respondents	Mix of respondents analysed	Number of responses	Mix of responses analysed	Mean Weighted AI[312]
Legitimate	1 unimodal	701	51%	255,691	49%	75.6
	2 'Silvey'	371	27%	139,339	27%	93.1
	3 bimodal	229	17%	98,469	19%	74.8
	4 erratic	39	3%	20,329	4%	61.9
	5 2 scores	24	2%	4,966	1%	76.8
	6 polarised	15	1%	4,784	1%	82.8
	7 split	5	0%	1,151	0%	82.2
	x <100	2114		76,489		78.6
Legitimate Total		3498		601,218		79.7
Illegitimate Total		4192		63,870		89.1
Grand Total		7690		665,088		80.8
Total analysed for response pattern		*1384*		*524,729*		*79.9*

Seen in figure 71, even just for the responses classified as 'legitimate', there are observable differences in the mean appreciation scores across the different response types. Types 1-3 are most worthy of attention because each group comprises of nearly 100,000 responses or more and together they account for 94% of the analysed responses. Group 1 respondents who distribute their scores in a 'unimodal' fashion have a resultant mean score of 7.56 and group 2, the 'bimodal' group, a mean score of 7.48. The difference from the 'Silvey' group mean at 9.31 is noteworthy within a 1-10 scale. The implication is that the resultant mean AI of a programme could be considerably affected by variations in the *proportion* of listeners of different response types, rather than actual 'quality' of the programme. Thus in an extreme example: a hypothetical comedy of 'average' quality, rated purely by people from the unimodal group, would achieve an AI of 76, while the 'Silvey' group would give the same comedy an AI of 93.

Different response pattern types result in different aggregate scores and therefore varying proportions of each of these types for any programme, could have an effect on the resulting AI score.

7.2.3 Do Demographic Differences Correlate to Response Patterns?

In the previous section, we can see that there are observable differences in the response patterns of individuals. Can we learn anything about the different groups? Again, we can just concentrate on the 3 most frequently observed response patterns where there are at least 229 respondents in a group. The group of particular interest is the 'Silvey' set, as their average appreciation score at 9.31 is far higher than that of the other two types.[313] We can look at whether basic demographic skews can be seen for the key groups:

---

312 AI is the expressed as 10 times the mean scores – see from p85.

313 N.B. – while 229 is a reasonable sample size for analysis, we are drilling down to lower demographic groups so some of the sub-groups consist of fewer than 20 respondents. While we can observe variances in mixes for this data set, it would be unwise to make any firm generalisations about the wider population from these figures.

Figure 72 LEGITIMATE RESPONDENTS, Radio 4 and 4 Extra – Respondent demographics and response types – Age group

Socal grade	All analysed respondents	mix	1 Unimodal respondents	mix	2 'Silvey' respondents	mix	3 Bimodal respondents	mix
Under 35	127	9%	86	12%	15	4%	19	8%
35-54s	383	28%	191	27%	101	27%	66	29%
55 and over	874	63%	424	60%	255	69%	144	63%
TOTAL	1384	100%	701	100%	371	100%	229	100%

Figure 72 shows that, compared to all analysed respondents, the make-up of the 'Silvey' group varies more in its age profile than that of the other two groups. We can see that these respondents skew slightly older than a typical legitimate respondent; the group consists of 69% 55+ whereas they account for only 63% of all respondents. Similarly, unimodal responses tend to skew younger than the other two types. So, for example, we could theorise that a Radio 4 comedy that appeals particularly to an older audience might result in a higher AI purely because it attracts a higher proportion of older listeners; this is a group who are more likely to use a response pattern that tends to have a higher appreciation score. The latter may thus be inflated due to audience mix and their scoring tendency, rather than due to a programme's actual 'quality'.

Figure 73 LEGITIMATE RESPONDENTS, Radio 4 and 4 Extra – Respondent demographics and response types – Social group

Socal grade	All analysed respondents	mix	1 Unimodal respondents	mix	2 'Silvey' respondents	mix	3 Bimodal respondents	mix
AB	676	49%	341	49%	170	46%	119	52%
C1	368	27%	194	28%	108	29%	49	21%
C2	154	11%	65	9%	45	12%	33	14%
DE	186	13%	101	14%	48	13%	28	12%
TOTAL	1,384	100%	701	100%	371	100%	229	100%

The 'Silvey' scoring pattern, is slightly less likely to be utilised by the ABs than is typical.

Figure 74 LEGITIMATE RESPONDENTS, Radio 4 and 4 Extra – Respondent demographics and response types – Gender

Gender	All analysed respondents	mix	1 Unimodal respondents	mix	2 'Silvey' respondents	mix	3 Bimodal respondents	mix
Men	744	54%	374	53%	176	47%	138	60%
Women	640	46%	327	47%	195	53%	91	40%
TOTAL	1,384	100%	701	100%	371	100%	229	100%

Figure 74 shows that women are more likely to rate according to the 'Silvey' pattern.

So these figures indicate that the 'Silvey' pattern, which gives a higher AI, is more likely to be used by older women outside the AB social class. What is particularly interesting is that these demographic skews broadly relate back to findings from earlier research, as discussed on p105. While one explanation has been that certain groups rate more highly because the output is more suited to them (North, 2010), could it in fact be that the differences are could be attributable in part to response patterns? This is expanded upon from p201.

There appear to be demographic skews between response patterns. Women, those not of high socio-economic status and older people are more likely to use a response pattern that results in a higher AI. This finding suggests a reason behind evidence in earlier research[314] that has found these groups rate more highly than average. However, it does not explain the motive behind that reason. For example, we might theorise that women rate programmes more highly partly *because* they are more likely to use a response pattern that results in a higher AI, however we do not know *why* they are more prone to do this.

---

314 See from p105.

7.3 Can AIs Still Be Used To Represent Appreciation?

The evidence indicates that overall distribution of appreciation scores is not unimodal, and at respondent level only around half of the sample distributes their ratings in such a manner. Conservative statistical views tell us that parametric analysis, such as the mean, should not be applied to such data.[315] Ehrenberg (1986, p9) accepts that means can be compared within non-standard distribution shapes but not across different shapes (i.e. mixture distributions) and we have seen earlier that, when we look at respondent level, our data has a wide range of different distribution shapes.[316]

Using the means of these disparately distributed appreciation scores taken from the Pulse survey is exactly what the BBC does in order to calculate the AI scores, which are then used as a measure of programme or channel quality. Statistically, validity is definitely compromised in doing this - ideally the data should be expressed through non-parametric summary figures.

Given that the use of AIs has been established procedure at the BBC for over 70 years, it might be difficult to change the culture. An alternative is to consider whether the mean might be a useful 'proxy' for a figure that *does* have meaning. The following section analyses the Radio 4 comedy appreciation scores (42,251 responses across 132 different programmes from 2446 respondent IDs) to assess the validity of use of mean AIs.[317]

7.3.1 Is The Mean A Valid Summary Figure For Comparison of Programme Quality?

The analysis of AI scores by respondent found that different respondents allocate their scores in different patterns. This section examines AI scores for individual programmes to assess whether for each programme level the scores aggregate to a unimodal distribution. They do not, and the following charts illustrate how disparate the distribution shapes can be:

Figure 75 LEGITIMATE RESPONSES Radio 4 comedy shows – score distribution 1 (n>100)
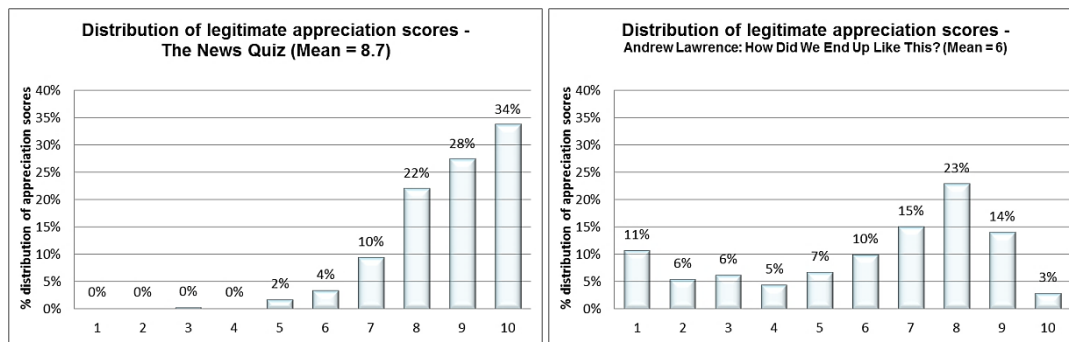
In the above example, the means are very different, so we might not be surprised to see varying shapes. However, even if their mean is roughly the same, two shows can still be very different in their spread of responses.

---

[315] See from p141.

[316] See from p180.

[317] Of the 132 programmes, 65 have 100 or more responses.
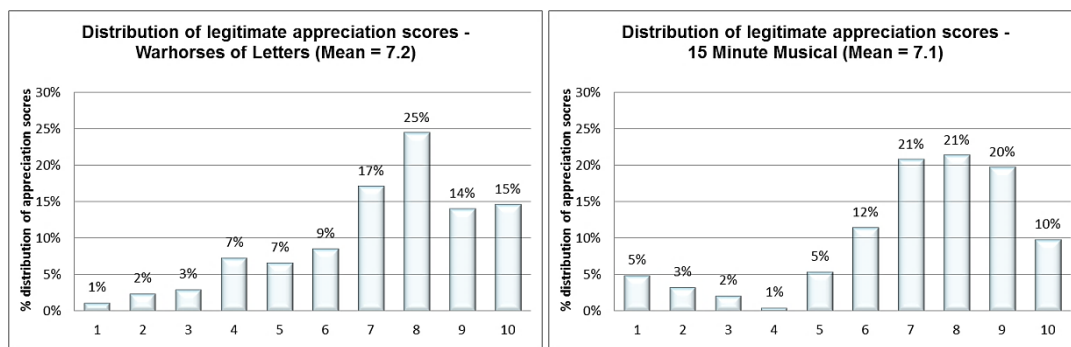
In figure 76 the AI scores for each of the two programmes are 71 and 72 respectively. At face value, to compare their means would indicate little difference between them. However, any broadcaster might be concerned, in particular, with the number of respondents who really *dislike* a programme. Looking at the mix of responses where the score is just 1, we can see that **Warhorses Of Letters** garnered just 3% of its ratings from the score of 1s and 2s whereas **15 Minute Musical** drew 8% of its ratings from these scores. If we were just comparing the means of these shows, as is done with AI scores, we might not understand the nuances of the responses. While the standard deviation can be a useful expression of a distribution's spread (see Bryman and Cramer, 1994, p88), once again, with our data having a variety of shapes – some of them not unimodal – it is not a statistically useful measure of comparison.[318]

The mean alone is not a good summary of Radio 4 comedy appreciation scores at programme level.

7.3.2 Can Appreciation Scores be Better Expressed than Just the AI (mean)?

The mode and the median are legitimate summaries of data that are not necessarily distributed in a unimodal fashion nor based on an ordinal scale (see Bryman and Cramer, 1994, p191). They are simple, common and easily understandable to the layperson. The Net Promoter Score (NPS)[319] is also non-parametric, and it is claimed that it can be applied to any satisfaction/dissatisfaction scale (Barwise and Meehan, 2011, p56). These figures are not currently calculated for the BBC Pulse survey but if we do so for the 2012 Radio 4 comedy data (legitimate scores only) the following can be found:

Mean	7.8	
Weighted mean	7.8[320]	(attempt to make the sample representative)
Mode	10	
Median	8	
Net Promoter	25	

---

[318] See from p156.

[319] See from p197.

[320] The weighted mean is the same as the mean for this dataset. Weighted means are used in this analysis unless otherwise stated

i The Mode

In the figures above, the mode is 10 (24% of all the scores given were 10s). However, programmers are concerned with being able to use summary figures in order to compare the relative 'goodness' of radio shows.[321] Is the mode useful for that purpose? Just because a response is common, it does not mean that it is typical. De Vaus (2002, p222) tells us that there are issues with using the mode,[322] so is the mode useful for evaluating programme appreciation scores?

The mode is simple: the most popular response. Below the distribution of two shows that both have a mode of 10 can be compared – figures 78 and 79:

Figure 78 LEGITIMATE RESPONSES Radio 4 comedy shows – individual shows – Clue (n=3431)
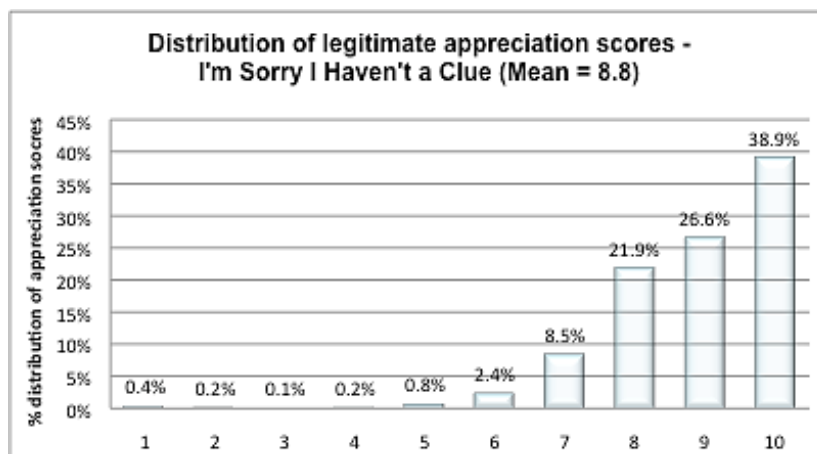
Figure 79 LEGITIMATE RESPONSES Radio 4 comedy shows – individual shows – Milton Jones (n=164)
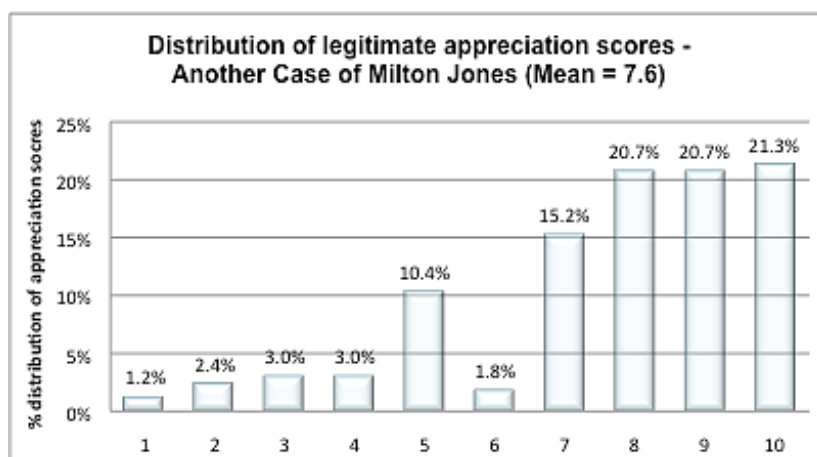
Clue (figure 78) is reasonably represented by a score of 10 as it is by far the most frequently allocated rating: 39% of all ratings. However, when we look at *Another Case of Milton Jones* (figure 79), while 10 is technically the mode it is so only by a narrow margin. In this case, 10 is not particularly representative as both 9 and 8 were very nearly as popular. Where scores are polarised, the mode is particularly unrepresentative:

---

[321] See from p88.

[322] De Vaus (2002, p222) details issue with using the mode: Picking the most common response does not tell us how typical it is. 2 Some distributions have more than one mode so it is not possible to use one figure to summarise these bimodal or multi-modal distributions. 3 The mode is very dependent on how the categories of a variable have been collapsed. If we combine some categories, we can get a lot of people in some categories, and thus this combined category would become the mode. In this sense, the mode is unstable and open to manipulation.

Distribution of legitimate appreciation scores - Count Arthur Strong's Radio Show! (Mean = 6.2)
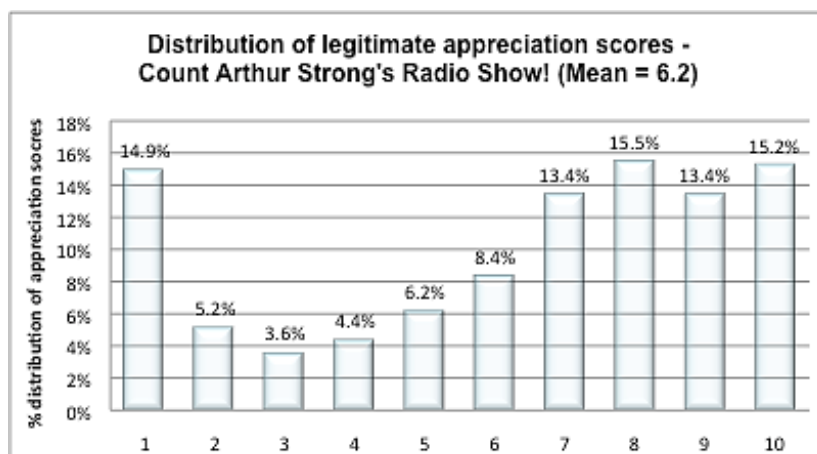
For *Count Arthur* above, the mode of 8 tells us nothing of the high proportion of listeners (30%) who are rating the show with a score of either 1 or 10 – a classic case of a comedy that polarises the audience.

It is possible to have more than one mode and, if so, which should be presented as the summary figure?

Figure 81 LEGITIMATE RESPONSES Radio 4 comedy shows – individual shows – Clare (n=481)

Distribution of legitimate appreciation scores - Clare in the Community (Mean = 7.4)
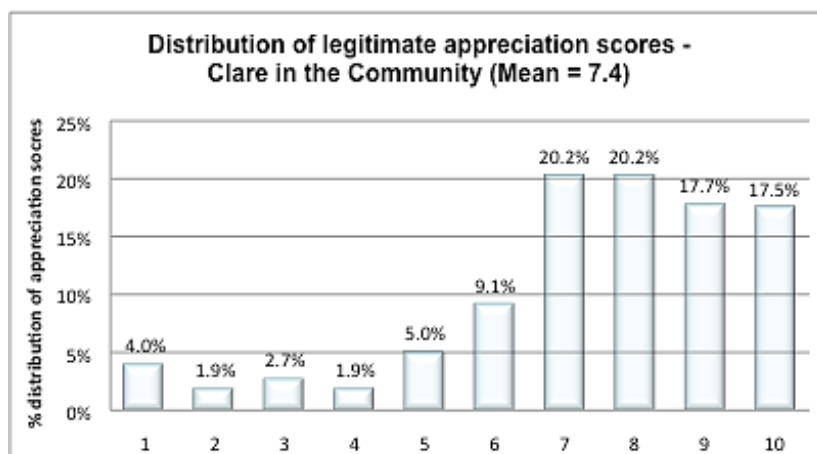
The distribution for *Clare in the Community* reveals two modes: 7 and 8. Analysis in Excel results in the system displaying the mode that it finds *first* from the list, so in this case the mode returned was 8. This is not necessarily an accurate summary of the data. Furthermore there could, in theory, be as many modes are there are points on the scale. For example, in figure 82 – *On the Hour* – there are three modes: 7, 8 and 9. How would this best be summarised? (NB: the example following is based on a sample size <50, and is therefore just for illustration).

Distribution of legitimate appreciation scores - On the Hour (Mean = 7)
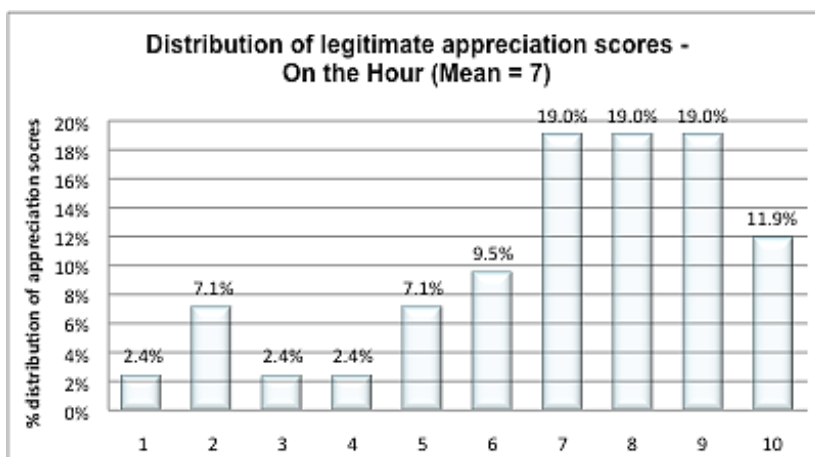
This analysis suggests that modal scores provide a poor representation of AI scores and this is in line with statistical theory which suggests modes can be easily manipulated simply by combining categories arbitrarily, and in this respect is unstable (De Vaus, 2002, p222).

The mode is not a good summary of Radio 4 comedy appreciation scores.

ii The Median

The Median is a 'rigorous but simple' procedure (McConway, 2010) to calculate a typical reading from ordinal or interval data (De Vaus, 2002, p361). It is calculated by finding the middle value in a ranked set of cases (Ehrenberg, 1986, p9). The median is preferable to the mode as it additionally takes into account the ordinal nature of the data and is therefore more informative (De Vaus, 2002, p223).

Across the 132 programmes, the medians were distributed as follows:

Figure 83 LEGITIMATE RESPONSES Radio 4 comedy shows – median score distribution

Count of Median	
Median	**Total**
5	1
6	6
7	53
7.5	4
8	59
8.5	1
9	8
Grand Total	**132**

If we were to use the median to summarise a programme's goodness, we would find that the majority of shows – 88% – were simply described by a 7, 7.5 or 8. While this is an 'appropriate' summary figure for ordinal data (De Vaus, 2002, p361), does it facilitate comparisons between different programmes? Both of the following 2 charts have the same median score of 8:
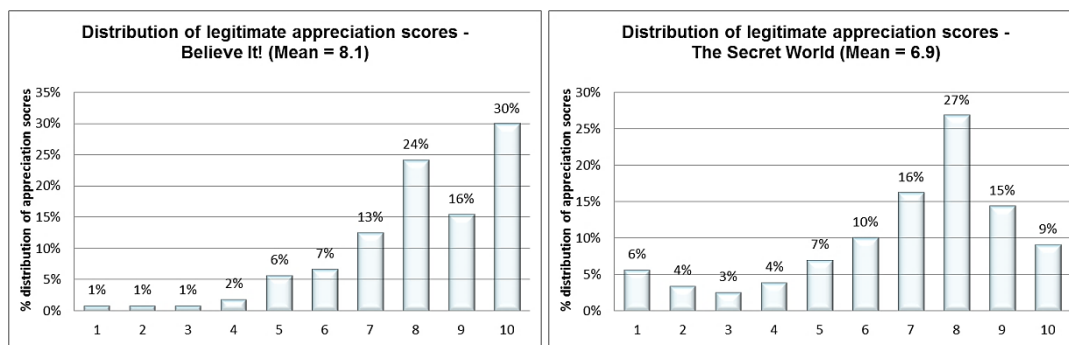
The two programmes above both received a ratings distribution with a median of 8. The first chart, **Believe It!**, has a bimodal distribution with a mode of 10. This is a very different distribution of response from **The Secret World** which, although also bimodal, is more polarised. It has a mode of 8 but has a second peak at 1. A radio programmer might be particularly interested in knowing about the polarising nature of **The Secret World**: i.e., 10% of respondents appeared to dislike it**,** which the median of 8 would not reveal. Thus using a median of 8 to summarise both of these datasets obscures information about the distribution for each programme. The median is a comparatively insensitive measure (Bryman and Cramer, 1994, p84).

> The median is not a good summary of Radio 4 comedy appreciation scores at programme level

iii The Net Promoter

The analysis thus far has advanced evidence that programme appreciation scores are not consistently distributed in a unimodal fashion and concluded that typical layman-friendly, non-parametric measures such as the mean, mode, median are inadequate summaries of the data for programme makers, regulators or the public to use. This section examines whether an alternative measure - the Net Promoter Score (NPS) would provide better representation of the data. The advantage of the Net Promoter Score is that it is statistically robust for datasets that aren't necessarily unimodal

NPS is a specific measure of customer advocacy based on asking consumers whether they would recommend a product or service to a friend or colleague with the likelihood rated on an 11-point scale of 0-10 (Kirby & Samson, 2008, p17). These scores are then split into groups as follows:

> 0-6 = detractors
> 7-8 = passives
> 9-10 = promoters

The NPS is calculated by subtracting the % number of detractors from the number of promoters. The score can range from -100 to +100.

This calculation principle could be applied to appreciation scores to create a single figure. Although the Pulse appreciation question asked of the respondents is not whether they recommend the programme, it is claimed that NPS-style calculations can be applied to any satisfaction/dissatisfaction scales (Barwise and Meehan, 2011, p56) including appreciation ratings (Collins, 2010a). NPS calculation would give an idea about the relationship between those that dislike the programme and those that really appreciate it (for appreciation scores, assuming detractor = 1-6 rather than 0-6 because of scale difference). The benefit would be that it would give greater importance to extreme scores compared to a mean calculation.

This section takes the example of programmes used in the previous sections to show how we might express the appreciation scores using the Net Promoter:

	NPS
All Radio 4	**26**
Radio 4 Comedy	**25**
15 Minute Musical (n=182)	8
Andrew Lawrence: ... (n=394)	-32
Another Case of Milton Jones (n=164)	21
Believe It! (n=103)	30
Clare in the Community (n=481)	11
Count Arthur Strong's Radio Show! (n=873)	-16
I'm Sorry I Haven't a Clue (n=3431)	60
On the Hour (n=42)	-2
The News Quiz (n=6529)	53
The Secret World (n=227)	-10
Warhorses of Letters (n=163)	2

Compared to the mean, the NPS provides a greater *range* of results. While the mean scores for the Radio 4 comedy shows in our dataset (with sample sizes of at least 100) range from 58 to 88, the Net Promoter Scores range from -38 to +61. As discussed (see from p145) a criticism of AIs was that they are not sensitive enough, tending to cluster around 60-80 (see Silvey, 1974, p67-68; Barwise and Ehrenberg, 1988, p50). Using Net Promoter Scores could contribute increased sensitivity.

However, despite this benefit there are a number of issues that preclude the use of the NPS over the AI (mean). There is the practical disadvantage of requiring the disruption of the AI system used for 70 years. In the chart above, the 'norm' for Radio 4 comedy is a Net Promoter Score of 25; moving from a figure of '81' to '25' could prove hard to understand for those not conversant with the theories or mathematics behind the move. Also, as with the mode,[323] Net Promoter Scores have instability with regard to the cut-off points for each group. Furthermore, NPS still has the same issue as the mean in that presenting it as a sole figure means that polarised scores can be obscured (Kirby and Samson, 2008, p17). By its very definition it is presenting a 'net' score so, if there are many people who love and many that hate a programme, the net result is the same as if *no* people love it or hate a programme; the net promoter score would be the same in both instances but each based on very different score distributions.

The Net Promoter Score is arguably more valid than the mean but can be unstable.

It produces a figure of greater sensitivity than the mean.

In terms of presentation as a single measure, it is subject to the same issues as the median, mode and mean whereby polarised scores are not well illustrated.

iv Presentation in Chart Form[324]

The mean, mode, median and NPS have been found to be inadequate to describe Radio 4 comedy programme appreciation on their own. To ensure that the collected data is useful, it needs to be presented in a way that allows the user to garner understanding of the spread of the scores, and understand whether a programme is inciting polarised responses. While a line chart (for example figure 49) does this, it would be difficult to compare many examples on a single chart. The following section discusses how the data might best be represented with a figure of central tendency presented *alongside* a representation of the spread of the data.

---

[323] See from p194 (De Vaus, 2002, p222).

[324] There are many ways to present the non-parametric data available, such as percentiles, the range, or box and whisker plots (Svensson, 2001, p47). However, after consideration and consultation with statistical authorities (McConway, 2013) a comparison chart showing the mixes of scale groups, such as figure 87 (i.e., the scores are grouped 1 and 2, 3-6, 7 and 8, 9, 10), indeed was confirmed as a clear way to summarise and convey the data. This style of presentation will henceforth be known a gromic (grouped mix comparison chart).

Firstly topline figures should be included to contextualise the data; anything higher than the 'norm' indicates something particularly appreciated, and scores below this indicate poorer overall appreciation.

Taking the NPS categories as a starting point to present the data spread (0-6, 7-8, 9-10) these cut-off points are arbitrary (Collins, 2010a). Kirby and Samson (2008, p17) highlight this issue: 'are people giving a 0 rating equally detracting as those with a 6? Intuitively: no'. For radio comedy, it could be useful for a producer to know the proportion of those who rate with a 1 or 2 specifically, as that might indicate a real dislike. Likewise, as North (2010) indicated, since a rating of 10 is measurably different from a 9,[325] radio producers might find it useful to distinguish between the highest scores. To analyse the data for Radio 4 comedy, therefore, it seems sensible to take the standard points but further segment some of the groups:[326]

<div style="text-align:center">

1-2 = ultra-detractors
3-6 = detractors
7-8 = passives
9 = promoters
10 = ultra-promoters

</div>

The charts below illustrate examples of how the data could be presented.

Figure 86 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Table presentation – individual shows

	ultra-detractors	detractors	passives	promoters	ultra-promoters	NPS	AI
All Radio 4 80 (n=509,098)	**2%**	**16%**	**39%**	**20%**	**24%**	**26**	**80**
Radio 4 Comedy 78 (n=42,251)	**4%**	**15%**	**37%**	**20%**	**24%**	**25**	**78**
15 Minute Musical 73 (n=182)	7%	20%	39%	24%	11%	8	73
Andrew Lawrence: .. 59. (n=394)	17%	31%	37%	13%	3%	-32	59
Another Case of Milton Jones 76 (n=164)	4%	18%	35%	21%	22%	21	76
Believe It! 82 (n=103)	2%	13%	38%	15%	31%	30	82
Clare in the Community 73 (n=481)	7%	17%	40%	17%	18%	11	73
Count Arthur Strong… 61 (n=873)	22%	23%	28%	13%	15%	-16	61
I'm Sorry I Haven't a Clue 88 (n=3431)	1%	4%	31%	27%	38%	60	88
On the Hour 69 (n=42)	11%	23%	34%	18%	14%	-2	69
The News Quiz 86 (n=6529)	1%	7%	32%	27%	34%	53	86
The Secret World 69 (n=227)	7%	25%	44%	16%	7%	-10	69
Warhorses of Letters 73 (n=163)	3%	25%	41%	16%	15%	2	73

Figure 86 includes a lot of data but does not convey it well. Figure 87 displays the same information, but the graphical representation makes the comparisons easier to read. The red/pink colours represent the lower scores (red = danger) and green, the higher (green = good). Only one summary figure is included in order to minimise confusion. Anticipating the discussion of analogies between NPS and AI (below),[327] the AI alone is used here.[328]

---

[325] See from p147.

[326] The Net Promoter Score will still be calculated from the difference between the 9-10s and the 1-6s.

[327] See from p201.

[328] This chart style will be referred to from herein as 'gromic' short for 'grouped mix comparison chart'.

Figure 87 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic presentation – individual shows
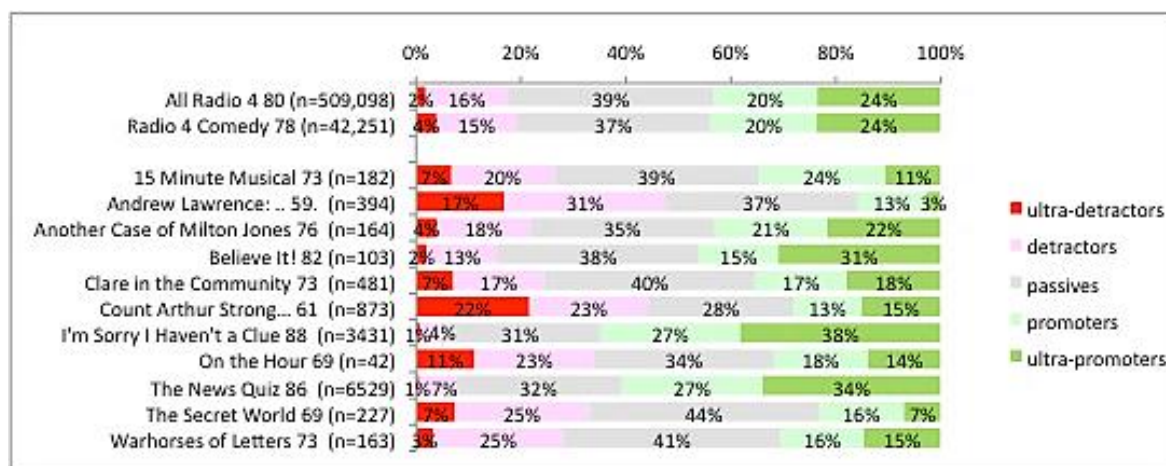
In both these examples, we can see how specific Radio 4 comedy shows compare both with each other and with the 'norms' (the 'All Radio 4' and 'Radio 4 aggregates). A good example for comparison is that between **Andrew Lawrence: How Did We End Up Like This?** and **Count Arthur**. If we compare the means alone, they are 59 and 61 respectively, i.e., very similar scores but much lower than the 78 'norm' for Radio 4 comedy. However, the presentation style in figure 87 reveals clearly that shows have very different response distributions.

Figure 88 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic presentation – individual shows (**Andrew Lawrence** and **Count Arthur**)
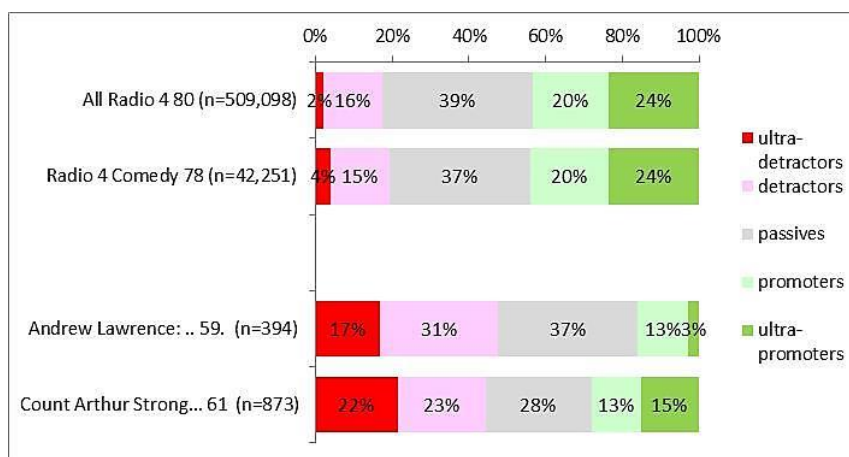
If we look at the % of ultra-detractors for each, there are more people who have a strong dislike for **Count Arthur**, i.e., 22% of people who heard **Count Arthur** gave it a score of 1 or 2, compared with 17% for **Andrew Lawrence.** On the face of it this might make **Count Arthur** a candidate for decommission. However, when we look at the ultra-promoters, i.e., the mix of 10s, we can see that 15% of all the scores were 10 whereas, for **Andrew Lawrence**, this was only 3%. Thus, while **Count Arthur** garners a lot of dislike it also gets a lot of love, whereas **Andrew Lawrence** has fewer who really dislike it but far fewer who love it.

The ability to look at the data in this fashion allows clear identification of 'polarising' programmes. This helps inform the academic and practitioner discussion of the potentially divisive nature of comedy. Earlier discussion (see from p48) shows how this debate has implications for BBC comedy commissioning – programmers and producers want not just popularist formats, but also programmes that could well incite dislike, to show that boundaries are being pushed and risks are being taken.[329]

---

[329] See from p48.

7.3.3 Are AIs a Proxy For Statistically Valid Metrics?

We have discussed how a conservative view tells us that the AI is theoretically not a valid metric when used across varying distribution shapes, [330] but could the mean scores be a good 'proxy' measure for metrics that are indeed statistically valid? One way to test this is by calculating the correlation[331] between mean AI scores and non-parametric measures:[332]

Figure 89 For all Radio 4 comedy programmes (132 series) – legitimate scores only[333]

Correlation between mean score and the <u>mode</u> –	0.66
Correlation between mean score and the <u>median</u> –	0.88
Correlation between mean score and the <u>Net Promoter Score</u> –	0.97

Thus, we can see that there is a positive correlation between the mean and common non-parametric, robust, summary metrics. By far the strongest correlation is seen between the mean and the Net Promoter Score. However, if we exclude any series that has fewer than 100 responses, the correlations strengthen:

Figure 90 For Radio 4 comedy programmes where responses >99 (64 series) – legitimate scores only

Correlation between mean score and the <u>mode</u> –	0.72
Correlation between mean score and the <u>median</u> –	0.92
Correlation between mean score and the <u>Net Promoter Score</u> –	0.99

The mean has an extremely high correlation of 0.99 with the Net Promoter Score. This suggests that for programme-level analysis, the mean and the Net Promoter Score are virtually interchangeable.

7.3.4 Does Response Pattern Dictate the Response?

We have found that there are observable differences in response patterns across respondents. In this respect it could be argued that the *primary* predictor of an appreciation score might not be the quality of the programme, but rather, the kind of response pattern towards which an individual tends. For example, respondent 1286593 (see response pattern on p183), never (within this dataset) allocates a score of 10, even when their verbatim responses are highly appreciative of a programme, 8 being their self-imposed maximum). However, as there *are* observable differences between the appreciation scores of different Radio 4 comedy programmes, this suggests that the nature of a programme *does* have an effect upon the appreciation score. Respondents appear to set a subjective 'range' of their own within the 1-10 available and use personal perceptions of what a good or bad score is. Thus, a 'good' show might garner a respondent's

---

[330] See from p141.

[331] Correlation measures the strength of the relationship between sets of numbers.

[332] See p197.

[333] The figures were exactly the same when calculated for the weighted or unweighted scores.

maximum score, which could be 10 for one person but 7 for another. A 'bad' show might likewise draw *relatively* bad scores. Take, for example respondent 1286593, in all her 1628 responses for Radio 4 and 4 Extra, she did not give any programme a score lower than 5. This can be contrasted with respondent 818838, who gave scores as low as 3. A 5 is 'bad' for respondent 1286593, but is perhaps not for respondent 818838, as it is two points higher than the lowest score they give.

> It can be argued that both the response type/pattern of the individual *and* the 'goodness' of a programme contribute to their chosen score. It seems likely that that the respondent rates the programme according to its 'goodness' but does so only within their response pattern (rating criteria), giving comparative responses rather than absolutes.

Effect of Response Patterns upon the Mean.

While it has been found that aggregate means are affected by the response pattern,[334] will subjective differences in individuals' response pattern types have a *consistent* effect on the mean? For example, if an individual person rates a programme, will the *degree* to which their scoring affects the mean be dependent upon their response pattern?

Figure 91 Radio 4 and 4 Extra – Distribution of all appreciation scores from two individual respondents – effect upon the mean (n>100 for each respondent)
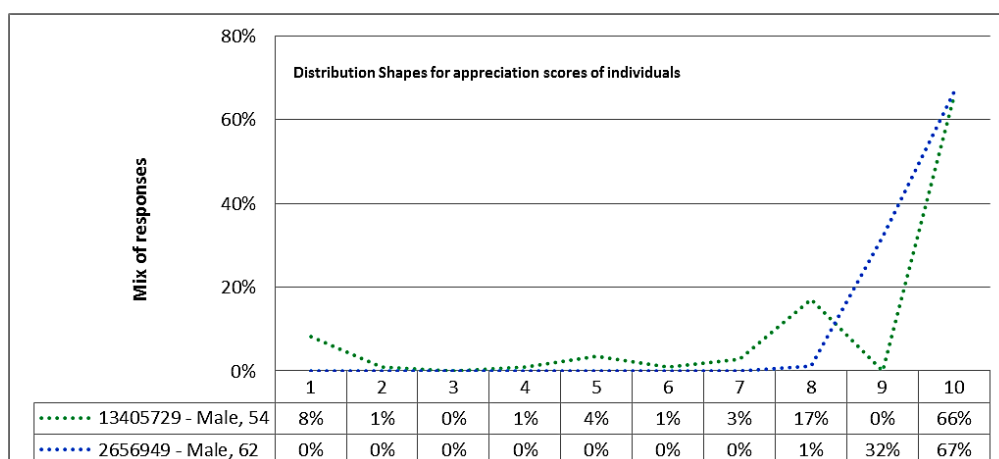
In figure 91, we can see that both respondents allocate a score of 10 for around two thirds of the radio programmes that they hear (within this dataset). Therefore a programme they both think is good will have a mean score of 10 (an AI of 100). When respondent 13405729 hears a programme he thinks is really bad (8% of the time) he uses the score of 1. This would bring the mean down to 5.5 (an AI of 55). However, should respondent 2656949 think poorly of a show, he appears only to use 8 as his lowest score and does this only 1% of the time. If he were to rate this programme 8, the mean would still be 9 (an AI of 90). Thus, in this (albeit extreme) example, a programme which has a consistent mix of people really liking or disliking it, the show could have an appreciation score ranging between 55 and 90 simply because of differences in response patterns.

In practice, is this likely to be a problem? What we don't know the extent to which the scoring is relative or absolute and, as discussed in the chapter on distribution,[335] there is no proof that attitudinal ratings are expressed in a linear fashion even using a numerical scale. We cannot know if there is much greater strength of feeling from an individual who gives a lowest score of 1 versus another whose lowest score is 8.

---

[334] See p190.

[335] See from p143.

However, while we cannot assess the sentiment behind the scores, we can show that an individual who uses the extremes of the scale will have a greater influence on the mean score than one who is more restrictive in their scoring. This could be problematic if those who score extremely tend to be certain demographic groups with specific interests. For example, if they are drawn to give low ratings for a specific type of programme, their scores could exercise undue influence on the programme aggregate. Although there is insufficient data to generalise about response patterns, the analysis did provide evidence that demographic groups have tendencies toward different response patterns.[336]

> While response patterns indicate that respondents are rating programmes on their relative goodness, at aggregate level, this may not adversely affect the result. However, respondents can have greater or lesser influence on the aggregate scores depending on their response patterns; more extreme responses can have a greater effect upon an AI.

7.3.5 Further Pulse Questions and their Relationship to Appreciation Scores

The Pulse survey includes further questions about the respondents' attitudes toward the programmes, along with questions about their listening situation, in addition to merely rating them. This section examines whether these metrics may also be useful to programmers in relation to Radio 4 comedy appreciation scores. Of the circa 42,000 Pulse responses relating to Radio 4 comedy, only circa 30,000 included responses to the further questions; this is still a considerable sample size for analysis.[337] The responses to each of these further questions are discussed below.[338]

i Further Question 1 – 'This was a high quality programme'

Respondents are asked the extent to which they agree with the statement: **This was a high quality programme**. The following response options are presented:

> **Agree strongly**
> **Agree slightly**
> **Disagree slightly**
> **Disagree strongly**
> **No strong views**

Figure 92 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – 'High Quality' compared to appreciation score

	1	2	3	4	5	6	7	8	9	10
% agree high quality	3%	5%	16%	22%	40%	53%	77%	89%	96%	98%

---
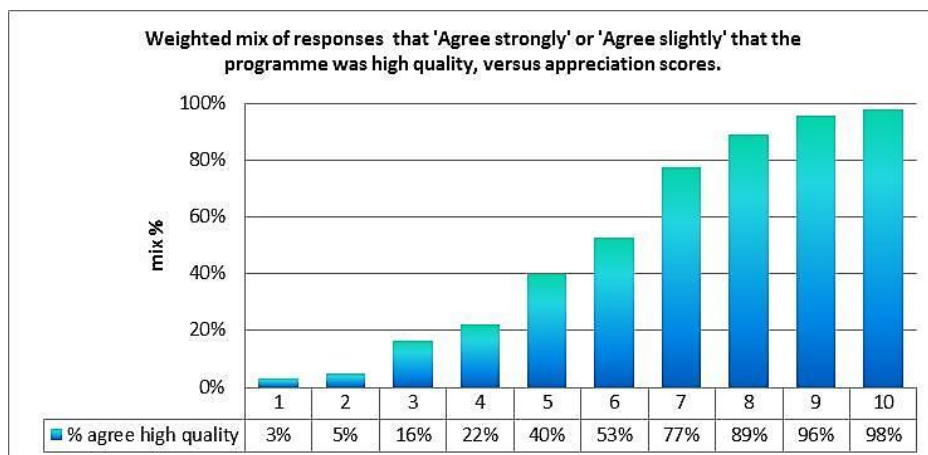
[336] See from p190.

[337] See from p85 for Pulse questions.

[338] Not all questions are asked of comedy programmes, e.g. '*This programme helped improve my understanding of politics*' is not asked.

Figure 92 shows that the mix of those who agree that a programme was high quality (total of 'agree strongly' and 'agree slightly') bears a very strong relationship with increasing appreciation scores: an extremely high correlation of 0.98.[339] In this respect, the quality question is not really giving us any further information, the properties being almost interchangeable. However, the slight flattening of the curve at the highest and lowest scores indicates that there *may* be something more complex than a simple linear relationship – as proposed by North (2010).

This is to be expected as it has been found that appreciation and quality are analogous (Davie, 2010).[340] The relationship is so strong that one could argue that there is no benefit from having the two measures; it appears that the respondents do not, or cannot, distinguish between them. This links back to Silvey's (1974, p115-116) claim that: 'most panel members found this distinction incomprehensible', and much later by Ang (1991, p145) who claimed that what is being measured is just 'general satisfaction', with attempts to break down the constituents being costly and difficult to interpret.[341] One could argue that this correlation indicates that quality may be a core constituent of appreciation.

> For Radio 4 comedy, perception of high quality has an extremely high correlation with high appreciation ratings. Respondents do not appear to differentiate between the measures and, as such, the measurement of both metrics may be unnecessary.

ii Further Question 2 – 'This programme felt original and different from most other radio programmes I've listened to'

This metric is circulated as an aggregate score in the corporate quarterly publication (for example: BBC, 2011a, p15) and is becoming increasingly widely disseminated to BBC staff. The audience research department send a weekly email to Radio 4 Editors that highlights AIs along with **original and different** scores, for example, (BBC email correspondence from BBC research, dated 06/12/2013):

> • *I'm Sorry I Haven't a Clue* received its highest AI of the series with an impressive 89 (+3) and an excellent original and different score of 70% (+10%).

> • *The Secret World* returns with its 4th series and receives an AI of 76 (+6 on last score).

> • *That Mitchell and Webb Sound* is back with its 5th series and pulled in an AI of 79 and distinctiveness score of 49%.

This commentary is alongside a chart plotting AI against this metric (ibid):

---

[339] The correlation calculation used here is Pearson's correlation rather than Spearman's correlation as Spearman delivered a 1.0 in this and most of the following 'further questions' analysis. Whilst, in theory Spearman should be used for non-parametric data, here we are looking at a relationship between the AI scores and the mix of those who agree, rather than comparing the line by line data. It is not possible to compare to the line by line data without attributing numerical values to the Likert style responses and these values would only be arbitrary.

[340] See from p127.

[341] See from p132.

Figure 93 BBC Audience Research – Quality and Distinctiveness Report – Radio Pulse 25th November 2013 – 1st December 2013 / TRP
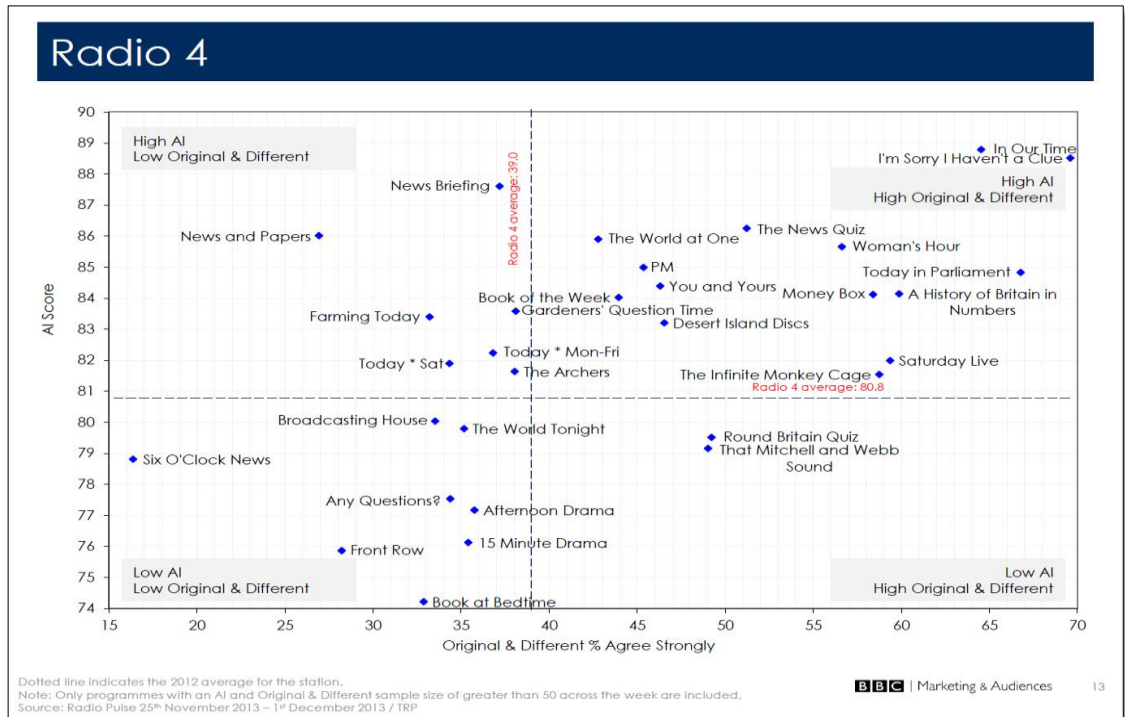
The chart above indicates that *Clue* scores the most highly for 'original and different' versus any other measured programme. The programme has been running largely unchanged in format and content for 40 years, so one wonders what this metric might really be measuring. It could merely be a comparison to other Radio 4 programmes rather than an absolute measure.

The chart below shows how the mix of 'original and different' varies with appreciation scores.

Figure 94 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – 'Original and different' compared to appreciation score

	1	2	3	4	5	6	7	8	9	10
% agree Original & Different	12%	19%	26%	35%	50%	57%	72%	84%	91%	95%
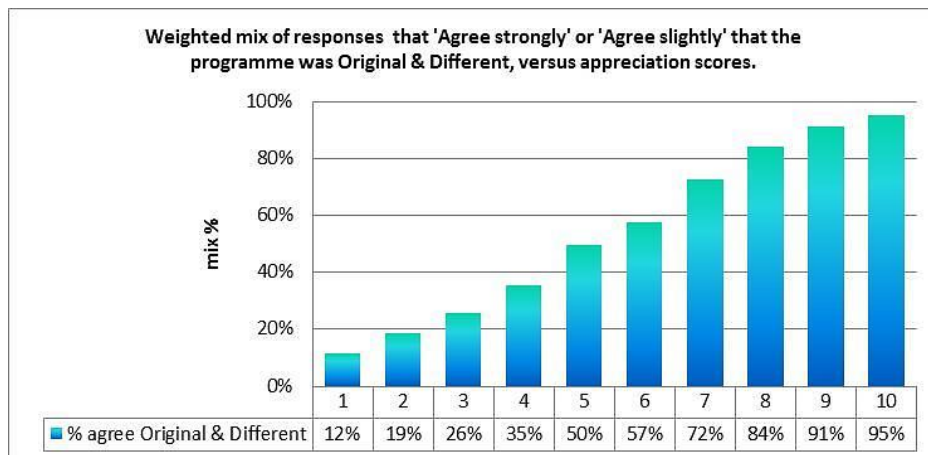
This clearly shows that as appreciation scores increase, so does the mix of respondents who agree that the programme was 'original and different'. The correlation between these measures is near perfect at 0.99. This indicates that if a respondent likes a programme, they will also think of it as original and different. Thus, like the measurement of perceived quality, this metric is not really telling us any additional information than can be seen from AIs alone. One could argue that this correlation indicates that the perception that a programme is original may be a core constituent of appreciation.

For Radio 4 comedy, perception of a programme being 'original and different' has an extremely high correlation with high appreciation ratings. Respondents do not appear to differentiate between the measures and therefore the measurement of both metrics may be unnecessary.

iii Further Question 3 – 'This programme felt fresh and new'

Respondents are also asked if they found the programme to be **fresh and new**. The chart below shows how the mix of agreement to this question varies with appreciation scores.

Figure 95 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – 'Fresh and new' compared to appreciation score
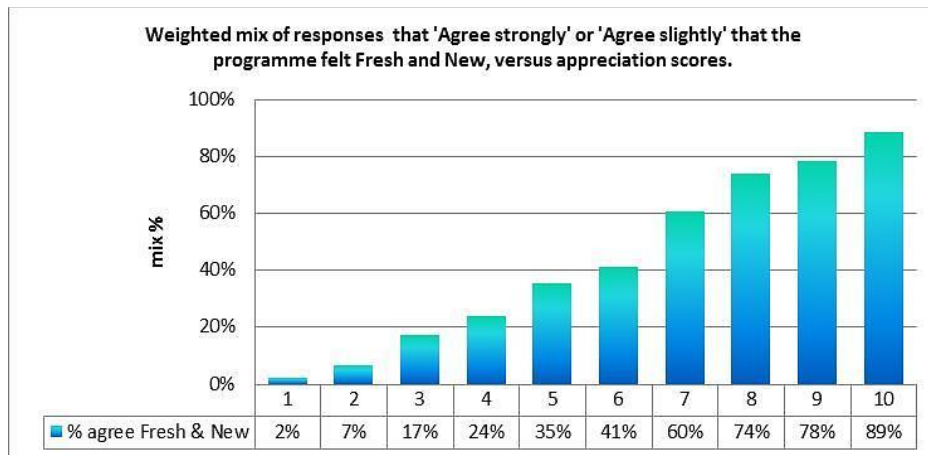
As with the previous two questions, as appreciation score increases so does the mix of agreement that the programme was fresh and new: a correlation of 0.99. Further on in the results discussion we find that appreciation increases broadly in line with the number of series that a show has.[342] This means that, perversely, the more series a show has, the more likely it is to get higher scores for 'fresh and new'.

Figure 96 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – 'Fresh and new' compared to number of series

Mix of responses that 'agree slightly' and 'agree strongly' that '*this programme felt fresh and new*'	
New (1-4 series)	62%
Established (5-10 series)	66%
Core (11 or more series)	73%

Figure 96 shows that Radio 4 comedy shows that have been running for many series are more likely to be considered 'fresh and new'. Again, the precision of the correlation between this metric and the appreciation score indicates that the respondents may merely see it as a positive thing to say about the programme rather than considering its exact meaning; 'fresh and new' could merely be a constituent part of appreciation.

For Radio 4 comedy, perception of a programme being 'fresh and new' increases along with increasing appreciation ratings. Since respondents do not appear to differentiate between the two, the measurement of both metrics may be superfluous. Series running for longer are more likely to be considered 'fresh and new'.

---

[342] See from p227.

iv Further Question 4 – 'It's the kind of programme I would talk to other people about'

A further question is: **It's the kind of programme I would talk to other people about**; **Agree strongly…** etc. This question could, arguably, be seen as most closely relating to that used in the Net Promoter Score: 'How likely is it that you would recommend company/brand x to a friend or colleague?' (Barwise and Meehan, 2011, p44-45).

Figure 97 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – 'Would talk to others' compared to appreciation score

	1	2	3	4	5	6	7	8	9	10
Agree' would talk to others	2%	9%	16%	16%	26%	26%	41%	55%	68%	81%
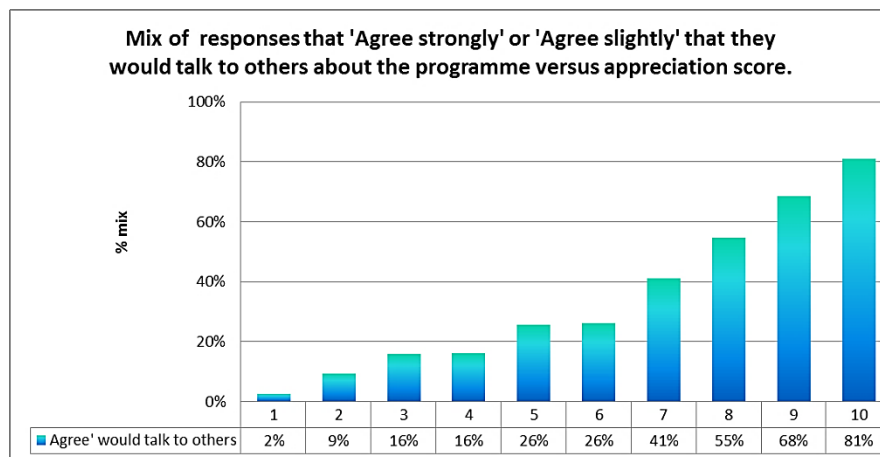
Again we find that as appreciation scores increase, the likelihood that people agree that they would talk to others about the show increases: a very high correlation of 0.94. Thus this question again elicits little more than we are already getting from appreciation scores, with respondents unable to distinguish between it and the general appreciation question.

> For Radio 4 comedy, perception of a programme as 'the kind of programme [the respondent] would talk to other people about' has an extremely high correlation with increasing appreciation ratings, suggesting the measurement of both metrics may be unnecessary.

v Further Question 5 – 'And how much effort did you make to listen to each of these programmes?'

Radio research has a requirement to understand how much attention is paid by the listener,[343] the Pulse survey attempts to establish this by asking the question: **And how much effort did you make to listen to each of these programmes?** For each show the respondent is given the following answers to choose from:

> **I made an extra special effort to listen to the programme**
> **I usually listen to this show as part of my usual routine**
> **I did not make any particular effort to listen to the show**

If a listener claims to make a special effort to listen to a programme, might their appreciation be higher?

---

[343] See from p79.

Figure 98 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – 'Effort' compared to appreciation score
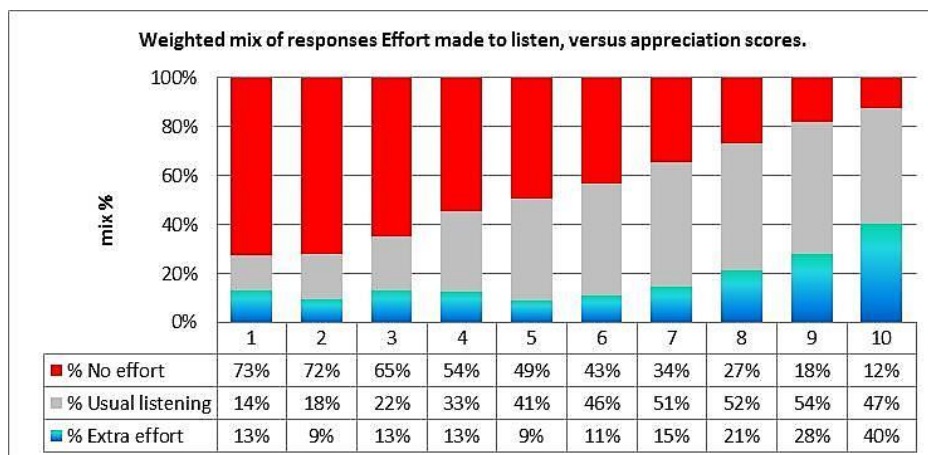
The chart above shows that there is not a perfect linear relationship between the increasing appreciation scores and the proportion of those claiming that they made 'extra effort' to listen (blue bars), but there is a positive correlation at 0.78. A stronger correlation exists between 'usual listening' and appreciation scores at 0.93 (grey bars). The data does not allow us to infer the causality behind this. Is a programme highly appreciated because it's part of usual listening, or is it part of usual listening because the respondent particularly likes it? However, the strongest correlation lies in a near perfect inverse linear relationship between appreciation and the proportion of claimed 'no effort' – a figure of -1.00 (red bars). Again, the causality is not known. Do listeners not like programmes to which they have made no effort to listen, or have they made no effort because it's something that they know that they don't like?

Claimed lack of effort and low appreciation scores have a near perfect correlation.

vi Further Question 6 – 'And how much of each programme did you listen to yesterday?'

A further question asks the respondent to quantify the amount of the programme that they heard and to estimate that on a 1-10 scale. What we don't know is how the respondent is interpreting this question. For example, if a respondent gives a score of 5, indicating that they may have 'listened' to around half of a programme, is it because they have had the show on for half the time, or because it was on for the full duration but the respondent only considered that they were concentrating for half the time, or even they were only half listening? While early radio users were commanded to listen with full concentration (see Shingler and Wieringa, 1998, px) we understand that, as a secondary or even tertiary medium (Barnard, 2000, p188), the level of attention paid can vary widely (see Mytton, 1999, p124-125).

This question asks how much was 'listened to' rather than 'heard' and as such implies a modicum of effort: 'an object of concentration' (Crisell, 1994, p15). Thus we might imagine that, as with the previous question asking respondents to indicate how much effort they made to listen, appreciation is higher when respondents are 'listening' to more of the programme.

Figure 99 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – How much of the programme 'Listened to' compared to appreciation score
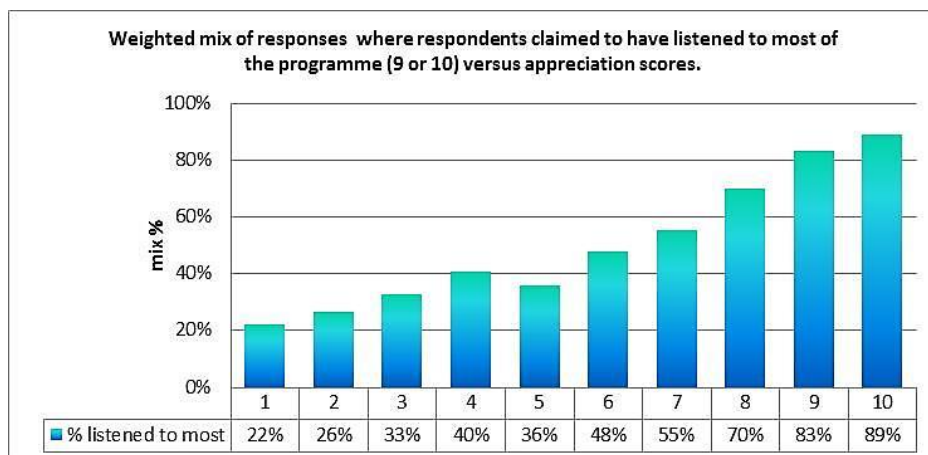
Figure 99 shows that as appreciation scores increase, the mix of those who respond with a '9' or a '10' to the 'listening' question increases. The correlation is very high at 0.97. Again, the causality is unclear: they may listen to much of the programme because they like it, or they may like it because they managed to hear most of it. Alternatively it could be that, as with earlier questions, the respondents have rated it highly as a general score and then allocated equivalent high scores to any further questions.

Respondents who claim to have heard all or nearly all of a programme also give high appreciation ratings. Asking how a programme is rated and then how much of a programme is heard gives near interchangeable results.

vii Further Question 7 – 'Was there anything in the programme that you personally found offensive?'

Our dataset includes a question regarding offence. Of the 30,473 responses to Radio 4 comedy (where the respondent went on to answer further questions in the survey in addition to just the appreciation rating) 579 (1.9%) said 'yes' (they were offended). Responses to programmes where listeners have heard something that they have considered offensive are far more likely to have lower scores, 45% being 1s and 2s compared to just 3% for those that did not cause offence (see chart below):

Figure 100 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Offence
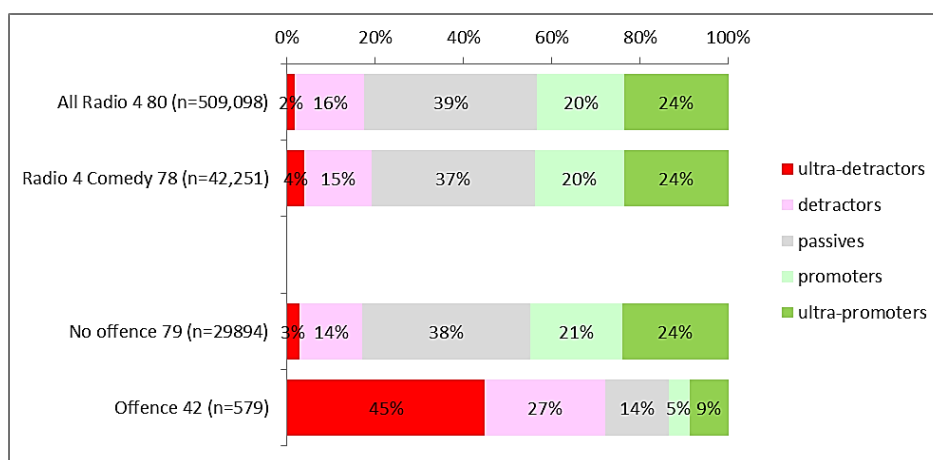
Where programmes *have* caused offence there is a huge effect on the mean score, the resulting AI being just 42 compared with 78 for all Radio 4 comedy.

Verbatim responses can illustrate the strength of feeling held when programmes are rated low and designated offensive:

That Mitchell and Webb Sound 22-Aug-12 Male 56 (respondent 613766)

'Yes' to being offensive, Appreciation score = 1
'Nothing to like. Disgusting, rude and racist! I am dissapointed [*sic*] that BBC Radio 4 can produce and broadcast such rubbish and crude programme on the air. It is offensive and you call that comedy? I fail to see the humour in its contents. Downright disrespectful and those two should shot!'

Andrew Lawrence 25-Oct-12 Male 44 (respondent 13190801)
'Yes' to being offensive, Appreciation score = 1
'Not funny, unintelligent, predictable rubbish. Not what one expects of Radio 4. Also jibes about vegetarians was very offensive to the members of our household who are vegetarian.'

The two examples above indicate that the listener has taken offence to aspects of the shows: crudeness and jokes about lifestyle choices respectively. Kahneman (2011, p302) talks of the concept whereby a single undesirable component can affect the view of the whole to a disproportionate extent.[344] It could be that the aspects the listener found offensive have coloured their whole view of the show, resulting in such a low appreciation score: '1' in both of these cases.

In the majority of examples, however, the selection of 'yes' to offensive appears to be merely an expression of 'unfunny' rather than the listener taking any particular offence:

Just a Minute 14-May-12 Male 64 (respondent 625830)
'Yes' to being offensive, Appreciation score = 1
'This programme has become weak and needs to be scrapped.'

Party 19-Sep-12 Male 54 (respondent 905171)
Yes' to being offensive, Appreciation score = 1
'That crap like this keeps coming back while ***Look Away Now*** hasn't! Apart from I'm Sorry I haven't a clue The News Quiz and the Now show there are NO decent radio comedy shows on BBC Radio 4. 4 Extra has funnier shows that are decades old! Will be making a call to BBC complaints about this rubbish being back again. YET MORE UNFUNNY RUBBISH!!!!!!! WORSE THAN A 5TH FORM REVIEW!!!! How on earth you got the audience to laugh at the pathetic unfunny lines is beyond me Perhaps they were looking at a funny film while the show was being broadcast or perhaps the floor manager had a machine gun! At least it is only 4 shows!'

Just a Minute 06-Aug-12 Male 60 (respondent 1054632)
Yes' to being offensive, Appreciation score = 1
'Just when you think it's safe to leave the radio on, this nauseating drivel lurches out, again! Parsons, especially has me reaching for the off button, and the sick bag if I'm not quick enough. This, for me represents the worst of Radio 4, long, long past its sell by date, smug, self-satisfied and not only nothing like as funny as it thinks it is but not funny at all.'

These examples relate to the theory discussed in the chapter on distribution regarding polarised responses; if comedy isn't funny, it can make people angry.[345] In these examples, the 'offence' response appears to be an expression of that anger.

Furthermore, some respondents mark programmes as offensive to make a point. For example, respondent 905171 (Male, 54) gave 84 responses to Radio 4 comedies where he went on to answer the additional Pulse questions and of these he claimed that 35 shows that he had heard were offensive. Of these 35, he used a number of the verbatim responses to lobby specifically for one programme to come back. Examples:

---

[344] See from p111.

[345] See p48.

Party   19-Sep-12   Male 54 (respondent 905171)
'Yes' to being offensive, Appreciation score = 1
'That crap like this keeps coming back while *Look Away Now* hasnt! [*sic*]'

Cabin Pressure   22-May-12   Male 53 (respondent 905171)
Yes' to being offensive, Appreciation score = 1
'That some bell-end comissioned [*sic*] it in the first place was bad enough but to make a second series when *Look Away Now* is waiting to return is a waste of licence payers money and time! GET RID OF IT NOW!'

Andrew Lawrence   25-Oct-12   Male 54 (respondent 905171)
Yes' to being offensive, Appreciation score = 1
'Why is the BBC commissioning new rubbish instead of good old tried and tested programmes like *Look Away Now*?'

What may be surprising is that for the 'offensive' programmes, there are still a proportion of high scores allocated to them; ratings of 10 still account for 9% of those given. Verbatim responses can illustrate the contradiction:

Clue   02-Jul-12   Male 56 (respondent 613766)

'Yes' to being offensive, Appreciation score = 10
'Hillarious [*sic*] a real treat for the most part. Jack Dee has to spoil things and go too far with his ceiling innuendo.'

The News Quiz   21-Dec-12   Female 49 (respondent 13263101)
'Yes' to being offensive, Appreciation score = 10
'Very funny. Some of the humour was more sexually explicit than usual which I don't like at the time of day I was listening especially as children might be listening.'

In these examples the listeners had found something that they didn't like within the programmes, but the rating of 10 indicates that it has not affected their overall enjoyment. This is in opposition to the aforementioned theory whereby a single aspect can ruin a whole.

> While a claim of 'offence' is more likely to connect with a lower appreciation score, it is not always the case. However, there is evidence to suggest that respondents can use the offence question to indicate dissatisfaction rather than actual offence. The metric clearly has a variety of meanings for respondents and may be difficult to interpret in relation to appreciation.

viii Further Question 8 – 'Where did you spend the most time listening to each of these programmes?'

This question seeks to identify where the respondent was when they listened to the show. The following options are given:

> **01 At home**
> **02 At work**
> **03 In the car**
> **04 Elsewhere**

The RAJAR survey also asks a similar question, but the results are not reported on as standard by either the RAJAR site (RAJAR.co.uk) or the BBC quarterly summary reports. The chart below shows how appreciation varies with each of these situations:

While it can be seen above that the locations do vary in terms of appreciation, this data for Radio 4 comedies relates, of course, just to the timeslots where they are broadcast. 54% of the responses come from 18:30 comedies. 'At home' is by far the greatest proportion of listening with 79% of the responses. Thus, while there are observable differences from alternative locations, their effect on the overall AI is relatively small.

'At work' gives comparatively more detractors (3-6s) but fewer ultra-detractors (1s and 2s). The resultant AI score is 5 points lower than for those listening at home. Unfortunately we have no information about whether these listeners are alone or with others, or on the nature of their work, so there's little that can be surmised from this information. The mere fact that the appreciation is so much lower is of interest and could deserve further investigation. Is this just the case for comedy? The chart below shows information on the location of listening for all Radio 4 genres excluding comedy:

Figure 102 LEGITIMATE WEIGHTED RESPONSES Radio 4 shows EXCLUDING Comedy – Gromic – Where programmes were 'listened to'
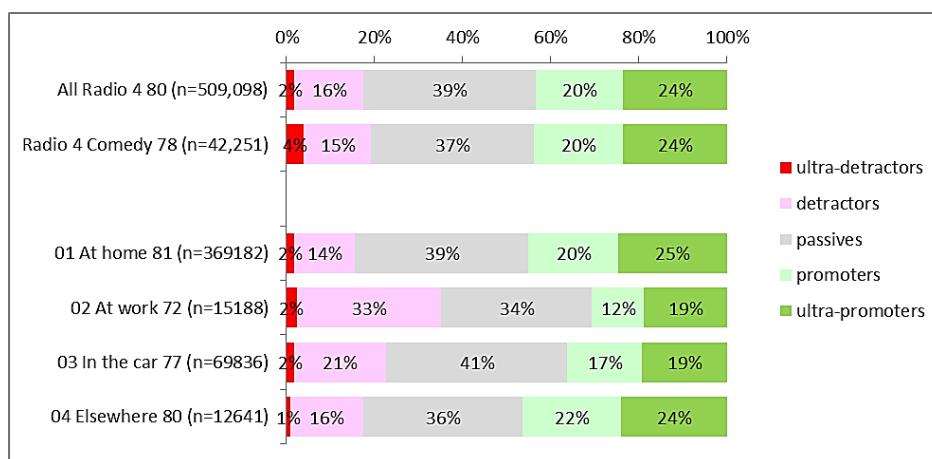
Here the difference is even more marked. Listening at work has an AI of 72 compared to 81 for those at home. However, might this have something to do with the mix of the programme genres? As it has been found that some genres attract higher appreciation by their very nature (see Carrie, 1997, p26), it could be that the type of programmes heard when people are at work may not be the type that would be most appreciated there. For example, an afternoon drama (14:15 – 15:00) could demand attention to get full appreciation, which might not be possible at work. For comedy, we can eliminate this effect by looking at just one timeslot that contains programmes limited to one sub-genre. Monday 18:30 attracts the highest

responses and consists just of the originations of comedy panel shows:[346] **Clue, JAM**, **The Unbelievable Truth** and **Museum of Curiosity.**

Considering these programmes, we can still see that when the programmes are heard at work, appreciation is lower overall; the AI is 81 versus 84 at home. However, this 3-point difference is smaller than the 5 points that we saw when looking at all comedy programmes and the 9 point difference in the non-comedy programmes. Thus it appears that listening to Radio 4 while working can have a detrimental effect on appreciation, but the extent of this effect is variable. The lessened reduction in the appreciation for these panel shows could be related to their 'dippability'. As discussed previously,[347] we might expect that programmes that are dippable – i.e., require a low level of attention or effort – could have a relatively higher appreciation in a situation where full attention is not possible. For example, if one is at work, a programme that can be dipped in and out of could be better background listening than programmes that demand full attention.

Another option for listening location is 'in the car'. Again, in all the charts above in this section, appreciation is lower than listening at home to some extent. This could be related to 'effort'. Earlier, it was seen that lack of effort correlated precisely with lower appreciation scores (p208). We compare the proportion of those who claim to have made no effort to listen in the home versus in the car:

The proportion of those who responded, 'I did not make any particular effort to listen to the show':

> Those listening at home = 66%
> Those listening in the car = 92%

Thus, respondents listening in the car were more likely to have just switched on the radio and heard the programme rather than having made a conscious decision to listen to it, and, apparently in consequence, have tended to give a lower appreciation score.

Another factor in play may be the actual audio. Reynolds (2011), albeit discussing radio drama rather than comedy, pointed out that producers do not consider the subtleties of audio for those listening in a car. If this is the case, respondents may not be able to pick-up on every single acoustic nuance of a programme that could otherwise be accessible to someone listening at home or on headphones. This might encourage lower appreciation.

---

[346] I.e., there are no repeat broadcasts.

[347] See from p102.

The appreciation for 'listening elsewhere' in the charts above can be higher or lower than listening at home. As we have no specifics for the nature of 'elsewhere', conclusions cannot be drawn from this.

There is reason to believe that the location of listening to Radio 4 comedy can affect appreciation. However, the extent of the influence is not consistent. The varying claimed effort required to listen to the programmes could be a factor.

ix Further Questions 9 –Verbatims

Word count as an indicator of engagement

Respondents are also given the chance to give open-ended responses to any programme they claim to have heard:

- **What did you think about this radio programme? Please write in what, if anything, you LIKED about it?**

- **And what, if anything, did you DISLIKE about it?**

Of the circa 42,000 Radio 4 comedy responses, circa 18,000 (43%) had some kind of verbatim response, ranging from just one word up to 142 words.

It was discussed earlier how the word count of the response might relate to the engagement of the respondent.[348] Could this proposed phenomenon be applicable to Radio 4 comedy verbatims?

Figure 104 LEGITIMATE UNWEIGHTED RESPONSES Radio 4 comedy programmes with further Pulse question responses – Verbatim word count – mean count compared to appreciation score
X axis = appreciation score, y axis = mean wordcount.
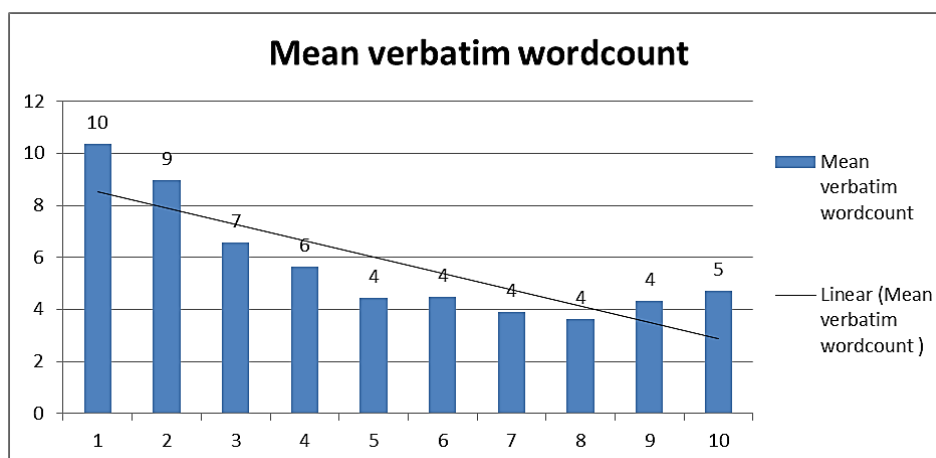
The chart above shows that Radio 4 comedies attract the largest word count if they have a lower appreciation. The relationship is not linear as there is also a peak at the other end, for scores of 9 or 10, giving a U-shaped distribution. This supports the claim of Collins (2010a): '*if something gets a really low score or a really high score, you get more detail about it*'. Thus, for Radio 4 comedy, if a listener has an extreme reaction to a show then, they are driven to give a written response, particularly if that reaction is negative. However, is this the case only for comedy?

---

[348] See from p132.

X axis = appreciation score, y axis = mean wordcount.

Non-comedy programmes also attract a higher verbatim word count for the lower appreciation scores with another peak at the highest ratings. Overall, the number of words written is slightly fewer than for comedy. The mean for comedy is 4.5 while that for non-comedy is 3.7. However, this is more marked when we just look at the lowest scores. For appreciation scores of 1s and 2s, the mean word count for comedy is 9.8 while that for non-comedy is 6.6. This supports the theory discussed earlier,[349] wherein if comedy is found not to be funny it does not just result in poor appreciation but actually incites anger. The verbatim responses allow the respondent to express the nature of that anger:

Rudy's Rare Records 09-Oct-12   Male 68 (respondent 641193)
Appreciation score = 1
'This was execrable. He needs to stick to gurning on his cookery show. The whole farago [*sic*] was offensive.'

Nick Mohammed in Bits 02-Dec-12   Female 56 (respondent 13498256)
Appreciation score = 1
'Utter puerile rubbish! I found nothing amusing about this – surely the audience laughter must have been "canned" or maybe it "looked" funnier than it sounded… As I said, it was utter puerile rubbish! I rely on Radio 4 to keep me entertained whilst I cook in the evenings as I refuse to have television in the kitchen. Sunday Dinner takes a lot of cooking so I am going to have to find an alternative radio station for the next 3 Sundays, or set up I-Player.'

The Simon Day Show 13-Dec-12   Male 61 (respondent 13513743)
Appreciation score = 2
'Is meant to be funny? A complete waste of time. If the BBC think this is worth spending our licence fee on then it should be shunted to one of the myriad of digital radio channels that nobody listens to.'

For Radio 4 the lowest appreciation scores are more likely to draw longer verbatim responses: particularly so for comedy.

---

[349] See from p48.

<u>Word use as an indicator of appreciation</u>

We have discussed how the 1-10 appreciation rating scale might give the *illusion* of a linear scale, but North (2010) indicated that research might indicate a more complex relationship.[350] Could word use in the verbatim responses begin to tell us about this relationship?

<u>Figure 106 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes – Likelihood of verbatim response including the word 'love' (total number of responses with verbatims =18013 – those containing 'love' = 722)</u>[351]
X axis = appreciation score, y axis = percentage of responses containing the word 'love'.
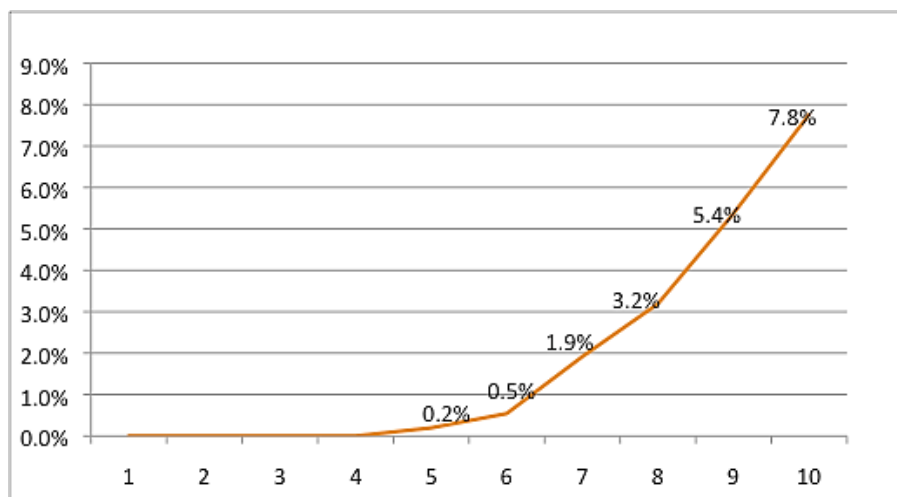
Figure 106 shows how the word 'love' and its derivatives are more likely to be used in verbatim responses when the appreciation scores are high. It shows a clear progression that may reflect strength of feeling in respect to numerical appreciation scores,[352] however it is not a simple, linear relationship.

This investigation may have been crude as 'love' might not necessarily be the very best word to use. Perhaps a combination of words might be used or, indeed, some other, more complex, 'sentiment analysis'. Further work could be of interest here as it could indicate the nature of the relationship between the numerical scale points and strength of feeling, and possibly form the beginnings of an attempt to give true 'validity' to appreciation scores. Relating back to McConway's comment:[353]

> *They have some internal transformation between that scale and the 10-points... What we don't know is if it's a linear scale.* McConway, 2013.

Thus, while we cannot say that appreciation is measured on anything more than an ordinal scale, further work in this area might be able to take it into an interval or even ratio scale. With the example above, could we conclude that a rating of 9 indicates 'love' around three times as much as a rating of 7 simply because the word 'love' is used around three times more often (5.4% of the time compared to 1.9% of the time). If we could verify the exact relationships between the scale points, it would then be possible to submit responses to parametric analysis with a confidence that is not currently present.

Further analysis of word use from verbatim responses may allow us to infer the true strength of feeling represented by points on a linear scale. I.e., it could lead to the quantification of the distances between scale points rather than applying an arbitrary equidistance. This *could* begin to approach the use of the responses in a statistically robust fashion.

---

[350] See from p147.

[351] All incidences where appreciation score was 7 or lower, the use of the word love was checked in case of it being used in the negative, e.g., 'I didn't love this'. Also, the incidences include derivatives of the word love, such as 'lovely'. Negative uses are not included in the figures.

[352] The data was also analysed to look into whether the word 'hate' might show the inverse effect, but the sample size was too small to be of use.

[353] See from p149.

7.3.6 Audience Size – Estimate From Number of Pulse Responses

Lack of data from RAJAR at programme-level means that we do not have visibility of how many people listen to specific episodes of a show, nor can we compare audience sizes between different series (data that would be of use to programmers).[354] However, we have discussed how evidence has been found that, for television, there is a correlation between the number of Pulse responses and the audience size[355] and there is no reason why this might not be the same for radio responses. Using the formulae proposed earlier (see from p93), Pulse data can be compared to general RAJAR figures to see if this is the case for Radio 4 comedy. To follow is an example using actual RAJAR and appreciation data.

In the RAJAR figures for Q4 2012, the average audience for the 18:30-19:00 weekday slot is 1,161,000 (L). That is, for the period running 17th Sept -16th Dec 2012, Monday to Friday, this is typically the number of people who were listening for at least 5 minutes according to their diary entries. The Pulse data tells us that across this same period for this time slot, there were 4,458 appreciation responses (n). As these responses are across the whole period, we need to divide the total by 65 (w) to get the average response each day as there were 65 weekdays during this time. This means that each episode received a mean of 69 responses (e). Therefore, we have 69 responses typically representing an audience of 1,161,000. This indicates that each response is representative of nearly 17,000 listeners (r).

$$n \div w = e \qquad\qquad L \div e = r$$
$$4{,}458 \div 65 = 69 \qquad\qquad 1{,}161{,}000 \div 69 = \underline{16{,}829}$$

Thus, to work out an estimate of how many listeners any show might get, all we would need to do is take the number of responses for a programme and multiply it by around 17,000. We can check this figure by validating it against another example. In theory, the number of responses is always proportionate to the number of listeners as measured by RAJAR. Applying this method to, for example, *Front Row,*[356] which also runs for half an hour in the early evening on Radio 4, Monday to Friday, over the same time period, it can be seen:

$$n \div w = e \qquad\qquad L \div e = r$$
$$1{,}174 \div 65 = 18 \qquad\qquad 680{,}000 \div 18 = \underline{37{,}777}$$

In this example, each listener response represents nearly 38,000 listeners: well over twice that found for 18:30 comedy. This lack of consistency means that, in fact, we are *not* able to use the number of responses as an absolute indicator of audience size. While we would expect the responses to be proportionate to the audience size, there could be a number of factors that affect the validity, such as:

- The sample sizes per episode are very small for *Front Row* (18) and can even be low for the 18:30 comedies (69). Figures of this size might be too small to give reasonable estimates.

- Response rate per episode could vary depending on the type of programme. In our example, *Front Row* is a long-running show that has a fairly standard format and style of content. Perhaps this means that the respondents are less likely to claim that they have heard it, either because they don't remember it, or they have nothing to say about it. This may relate to our finding that comedy attracts longer verbatim responses than other types of programme[357] - maybe comedy prompts people to respond, in one form or another, more than other genres. Thus, if a programme has fewer responses,

---

354 See from p81.

355 See from p93.

356 Front Row is a BBC Radio 4 arts programme that TXs 19:15-19:45 Monday to Friday throughout all of the measured RAJAR period.

357 See from p214.

we are unable to say whether this is because it has fewer listeners or just because it has a lower response rate.

- Response rates could also be affected by any number of variables such as the demographics of the respondents or even the TX slot.

Although, in the example comparing 18:30 comedies and *Front Row,* we cannot prove a relationship between audience size and number of responses across different genres, we have reason to believe a relationship is there as found by (Van Meurs, 2008). It is not contentious to theorise that there may be consistency *within* a certain genre or sub-genre and for a specific time slot, meaning that the more people who heard a Radio 4 comedy, the more Pulse responses it would be likely to receive - as theorised by Carrie (1997, p3).[358] The chart below shows all the responses for Radio 4 shows that TXed on a Monday 18:30-19:00 throughout 2012.

Radio 4 - Monday 1830 TXs during 2012	number of episodes in 2012	AI	mix of '10s'	total no. of appreciation responses	mean no. of responses per episode
I'm Sorry I Haven't a Clue	12	87	36%	1934	161
Just a Minute	22	83	30%	2457	112
The Unbelievable Truth	12	82	21%	1312	109
The Museum of Curiosity	6	76	17%	508	85

As each of these programmes was broadcast in the same slot, we can assume that, in general, the audience available to listen was fairly static.[359] However, we can see that the number of respondents to each series varies significantly to the extent that *Clue* typically gets almost twice as many Pulse responses as *The Museum Of Curiosity* (161 versus 85). This could indicate that *Clue* has twice as many listeners and therefore gets twice as many responses. We can see that a greater proportion of respondents rate the programme as a '10' (and the AI increases). The response rate is also relatively higher. If we look at the correlation, we find that the mix of 10s and the mean number of responses per episode has a high correlation of 0.91. This means that there is a definite relationship between appreciation and response level and, by inference, appreciation and audience size, within Monday 18:30 comedies.

Researchers such as Menneer (1987) and Carrie (1997) have found that there is a relationship between higher appreciation and higher audience size, particularly for entertainment shows, once programme types, audience make-up and scheduling differences had been accounted for.[360] Our figures account for these differences as they are all panel shows sitting at the same time of the day, with the same available audience. This high appreciation reinforces the notion that *Clue* is getting a bigger audience than *The Museum Of Curiosity*.

Returning to the *Front Row* example it seems that, while we cannot use the number of responses to indicate an audience size when comparing different genres, there is evidence to suggest that we could compare programmes of a similar genre that are TXing in a specific slot (as in the chart above).

Without weekly RAJAR figures we don't have audience size information at episode level, but we do have access to online listening figures that relate back to programme-level information. We can add these into this analysis:

---

[358] See from p91.

[359] See from p84 – scheduling is a major contributor to audience size.

[360] See from p91.

Radio 4 - Monday 1830 TXs during 2012	number of episodes in 2012	AI	mix of '10s'	total no. of appreciation responses	mean no. of responses per episode	total no. of online listens k	mean no. of online listens per episode k
I'm Sorry I Haven't a Clue	12	87	36%	1934	161	2063	172
Just a Minute	22	83	30%	2457	112	2222	101
The Unbelievable Truth	12	82	21%	1312	109	1829	152
The Museum of Curiosity	6	76	17%	508	85	354	59

The chart above shows in the final column the mean number of online listeners per episode for each of the Monday 18:30 comedies.[362] Unlike the number of responses, online listening does not necessarily increase with a higher appreciation. **JAM**, for example, has a far higher mix of scores of '10' than **The Unbelievable Truth** (30% as opposed to 21%) but it has a lower level of online listening (101,000 per episode as opposed to 152,000 per episode respectively). Appreciation is therefore not proven as a predictor of online listening.

> While there is evidence that the number of responses is not an absolute predictor of audience size, there is reason to believe that this figure can give an indication of relative audience size for comparable programmes, i.e., if they are of the same genre and in the same slot. This initial analysis also shows that appreciation may be a predictor of linear audience size.

7.4 Summary of Findings Related to Research Question 1

Identifying individuals who gave illegitimate scores also prompted the observation of varying response patterns at respondent level. While the aggregate distribution was bimodal, the respondents adopted a variety of response patterns. We might have expected respondents to score in a unimodal hump-backed distribution but only 51% of respondents analysed were found to have distributed their ratings in this fashion. Of that 51%, there was no consistency to the skew, central tendency or steepness of the curves. The remainder used a variety of response patterns. These responses, when aggregated to programme level, also resulted in little consistency in distribution shape. This analysis suggests that the mean AI scores may not be a statistically legitimate summary of the data. Presenting the data in a format that allows the users to clearly see the spread of the data may lead to improved decision-making as it would allow an improved understanding of divisive programming – important, and even desirable, for comedy in particular.[363]

The analysis indicated that different response patterns resulted in varying mean scores. This indicates that if, as is suggested by the data, some response patterns are more likely to be adopted by certain demographic groups, skews in listening by these groups could have a disproportionate influence on the resulting AIs. These response type group skews relate to standard industry findings where respondents who are older, of a lower socio-economic status or female, are more likely to give higher appreciation scores.

While using a mean score for the AI is not statistically representative of the data it purports to represent, when compared to the more robust Net Promoter Score it appears to be a good proxy measure. We *can* say, very generally, that a programme with a higher AI or NPS is more 'appreciated' than a programme with a lower AI (if sample sizes are appropriate). However, because using this single measure will obscure lower-level detail, neither metric should be reported without referencing the spread of the data.

---

[361] Source – http://analytics.reporting.omg.bbc.co.uk/analytics. Online listening = the sum of simulcasts and catch-up.

[362] What we must bear in mind is that the online listening could relate to the Monday 18:30 TX or the Sunday 12:00 repeats. All of the programmes in this chart have the repeat so any effect from this would be seen for all of the series.

[363] See from p48.

Agreement with many of the additional Pulse questions asked of the respondents correlates extremely strongly (one of which is a negative correlation) with appreciation:

Figure 109 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy programmes – Further Pulse questions (Agree Strongly or Agree Slightly) correlation with appreciation – summary

0.98   *This was a high quality programme*
0.99   *This programme felt original and different from most other radio programmes I've listened to.*
0.99   *This programme felt fresh and new.*
0.94   *It's the kind of programme I would talk to other people about.*
-1.00   *And how much effort did you make to listen to each of these programmes?*
0.97   *And how much of each programme did you listen to yesterday?*

This means that if a programme has a high appreciation rating, we can also conclude that it is perceived as: high quality; original and different; fresh and new; more likely to be a programme that respondents would claim to speak to others about; that they made more effort to listen to; and that they listened to more of. This supports Silvey (1974, p115-116), Ang (1991, p145) and Johnson (1992) who claim that appreciation ratings are just a general measure of satisfaction, that respondents themselves find hard to distinguish between.[364] Arguably, it could be theorised these findings indicate that these elements all contribute to the overall appreciation of a programme.

As the correlations are so high (and if it were the case across all of BBC radio and television) one could reason that there is redundancy across these questions and perhaps only the one appreciation question is needed. Having fewer questions on the survey could improve completion rates (Best & Krueger, 2008, p223). A further argument for eliminating these duplicating questions is that programmers do not generally have access to the results of them over and above the AI and the sometimes 'original and different' question. If they are used at all, it is by the research teams but without dissemination to programming teams therefore having no usefulness to programming decisions.

The Pulse questions that do seem to elicit additional information about the response to the programmes are those relating to offence, listening location and verbatim responses: the latter in particular, providing richer understanding of the programmes' perceived quality as there is evidence that word count correlates with engagement. The potential of verbatims may be further unlocked with linguistic analysis, as it could help us to understand the nature of feelings in relation to a numerical scale. The implications of this could be far wider than just the BBC's survey.[365]

As RAJAR is unable to provide programme- or episode-level detail about audience size, the exact behaviour of the listeners is unquantifiable via this route. This is unfortunate, as measurement of behaviour (albeit with the caveat that RAJAR diaries are still just claimed listening), rather than the purely subjective measurement of appreciation, could arguably be a better measure of a programme. *If* it is the case that people listen to what they like, more people listening means it's more liked. Within a specific slot, it appears that we may be able to ascertain relative audience sizes. The number of appreciation responses appears to reflect the number of listeners.

---

[364] See from p132.
[365] See from p262.

7.5 Answer to Research Question 1[366]

Presenting programme appreciation *merely* as a mean score is not a statistically valid representation of the data, as response patterns are not consistent and are not necessarily unimodal or hump-backed. In particular it provides a poor understanding of relative comedy programme performances. While the mean correlates with the arguable more robust NPS which could offer greater sensitivity for summary scores, both metrics still obscure useful information about the spread of the data and can be misleading.

Given the reliance on AI data in radio programme commissioning (see from p88), mean AI's are likely to lead to sub-optimal decision-making. The analysis has shown how AI data can be disaggregated to generate information about the distribution of the scores which will be more useful in commissioning and scheduling decisions, and may also be useful in production decisions. The number of responses may be able to provide information about the behaviour of the respondents as well as about their subjective views. Verbatim responses may, in particular, have the potential to offer greater depth of understanding in this area.

---

[366] See p160.

*And our AIs on **Steptoe** and **Hancock** were very high. But, it doesn't matter how high it was, there is always 5% of the population that hated it. It doesn't matter how successful the show is. You can have an AI of 84, which is astronomical, but there'll always be 5% that didn't like it, and we found out this is true of everything.*

(Simpson – Galton and Simpson, 2012)

Chapter 8 deals with the second research question, which asks whether comedy, generally thought of as a polarising genre, is particularly misrepresented by the mean as a measure of performance. Here it is found that, although it is indeed a particularly polarising in the sense that the most loved and hated programmes are comedies, religion appears to be the most divisive when taken as an aggregate.

RQ2: Is the distribution of appreciation scores for Radio 4 comedy particularly polarised, and therefore an extreme case of the problems with using AI scores as a measure of performance?

8.1 Is Comedy Polarising?

Chapter 2 concluded that comedy is a good test of the value of mean AI scores because of a common view within the literature that comedy can provoke extreme audience responses compared to other genres.[367]

If we compare different genres on Radio 4, we can see if comedy is indeed exceptional.

Figure 110 LEGITIMATE WEIGHTED RESPONSES Radio 4 shows – Gromic – Genres[368]
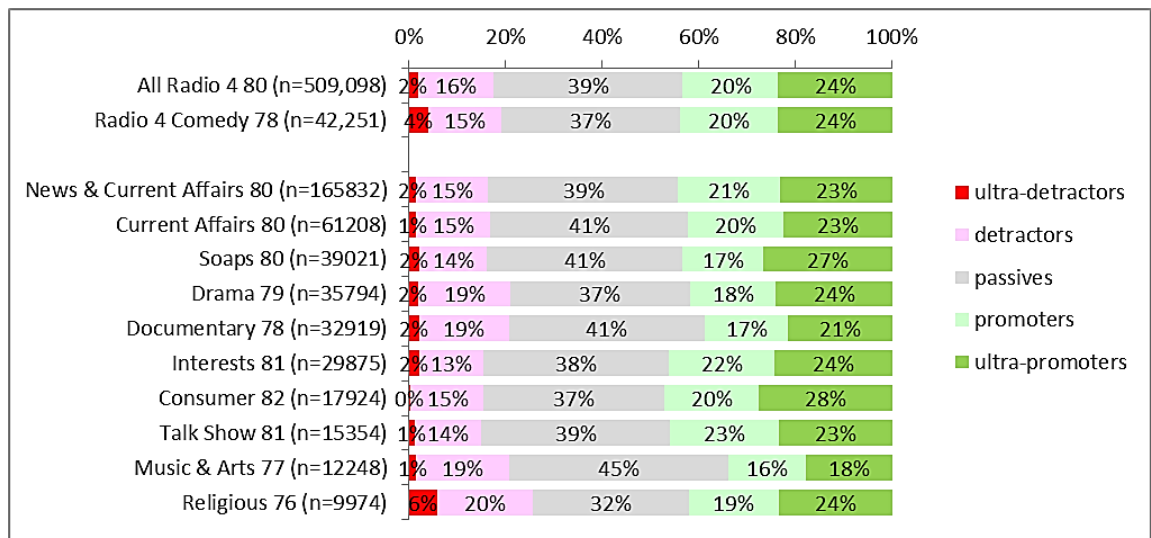
Looking at the figure above, 'Radio 4 Comedy' has a higher proportion of ultra-detractors, i.e., ratings of 1s and 2s, than Radio 4 as a whole: 4% versus 2%. However, we can see that it is actually *religion* that attracts not only the highest number of ultra-detractors, with 6% of all religious programmes receiving a score of 1 or 2, but also the highest number of detractors, with a further 20% of its scores being between 3 and 6. This is a higher proportion than that seen for comedy shows, so we might conclude that religion is in fact more polarising than comedy, something that can be illustrated when we look at the verbatim responses for religious programmes:

---

[367] See from p48.

[368] The genres of all non-comedy programmes are taken from the source Pulse data. These genres are the top ten in terms of responses and account for 90% of the responses excluding comedy.

Prayer for the Day,   05-Mar-12   Female, 60 (respondent 952584)
Appreciation score = 1
'Religion should be banned from BBC'

Prayer for the Day,   21-Mar-12   Male, 64 (respondent 1189099)
Appreciation score = 1
'Totally irrelevant and out of place on the radio. Should be shuffled off to a really early morning slot for the braindead who believe in religion and prayers – I actually find it offensive'

Sunday Worship,   08-Jul-12   Male, 52 (respondent 1057421)
Appreciation score = 1
'People who want to hear this load of cobblers should go to Church.'

The low scores allocated by some respondents appear to be due to the genre itself rather than the actual quality of the rated programme. *Prayer for the Day* arguably attracts the most vitriol as it is broadcast within the *Today* programme and can be unavoidable for those who are tuning in just to hear news coverage, whereas something like *Sunday Worship* can be more easily circumvented.

However, even though religion as a genre attracts the most ultra-detractors at an aggregate level, looking to specific series, we find that the most disliked shows on Radio 4 are, in fact, comedies:

Figure 111 LEGITIMATE WEIGHTED RESPONSES Radio 4 shows – Gromic – R4 programmes with highest ultra-detractors (n>99)
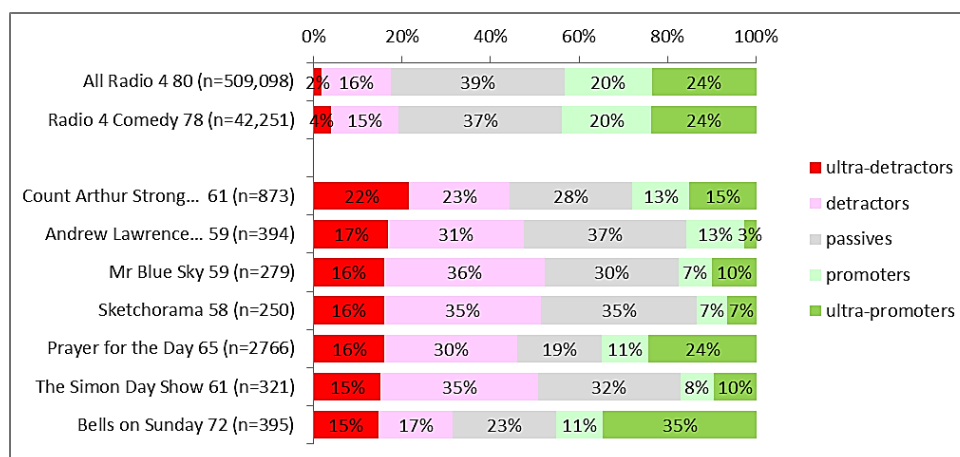
If we look at the Radio 4 programmes with the highest number of ultra-detractors, the top 4 and the 6th are all comedies. The 5th and 7th are religious. Also, whereas *Prayer for the Day* and *Bells on Sunday* both have a large number of ultra-detractors, there is a large proportion of scores of 10 (ultra-promoters) for these programmes: far higher than for the comedies.

Whereas detrimental responses for religion often appear to be based on the genre, for comedy they seem to be very programme specific:

Count Arthur   20-Mar-12   Male, 71 (respondent, 718487)
Appreciation score = 1
'It continues to be the worst alleged comedy programme on Radio 4 and is totally unworthy of a quality station – how anyone can pay for this rubbish and suggest that it is comedy is totally beyond me.'

Count Arthur   27-Mar-12   Male, 43 (respondent, 10677121)
Appreciation score = 1
'Unamusing, badly written, nasty exploitation of a stereotype, totally lacking in humour. Was asked by someone to give this programme a second chance – it continues to be utter drivel. The cliche "what a waste of licence payers' money" is particularly apt.'

Count Arthur   07-Nov-12   Female, 62 (respondent, 606654)
Appreciation score = 1
'Truly awful. Dreadful, banal and stupid.'

This evidence that programmes receive many poor responses from any one individual goes against the theory of purchaser bias – that people tend to consume what they like.[369] For example, there were 12 respondents who gave **Count Arthur** a rating of 1 or 2 for 4 episodes or more; it seemed that they disliked the show but continued to listen to it. Respondent 2580571 (Male, 65) actually rated 9 episodes of the show, every time giving it a score of 1 and often adding the comment 'was it supposed to be funny?' There is, of course, the chance that such people do not have control over the content to which they are listening, but that data is not available. Another explanation is inertia; heavy listeners may just leave Radio 4 on for long periods of time, not making the effort to turn off or over if there are any programmes they dislike. '*I get the impression, by and large, a fixed number of people will listen to comedy at six thirty and they either like it or don't*' (Schlesinger, 2011). This relates to Menneer's (2014) theory[370] whereby it may be the case that it can be difficult for listeners to predict what comedy they might enjoy, thus, there may be incidences where the comedy heard is not ideal for the listener.

So, it has been shown that comedies can solicit the most negative responses, but what of the most positive ones? The genre summary chart figure on p222 shows that 'Consumer' programmes attract the highest proportion of scores of 10 (28% of the responses). Comedy achieves just 24%, no higher than is typical for all of Radio 4. However, again, when we look at programme-level, the detail tells a different story. The following chart shows the programmes that received the highest proportion of 10s, i.e., had the most ultra-promoters:

Figure 112 LEGITIMATE WEIGHTED RESPONSES Radio 4 shows – Gromic – R4 programmes with highest ultra-promoters (n>99)[371]
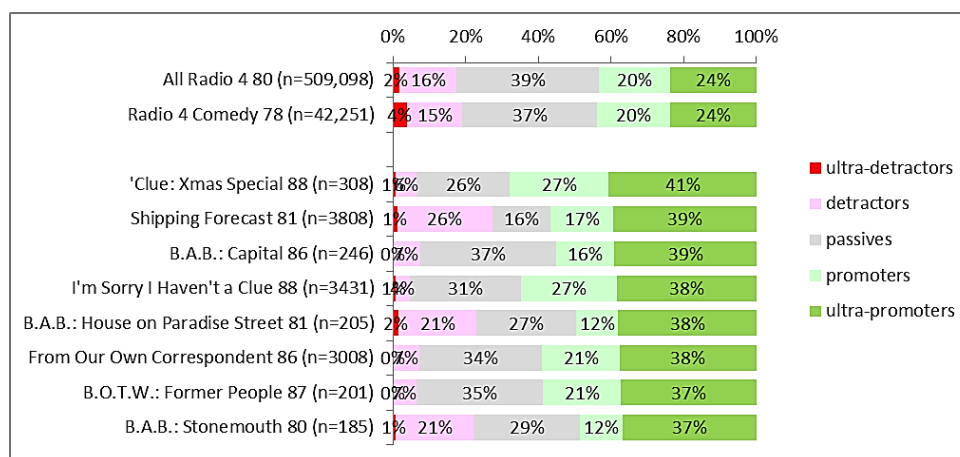
We can see that the most 'loved' programme was comedy; the *I'm Sorry I Haven't A Clue Christmas Special* garnered 10s from 41% of the respondents who rated it. This finding is in line with a theory from BBC audience research (2009b) claiming that special one-off programmes tend to get higher appreciation scores than regular programmes.

However, it isn't only comedy that garners a high mix of 10s. The *Shipping Forecast* has the second highest mix of 10s and a *Book at Bedtime* reading is in third place. While consumer programmes might get the highest proportion of 10s at an aggregate level, individual programmes do on feature in the top 10.

Comedy on Radio 4 is not the most polarising at an aggregate level, but at programme-level it is comedies that initiate the very highest proportion of very high or very low scores.

---

[369] See p152 – consumption bias.

[370] See from p48.

[371] B.A.B. denotes Book at Bedtime and B.O.T.W. denotes Book of the Week.

8.2 Answer to Research Question 2

The programmes on Radio 4 that attract the highest proportion of 10s and the highest proportion of 1s and 2s respectively are both comedy shows. However, while individual comedy shows can elicit extreme responses, religious programmes appear to attract polarised scores for the whole genre.

Where scores are more polarised, the mean is less representative as a summary figure. Therefore, comedy is particularly poorly represented by the AI score compared to the whole of Radio 4. There are implications for programming as they would not be made aware of the proportion of extreme responses. For example, *Count Arthur* may have a similar AI score to *Andrew Lawrence* (61 versus 59),[372] but what this summary figure doesn't tell us is that *Count Arthur* has a far higher proportion of extreme responses (ultra): 37% versus 20% respectively.[373]

---

[372] See figure 87.

[373]

Chapter 09 – Research Question 3 – Influence of (Non-Genre) factors on Appreciation Scores

It's not for nothing that I have, for nigh on thirty years, been chairman of a BBC radio show called
I'm Sorry I Haven't A Clue. *(Actually it very nearly is for nothing, but that's between me and the Corporation).*

(Humphrey Lyttlelton, 2006, pi)

As noted in Chapter 4, AI scores are used within the BBC as an indicator of the quality of programmes and of the performance of whole networks and channels. The analysis thus far has shown how the AI scores may be affected by a range of factors internal to the appreciation process (particularly the way different sections of the audience develop non-unimodal response patterns) as well as features of the programme and genre (such as the 'polarising' nature of comedy). However it may be that AI scores are affected by a range of factors external to either the process or the content of the programme rated. The final research question examines this possibility by asking if independent variables affect the appreciation scores regardless of the quality and sub-genre of a radio comedy programme. This chapter attempts to address this question by considering a number of factors relating to the production and scheduling of the programme (cost, series, familiarity of talent) and a range of demographic variables that are available for analysis through the Pulse survey. The data is presented using the chart developed that represents the mean score alongside the distribution of those scores – 'gromic'.[374] Aspects such as familiarity, series number and demographics do appear to affect the ratings. Finally this chapter illustrates how the findings might be applied to improve programme evaluation, using an example programme.

> RQ3: Are there underlying independent variables affecting appreciation ratings that can impact on comparisons of the distribution of appreciation scores across various Radio 4 comedy sub-genres?

The discussion has established that it might be expected that there are differences in appreciation for different comedy sub-genres, although the precise extent of those differences is not predictable.[375] While Carrie (1997, p131) found differences in AIs for television entertainment sub-genres, Barwise (2013) posits that those differences are not '*big*'.

Figure 113 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Sub-genre[376]

---
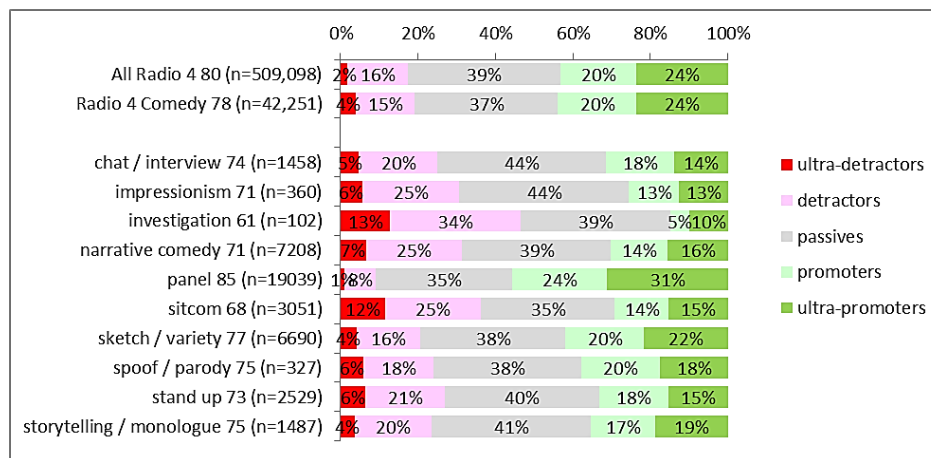
[374] See from p198.

[375] See from p114.

[376] See from p164 regarding allocation of sub-genres.

The figure above shows that there are observable differences in the score distribution for different sub-genres. 'Investigation', a small group that just included one series – *Tina C's Global Depression Tour* – fares poorly with the highest number of low scores (47% were 6 or lower). It also had the lowest proportion of promoters with just 15% of its scores being 9s and 10s. In comparison, 'panel' is clearly the most popular genre; it includes *Clue*, *Just a Minute* and *The News Quiz*. A mere 9% of its scores are 6 or lower while 55% are 9s and 10s.

The popularity of panel shows might not surprise us. They allow the comedians to be spontaneous while still retaining the producers' option to edit the recording to select only the best bits of the show, which can be identified by the response of the studio audience.[377]

> *You've got a lot more liberty on a panel show because you can talk for longer and it doesn't go in the final show.* (Lock, 2011)

> *They're cheap to make and they do deliver, so are terribly popular.* (Naismith, 2010)

They are also 'dippable',[378] meaning they do not necessarily demand much of their listeners. This could be of particular importance for a secondary medium.[379]

> On initial inspection, sub-genre appears to be an independent variable influencing appreciation scores

When using appreciation to understand the performance of a Radio 4 comedy, a programmer may compare the ratings to that of a similar sub-genre.[380] But there may be many independent variables, which affect radio comedy appreciation scores.[381] The following sections analyse the dataset to try to determine which other variables may influence appreciation scores.

9.1 Context - Familiarity – Series and Episode Number

It has been discussed how elements of familiarity could affect the appreciation for a programme.[382]

> *People tend to like things that they hear every week.* (Dare, 2012)

> New comedies tend to get higher AIs as audiences get used to the programme. (BBC Audience Research, 2009b)

This section examines the data set to see whether audiences tend to appreciate characters, situations and formats that they have simply become used to – familiarity being a factor that could theoretically increase appreciation.

---

[377] See from p60.
[378] See from p102.
[379] See from p78.
[380] See from p126
[381] See from p95
[382] See from p98.

Figure 114 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Number of series (just originations)
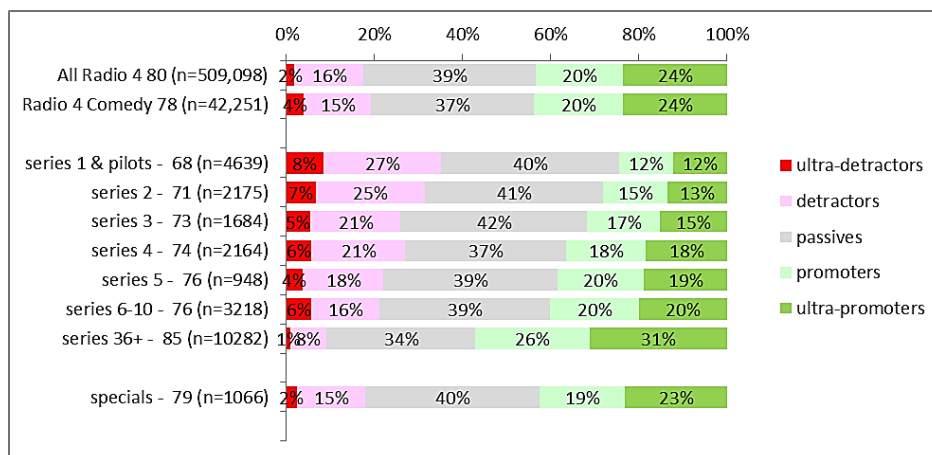
Figure 114 shows that as more series of radio comedy programmes are broadcast they tend to have fewer detractors and more promoters. Although there seems to be a blip at 6-10 series, this may be because this happens to be where **Count Arthur** falls and we have already shown that, across the whole of Radio 4, this is the programme that exceptionally attracts the most detractors. Stripping this programme out of the figures, the chart looks like this:

Figure 115 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Number of series (just originations) – Excluding Count Arthur
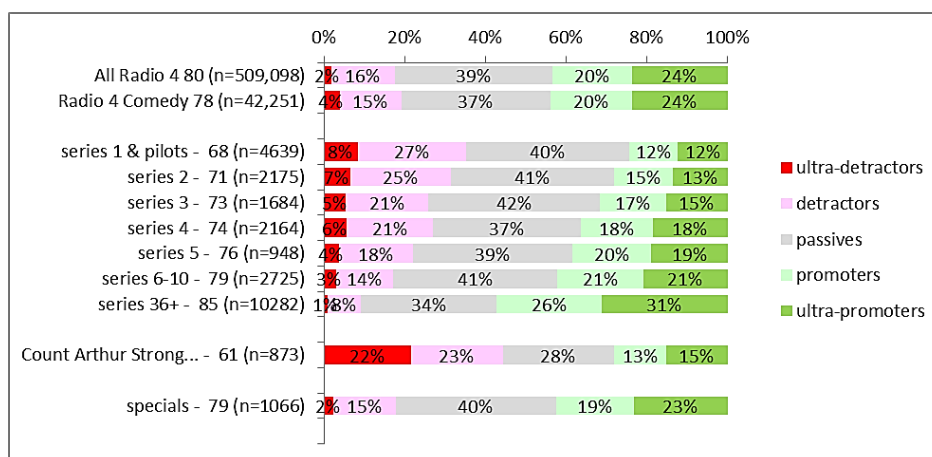
Without **Count Arthur**, the trends are even more clearly defined. The AI and the mix of promoters both increase steadily as the number of series increases while the proportion of detractors decreases. Thus, a programme in its first series could not be expected to achieve an AI of 85 when the 'norm' is 68. For example, the AI of the first ever episode of **JAM** in 1967 (albeit based on a different scale type) was **57** (BBC Caversham Archives), but it has an aggregate of **85** throughout 2012 on Radio 4. Likewise, **Clue** debuted at just **49** in 1972 (ibid) compared to **88** in 2012.

While this might be a good guide, that is all it is. Increasing number of series does not mean that we can *always* expect a low number of detractors and a high number of promoters, as can be seen from **Count Arthur**, which was in its 7th series during 2012.

However, this may not mean that, shows actually get measurably 'better' as they have more series. Most radio comedy shows (though not all) have the same format, the same writers, performers and producers throughout their run and sometimes fundamentally even the same jokes. At this point we should come back

to the way the BBC uses AI scores as a measure for quality:[383] while it has been deemed proven that appreciation ratings for Radio 4 comedies correlate extremely highly with evaluations of quality,[384] it is unknown *exactly* what respondents are evaluating when they 'rate' the programme they heard.[385] The figure above indicates that familiarity may be one of the factors involved.

A further consideration is that only programmes that are perceived to be 'good', or perhaps have the potential to be good, will get further series. Appreciation will increase as 'poorer' programmes fail to be recommissioned and are thus a decreasing element of the aggregate. Also, as posited by Collins (2010a), there are additional acknowledged phenomena that can be attributed to increasing appreciation. A steady growth in BBC AIs has been seen (albeit, more so for television than radio), attributed to increased media choice allowing people to consume only what they want. At programme-level this can mean that, increasingly, if a listener does not like a radio comedy show then they need not listen again in the future. Thus the proportion of detractors declines, resulting in a higher proportion of promoters and, in turn, a higher AI. This shows that, while we can say that incremental series generally exhibit higher appreciation, there may be a number of factors that may be involved in this, not *just* that familiarity breeds liking. Unfortunately, the dataset we have is limited to one year so we are unable to analyse across a number of series to see if respondents who give lower ratings are less likely to listen to future series. However, we can examine this aspect on a shorter term, for episodes within series.

Figure 116 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Episodes – just first series originations with four episodes each

Figure 116 shows aggregation of all the Radio 4 comedy shows that had their first series in 2012 and ran for 4 episodes, the most typical number of episodes for a new series. There were 22 different shows that fit these criteria. At this level, there appears to be no obvious relationship between the number of episodes and the appreciation score. However, the number of respondents (n) for each episode reduces – from 1039 to 928 – as the number of episodes increase — and this could relate to listeners who dislike the show ceasing to listen. The proportion of those who rated the programmes 3-6 (pale pink detractors) decreases with time, so we might infer that people who are unimpressed are less likely to come back. However, the proportion of ultra-detractors does not show a steady decrease. It could be argued that if people really dislike a programme, they might wish to listen again in order to check how bad it really was. Do people actively want to hear something that they can complain about? Some people will certainly come back again and again to a programme that

[383] See from p88.

[384] See from p203.

[385] See from p126.

they rate very poorly.[386] Alternatively, we might consider that some people are coming into the series having not heard earlier episodes and with less introductory context, meaning that their level of dislike could be higher than if they were hearing the very first episode. We can account for this factor if we solely consider those who claim to have heard *all* the episodes:

Figure 117 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – episodes – just first series originations with four episodes each – Only respondents hearing all episodes

Figure 117 shows how appreciation scores progress throughout a series – albeit using a relatively small sample size – for those who heard *every* episode of a series. There is no obvious increase in appreciation across these 22 series so we cannot say that familiarity appeared to increase appreciation. This finding contrasts with earlier findings, such as the work of Carrie (1997, p206) who found that the more television episodes watched, the higher the appreciation.[387] This indicates that while appreciation increases across a number of series (see figure 115) – a timescale of a number of years – we do not see it over weeks.

If people consume the media that they like (Carrie, 1997, p24), then we might expect higher appreciation scores to relate to higher frequency of listening. Carrie (p207) did indeed find that, when considering television AIs, 'in all cases, frequency of viewing is strongly correlated with individuals' appreciation.'

Figure 118 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – episodes – just first series originations with 4 episodes First episode only – respondents hearing all episodes vs respondents hearing just the first episode
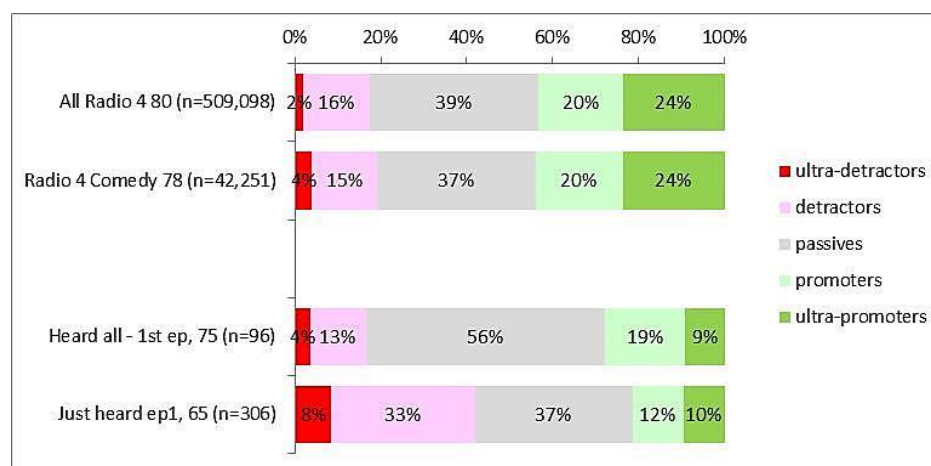
Figure 118 helps examine whether poor appreciation scores for a first episode mean that a person is less likely to go on to listen further, illustrating the difference in the ratings between listeners who gave a

[386] See from p223.

[387] See from p105.

response to every episode of a series versus those who just rated the first episode. Those who rated only one episode were far more likely (42% compared to 17%) to have given a low score (1-6). Something similar occurs with promoters too; those who marked highly (9s and 10s) were more likely to go on and rate the whole series (28% vs 22%).

> As the number of series increases, aggregate appreciation has a tendency to increase as fewer people give low scores and more give higher but his is not observable within a series, nor is it universal as some programmes (for example, *Count Arthur*) can be exceptions to this pattern.

9.2 Context - Familiarity – Repeats Versus Originations

A further variable which might affect appreciation may be the distinction between programmes which are originations, (TXing for the first time) and repeats. We can compare the appreciation scores for repeats versus originations. However, the analysis has shown that the very longest-running series are prone to particularly high appreciation scores,[388] and it is these programmes that receive narrative repeats[389] on Radio 4. To see the real effect of repeats on appreciation scores, these broadcasts have been excluded from the following chart, and the remaining data split into originations and ad hoc repeats.[390]

Figure 119 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Gromic – Originations vs ad hoc repeats – Excluding programmes with narrative repeats
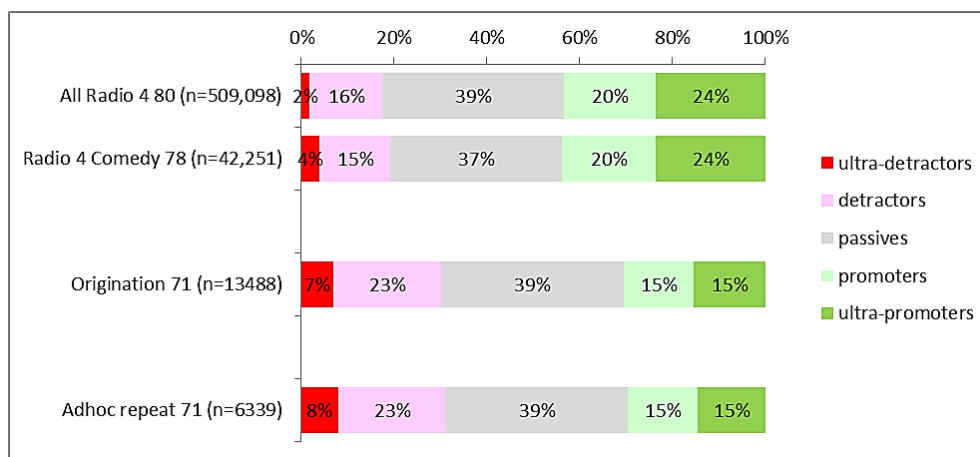
Figure 119 illustrates that there is virtually no difference between the aggregate AI of repeats and originations. Thus, we do not need to take repeat status into consideration when analysing genre differences. This could be contrary to expectations if we consider that some of the listeners may have heard a repeated programme before, giving them a higher level of familiarity.[391] It could be that this positive effect of familiarity (if there is one) may be countered by a level of dissatisfaction from others in having to hear a programme again. However, the fact that a programme is a repeat does not mean that it has been heard before by a particular respondent; repeats tend to be placed in slots in order to reach new audiences compared to the origination.

> At an aggregate level, there is little difference in appreciation for originations versus ad hoc repeats.

---

[388] See p228.

[389] A narrative repeat is a repeat that TXs within a week of the origination.

[390] An ad hoc repeat is a repeat that TXs more than a week after origination – typically around 18 to 36 months later.

[391] See from p98.

9.3 Context - Familiarity – Promotions

Familiarity of a programme could also relate to 'advertising',[392] and we can investigate this factor by analysing responses for Radio 4 comedies which are selected to have on-air promotions (programme trails).[393] To consider the effect of trails alone, we need to exclude long-running shows as they would have a pre-existing level of familiarity; thus the chart below shows only appreciation for programmes whose *first* series or pilot premiered during 2012. These have been split into programmes that had trails on Radio 4 ahead of their TX and those that did not.

Figure 120 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Gromic – Trails – First series first episodes and pilots with and without on-air support
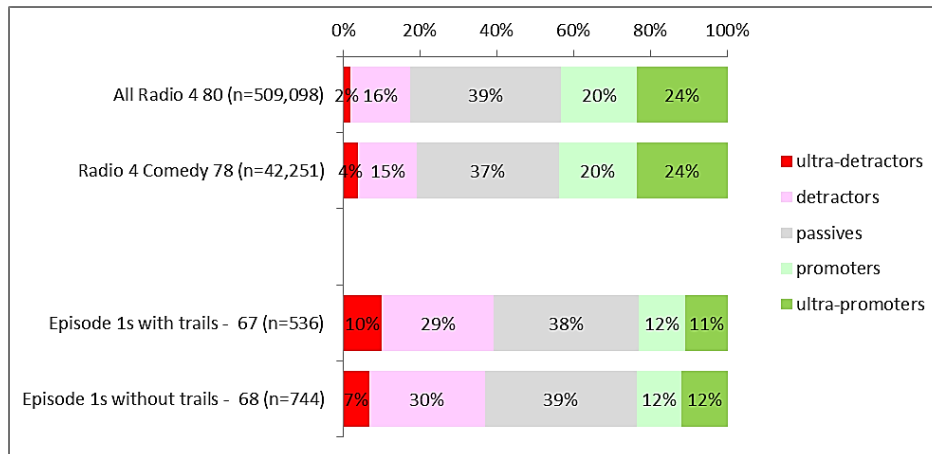
While we might expect increased familiarity from the trails to have a positive effect upon the appreciation for a programme, this is not borne out in the data; if anything the opposite is true. Programmes that had trails received a higher proportion of scores of 1 and 2 (10% vs 7%) than those that did not, and slightly fewer 10s. Programmes receiving trails actually had a lower aggregate AI: 67 versus 68.

There could be a number of reasons why trailed programmes do not see any increased appreciation:

- The raison d'etre of programme trails (as opposed to brand trails) is primarily to make people listen, particularly people who might not otherwise. This means that the process may attract listeners who might not typically listen to radio comedy, and they could possibly be more likely to dislike the genre.

- The choice of which programmes to promote is the responsibility of the Radio 4 Controller, with additional input from the Head of Planning and Scheduling and the Presentation Editor. There may be an element of bias in which programmes are selected for promotion. For example, if it were the strategy to promote newer, edgier programmes over safer ones, we might expect more resistance from listeners to these programmes.

- While it might be hoped that exposure from the trails will increase appreciation, we have no information about the quality of the trails. Hubbard (2012) claims that comedy trails, in particular, can present their own problems:

  > With comedy trails, you have to be very careful that you're not playing the same gag over and over again, so everybody can hear the punchline coming from Kent.

This dataset provides no evidence that familiarity through trailing increases appreciation. There is little difference in the aggregates scores for programmes that had on-air trails versus those that did not.

---

[392] See from p98.

[393] Heavy Radio 4 trail campaigns will reach around 8 million adults – the majority of the 11 million Radio 4 listeners. Figure supplied by Radio 4 media planning.

9.4 Context - Familiarity – Talent

We have discussed how well-known talent can be a component of familiarity in Radio 4 comedies, as an element of risk mitigation.[394] Where a new format is involved, it may be possible that a listener is more likely to be appreciative of the programme if the talent involved is already known to them. Figure 121 shows the analysis of first series or pilots within our dataset. In attempting to compare those with known talent to those whose talent is comparatively less familiar, each programme was discussed with the Radio 4 Commissioning Editor to ascertain whether the programmes, although new to Radio 4, might have contributors who were already generally known to the listeners. Programmes with well-known talent were so categorised because they included people such as:

> Alison Steadman, Anthony Head, David Nobbs, Frank Skinner, Gyles Brandreth, Hugh Bonneville, Jack Docherty, Juliet Stevenson, June Whitfield, Miles Jupp, Nicholas Lyndhurst, Reece Shearsmith, Richard Wilson, Stephen K. Amos.

Figure 121 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Talent – First series and pilots
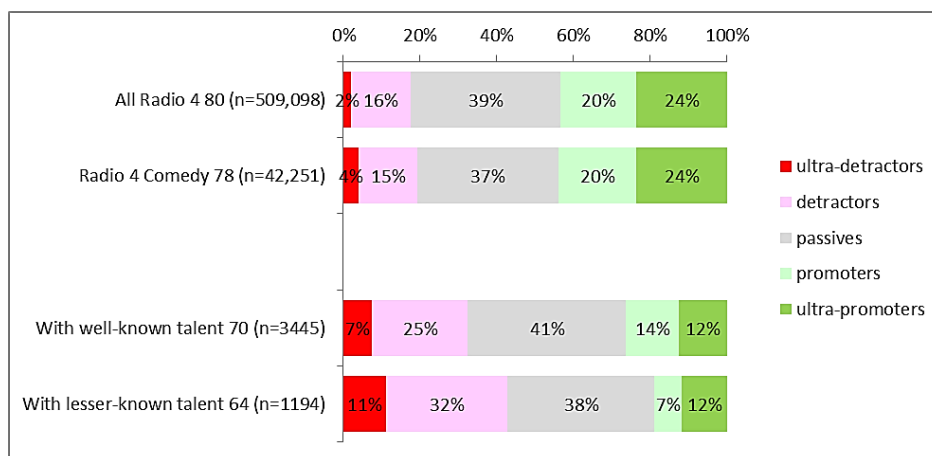
For this dataset, programmes that included well-known talent have a higher proportion of promoters (26% vs 19%) and fewer detractors (32% vs 43%). The net effect is an increase of +6 points on the AI, suggesting that the presence of known talent in a new radio comedy series increases mean appreciation. However, there are a number of factors to consider. Firstly, the categorisation of 'talent' was subjective. For example, *The Guns of Adam Riches* was not included as containing known talent because, despite the fact that Adam Riches had won the Edinburgh Comedy Award in 2011, Riches was not necessarily well-known to the Radio 4 audience. Secondly, there is the possibility that experienced performers could be more likely to be involved only in programmes that actually are *better*. They might have their pick of the scripts and the available producers, whereas this might not be the case for newer talent. Therefore, it might not be the talent alone that is necessarily drawing the appreciation. Causality is unclear. Are programmes higher quality and more appreciated because they contain known talent, or do they include known talent because they are superior?

> Evidence suggests that known talent in a radio comedy programme is linked with higher appreciation but causality is not clear.

---

[394] See from p99.

9.5 Context – Expectations – Cost of the Programme

Another explanation for the relationship between well-known talent and higher appreciation might be higher investment in the programmes. Chapter 4 discussed audience expectations when listening to a Radio 4 comedy;[395] a programme with higher production values and more costly talent might be more likely to meet or even exceed the expectations of the listeners.

The set price for Radio 4 comedy shows in 2012 was £11,300 per half hour,[396] and in this respect the network has no visibility of the actual cash investment into each programme. However, while this is the average paid, the Radio Comedy department monitor the exact costs associated with each programme. Thus we can compare the cost per 28 minutes with the appreciation; in the case below this is plotted against the proportion of 9s and 10s (promoters) that each programme received.

Figure 122 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Scatter – Programme cost vs proportion of 9s and 10s – Radio Comedy dept, n>99
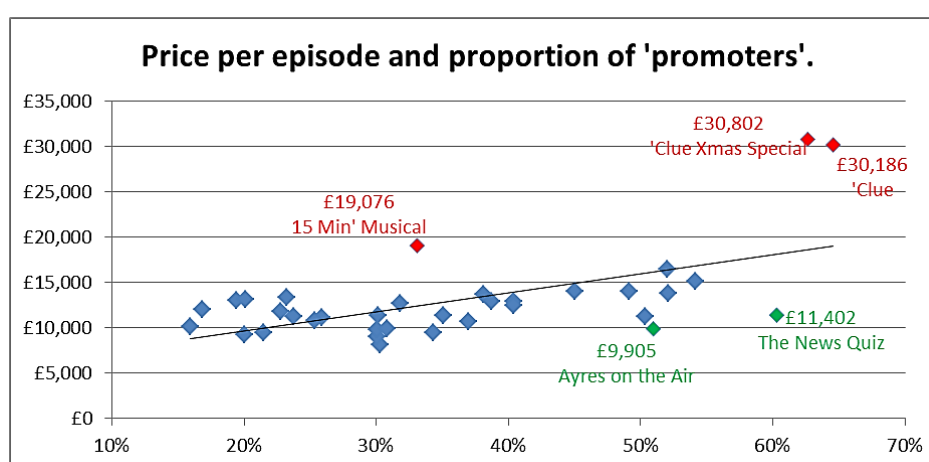
The trend line indicates that there is a positive relationship between the cost of a programme and the proportion of people who give it a 9 or 10 appreciation score. In other words the more expensive a programme is to make, the greater the proportion of people who really like it. Programmes below the trend line are more 'efficient' in attracting high appreciation, while conversely those above the line are more inefficient. For example, we can see that *15 Minute Musical* has a particularly high cost of £19,076 per 28' but only 33% of its appreciation scores are 9s and 10s. This compares unfavourably with a programme such as *Ayres on the Air*, which is around half the cost (for the equivalent duration) at £9,905 but garners 9s and 10s for 51% of its ratings. Similarly, *Clue* attracts the highest proportion of those who love it, but does so at an extremely high cost compared to something like *The News Quiz*, which is around a third of the price but still with nearly as high a proportion of 9s and 10 as *Clue*.

A similar, yet inverse, effect can be seen when we compare cost to low appreciation ratings (1s and 2s).

---

[395] See from p96.

[396] See from p44.

Price per episode and proportion of 'ultra-detractors'
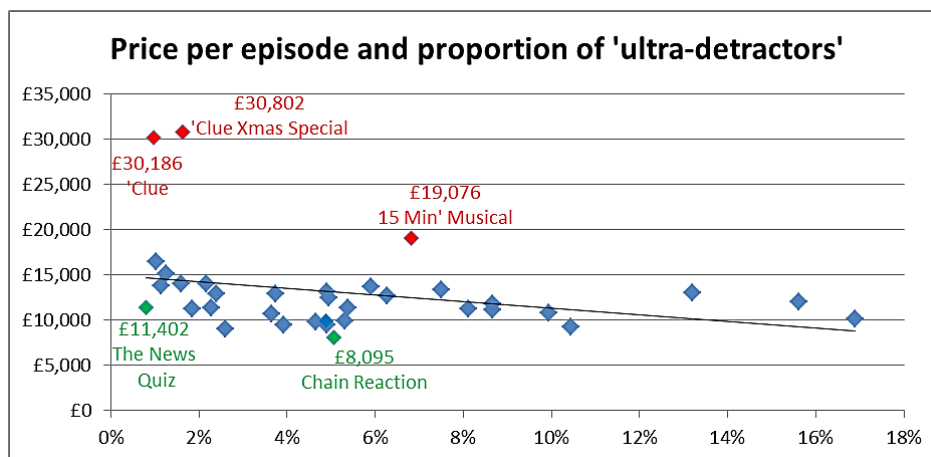
The direction of the trend line indicates where more money spent on a programme means that there will be fewer very low appreciation scores, or 'ultra-detractors'. Looking at *15 Minute Musical* again, its cost is more than twice that of *Chain Reaction* but *Chain Reaction* only attracts 1s and 2s 5% of the time whilst the figure for *15 Minute Musical* is higher at 7%.

However, we have seen earlier that there may be a strong relationship between the number of series a programme has and its appreciation. Therefore, we need to look at programmes that have had a comparable number of series.

Figure 124 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Scatter – programme cost vs proportion of 9s and 10s – Radio Comedy dept, n>99 – first series and pilots only

Price per episode and proportion of 'promoters' 1st series.
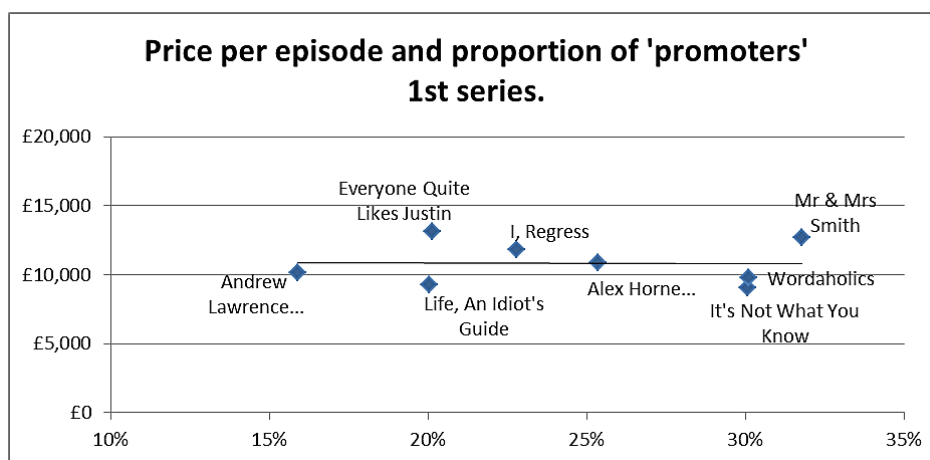
When we consider only a first series, the positive relationship between cost and 9 and 10 scores is not apparent. Although the figure above allows us to see which programmes are most cost/appreciation efficient, we cannot say that spending more money on a radio comedy show will result in higher appreciation. *Everyone Quite Likes Justin* and *Mr and Mrs Smith* both have very similar costs, but the latter has a far higher proportion of people who rate the programme highly.

Higher production costs do not necessarily have a correlation with improved appreciation.

9.6 Context – Expectations – Time of Day

We have considered that the time of day may have an effect upon appreciation.[397] Preunkert (2009), looking at AIs for German television, found that appreciation was lower in the morning. Is this phenomenon observable for Radio 4 comedy? The chart below splits the comedy originations by the daypart in which they were broadcast. As it has already been established that there is an effect linked to the number of series, the data has been further split by level of programme establishment, i.e.: 'new' = up to 4 series; 'established' = 5-10 series; and 'core' = 11 or more series.[398]

Figure 125 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Gromic – Time of day split by level of establishment
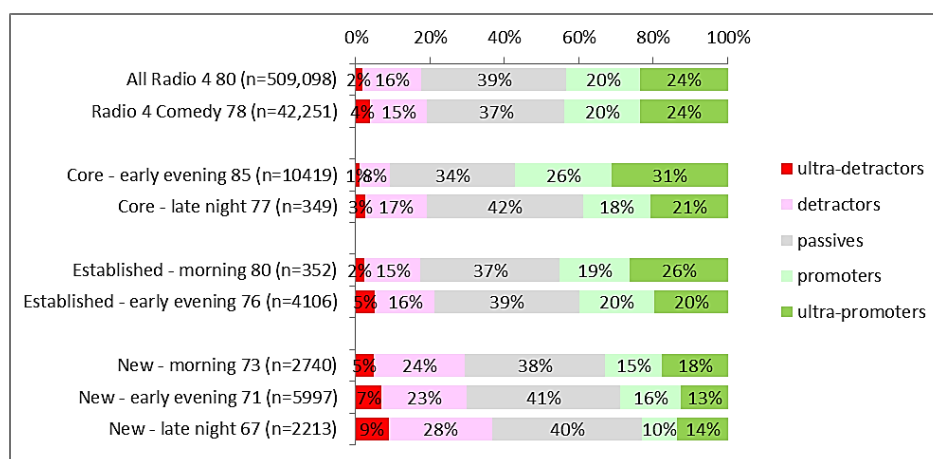
Figure 125 shows, contrary to Preunkert's findings, that programmes TXing earlier in the day tend to attract higher appreciation scores and appreciation seems to decrease as programmes are placed later in the day. This may relate to the requirements for the different slots. The guidelines for the comedy slots (BBC Radio 4 Commissioning Guidelines Spring 2011)[399] state that the 11:30 morning programmes are commissioned with the idea of consisting of: 'a plot or structure that makes it easy for listeners to follow the narrative – this is still a busy time of day' (p50), while 23:00 evening programmes are more likely to be, 'new innovative formats that will bring surprise to the slot both in terms of content and sound or subverting a familiar genre' (p55). We might therefore expect relatively fewer low scores in the mornings and relatively more dislike for more challenging formats that are heard later in the day.

We may be able to obtain absolute figures by looking at Radio 4 Extra (R4X) figures. On this network, comedy shows are nearly all repeats and programmes tend to be repeated throughout the day on a loop, i.e., the same programme can be heard more than once per day. On 11th June 2012, for example, *Steptoe and Son* could be heard at 09:00, 13:00 and 20:00. This contrasts with Radio 4 where programmes are bought to originate in specific slots. Thus, using the R4X appreciation scores, we can attribute daypart differences to other independent variables, rather than content.

---

[397] See p102.

[398] Whilst this segmentation is in essence arbitrary, it does reflect attitudes of the programmers.

[399] Spring 2011 would have been the commissioning round when the majority of 2012 programmes were commissioned.
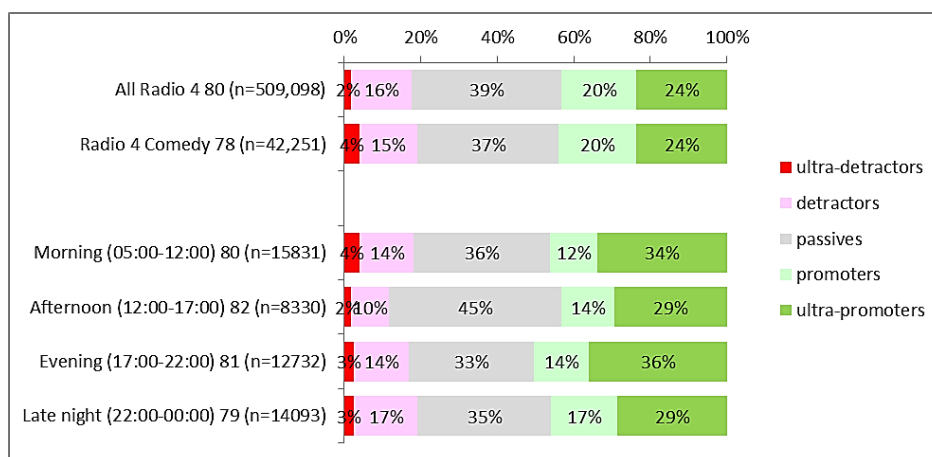
Although the aggregate AI scores for R4X comedy by daypart only vary between 79 and 82, looking at the score distribution there appears to be noteworthy variations. In the afternoon, detractors and promoters (i.e., those who like or dislike a programme) account for 55% of the ratings. However, in the evening this increases to 67%. While it would be hard to prove causality (as Preunkert, 2009, also found) we might theorise that: this variance occurs either because of variations in audience composition,[400] that it is attributable to changes in the propensity of group listening[401] or because comedy may work better at certain times of the day[402] (usually the evenings, e.g., Cantril and Allport, 1935, p222; Bennett, 2011).

There has been some research on the role of 'adjacencies' as an independent variable for appreciation,[403] i.e., which programmes lead into and out of the shows being considered. For Radio 4 comedies, the type of programmes adjacent to these programmes are mostly unchanging for each time slot. For example, the 18:30 shows are always preceded by *The News* and always followed by *The Archers*. Hence, there is no way to separate the effect of programme adjacencies from the effect of the time of TX for the Radio 4 comedies.[404]

While data for Radio 4 comedy shows indicated that appreciation is higher earlier in the day, this may be explained by the content of the programmes. Radio 4 Extra analysis indicates that there may be absolute differences in appreciation across different dayparts, but there is no clear progression observable.

9.7 Context – Expectations – Dippable or Not

Radio programmes can vary in structure in terms of being 'dippable' or not,[405] i.e., whether they can be enjoyed without full attention. Panel games, for example, support casual listening since there is no overriding narrative across the programme. Radio is recognised as a secondary or even tertiary medium,[406] and we might expect programmes which are more dippable to have higher appreciation since enjoyment may be less compromised when the listener is also involved in an alternative and primary task.

---

[400] See from p102.

[401] See from p100.

[402] See from p102.

[403] See p103.

[404] Further work on interstitial adjacencies could have been conducted (as per Benson and Perry, 2006) but it was outside the scope of this research.

[405] See from p102.

[406] See from p78.
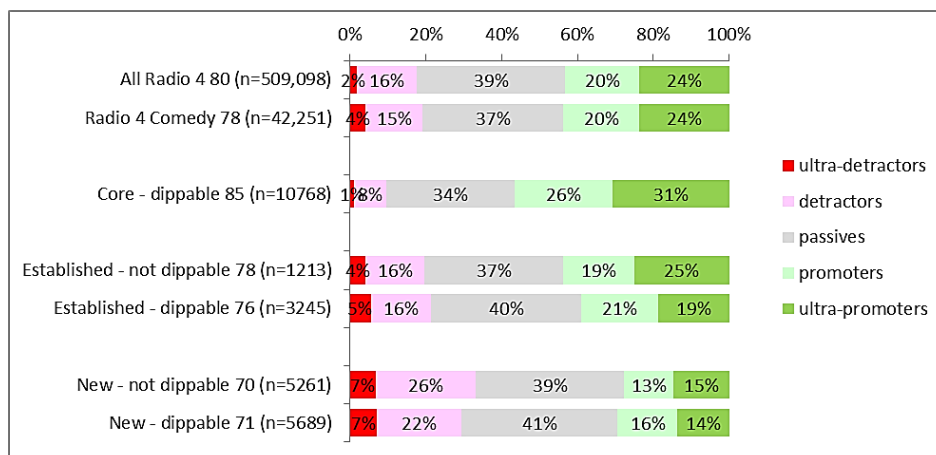
The chart above shows appreciation for the different levels of programme establishment split by whether they were dippable or not. All of the 'core' programmes are dippable panel shows, programmes such as *Clue* and *JAM*, and are highly appreciated. However, looking at the 'established' programmes we can see that appreciation is higher for programmes that are *not* dippable, with a greater proportion of promoters and fewer detractors. Programmes in this group include *A Charles Paris Mystery* and *Ed Reardon's Week*; these are examples of shows with a narrative arc that would benefit from full attention. For new programmes, there is a lesser difference between the two types.

As mentioned in chapter 2,[407] commissioners seek different types of programmes for different times of day and there is evidence that dippability is a consideration. For example Counterpoint Research (1997c, p6) indicated that ideal early evening radio '*can be dipped into, and easily left behind*'. This view is supported by Raphael (2011): '*It has to be dippable*'. Early evening in particular is s a time when listeners are busy with primary tasks with the radio taking second place (Counterpoint Research, 1997c, p6). The data allows us to compare dippability across dayparts:
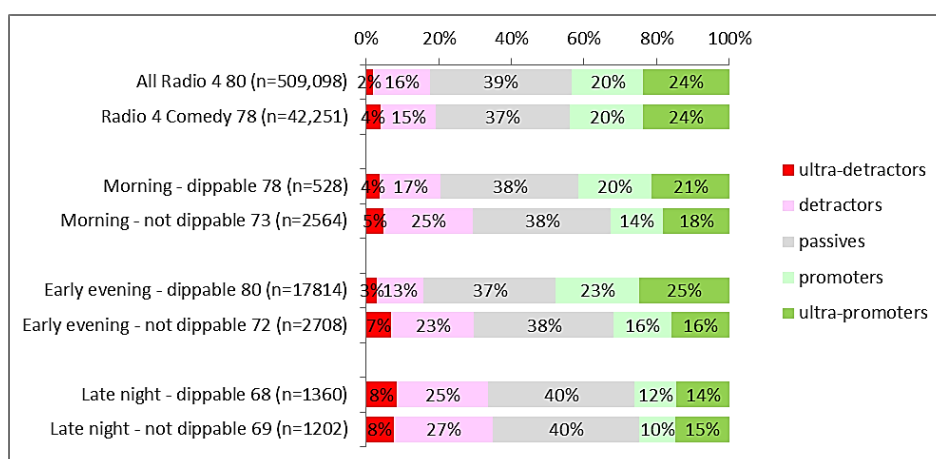
Figure 128 shows that, for early evening, dippable programmes do indeed get higher appreciation scores, having more promoters and fewer detractors. However, this is very much driven by the 'core' programmes that have been running for many years. If we strip these out, we see the following:

---

407 See from p37.
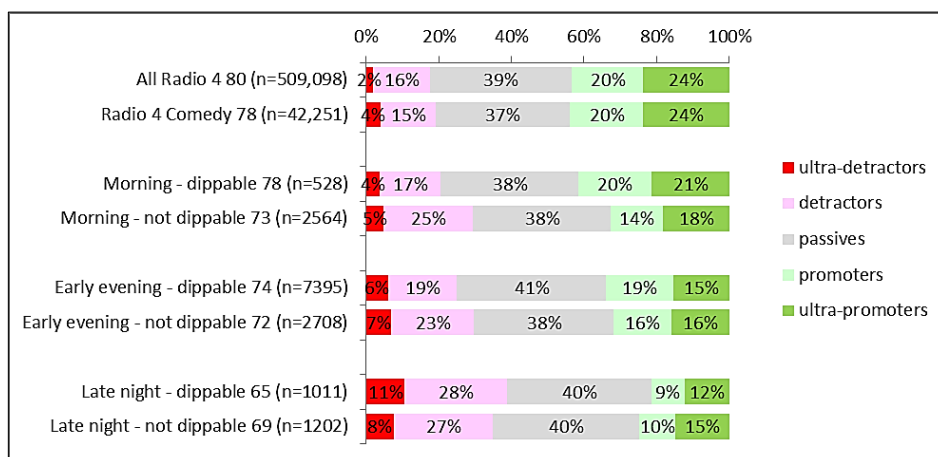
Without the 'core' programmes, dippable programmes have a slightly higher appreciation, but it is actually in the morning where dippable programmes are comparatively more highly appreciated. Late night, where people are less likely to be engaged in primary activities, dippable programmes actually have a lower appreciation. This relates to the discussion about the *utility* of a radio programme – see from .

> A programme's dippability appears to be a factor in appreciation but the effect can be positive or negative depending on the daypart.

9.8 Context – Listening With Others

We have discussed how comedy appreciation can be affected by group listening in the chapter on appreciation.[408] Unfortunately the Pulse survey does not hold information on whether the respondents are listening alone or not. The nearest information that we might use for this aspect is discussed later in the section: 'With or Without Children'. Whether or not a show has a laughter track could also relate to listening with others, which is covered in the section to follow.

9.9 Context – Studio Audience

It has been found that studio audiences are seen as a proxy for a group and can, on occasion, enhance appreciation.[409] Is the sound of laughter a factor in the appreciation for Radio 4 comedies?

---

[408] See .

[409] See .

Figure 130 shows that the core programmes are all shows with a studio audience. For established programmes, shows with audiences actually have a slightly lower appreciation. However, for newer programmes, it is the other way round; newer shows with audiences have slightly higher appreciation scores – i.e., fewer detractors and more promoters. Is it the case that newer shows benefit from having an audience in their early days, the heard laughter of others telling the listener that it is indeed funny? Perhaps before a comedy has time to bed in[410] it could be an aid to engagement and appreciation. We can consider this if we look purely at first episodes:

Figure 131 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Gromic – With and without studio audience – Just first series and pilots – very first episodes only
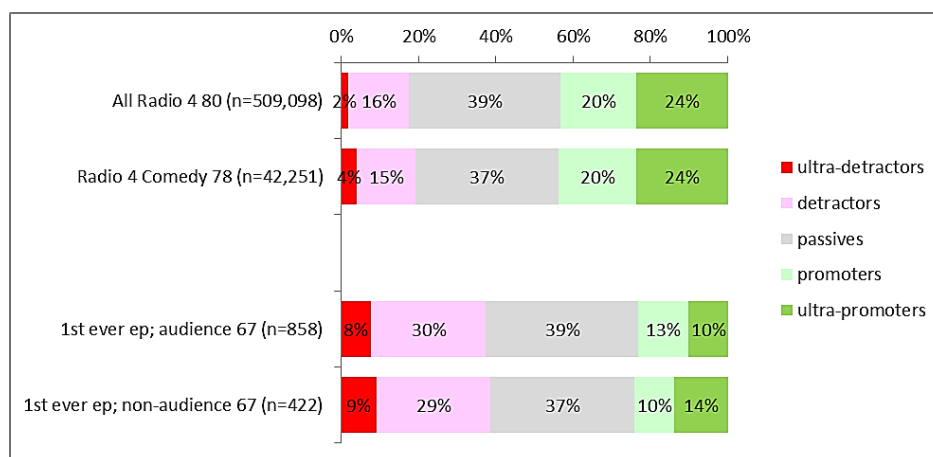
The purpose of the audience is to break down the barriers and increase intimacy for the listeners (McQuail, 1997, p117), and aid mirth by highlighting where the jokes are (Martin and Gray, 1996, p221). If studio audience laughter does successfully do this, we might expect the very first episodes of a new show to benefit from having an audible audience. Yet figure 131 shows that the very first episodes of all the shows had a mean aggregate score of 67, regardless of whether they had an audience or not. The main observable difference is that it is actually the non-audience shows that attract a higher proportion of ultra-promoters and ultra-detractors: 23% as opposed to 18%. Shows with audiences seem slightly less likely to attract extreme scores.

Looking at aggregate scores here may be of limited use. Lieberman et al (2009, p502 p512) highlight the suggestion that to understand the effect of studio audience laughter, episodes within a series need to be judged individually as varying narrative structures can be influenced positively or negatively. Despite the reasonable size of our dataset, drilling down to series/episode level reduces sample sizes to unusable levels, often fewer than fifty responses.

There is no clear evidence whether a studio audience improves appreciation or not. This is not unexpected as humour research on this topic is often inconclusive or inconsistent.[411]

---

[410] See from p126.

[411] See from p101.

9.10 People – The Demography of the Audience

As discussed, there is research indicating a relationship between demographic variables and comedy appreciation.[412] The Pulse survey holds gender, age and social status data on each respondent, so it is possible to analyse these aspects in relation to their responses (albeit within the understanding that response pattern type skews may be partly contributing to the differences).[413]

9.10.1 Gender

Figure 132 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Gender comparison
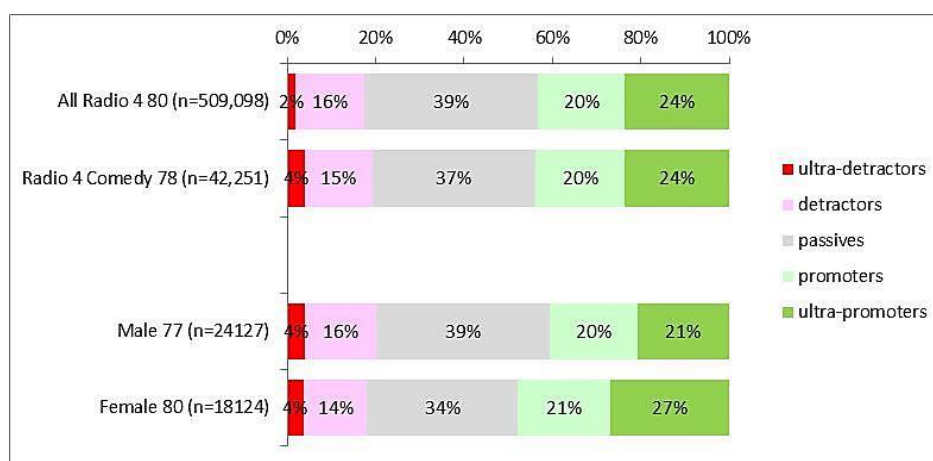
The chart above shows the differences between men and women's scores for all of the Radio 4 comedies. Overall, men give a lower AI (77 versus 80) supporting the findings of: Menneer (1987, p254), Carrie (1997, p24), and BBC research (BBC Audience Research, 2009b) that claims that women tend to rate 2-3 points higher than men. While this difference is relatively small, our analysis shows that it is derived from a differently shaped distribution. Women are far more likely to give scores of 10 compared to men; 27% of their ratings are 10s compared to just 21% for men. This in turn may reflect gender specific response patterns. Chapter 7 demonstrated that women have more of a tendency to mark in a 'Silvey' pattern than men; 30% of the women analysed marked in a Silvey pattern, as opposed to 24% of the men.[414] As the Silvey pattern gives an aggregate AI that is far higher than the other patterns – 93 versus 76 for unimodal[415] – the mere fact that women have a greater tendency to use this pattern means that their AI is likely to be higher.

Is it the case that women just give higher ratings generally? In one study, Studer (2011, p1-p10) found that (as in an earlier study) when measuring happiness, women gave higher ratings. However, when the scale was changed to a continuous (VAS) scale, the difference between men and women disappeared. This implies that the gender difference in ratings can be altered by the scale used.

North (2010) claims that these differences may not be *just* because women rate more highly than men as it could be partly due to the mix of broadcasting: '*It's the composition of the choice of programme*'. The figure above is an aggregate across a range of programmes but it could be that the gender mix varies from programme to programme, which could affect the aggregate. We can strip out the '*mix effect*' if we just look at individual programmes:

---

[412] See from p105.

[413] See from p190.

[414] See from p190.

[415] See from p190.
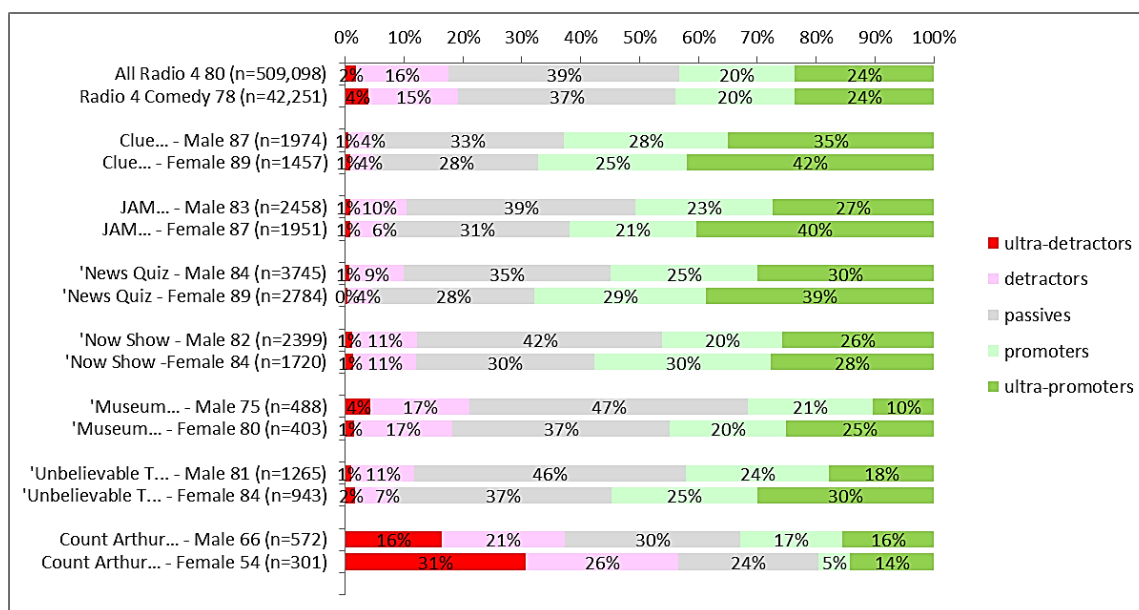
The chart above illustrates examples of Radio 4 comedies that had particularly large sample sizes (n>800).[416] For the first 6 examples, there is always a greater number of female promoters than male promoters (9s and 10s) and roughly equal numbers of detractors (1s-6s). This means that for these programmes the AI score is higher for women by between 2 and 5 points. However, in the last example – **Count Arthur** – women rate it more poorly than men, the percentage of ultra-detractors being particularly marked. Women are nearly twice as likely to give scores of 1s and 2s than men. The resultant AI score is 12 points lower for women than for men (54 versus 66). Therefore, looking just at topline figures, there is an indication that women do generally rate Radio 4 comedies more highly than men, but there can be exceptions. Despite inconsistencies in this area of research,[417] some studies have found that different types of comedy appeal to a greater or lesser extent to each gender (for example, Sanner et al, 2001, p4-5). We saw earlier that **Count Arthur** is the Radio 4 programme that attracts the most ultra-detractors,[418] and the figure above indicates that there must be elements of the programme that are particularly distasteful to female listeners.

The Gender of the Performer

The listener can't always be sure of the age, ethnicity or sexuality of the performer on the radio, but their gender is usually evident. There is evidence that there may be bias against female comics among male audiences in particular (Morinan, 2010b).[419] The chart below shows the appreciation scores for all Radio 4 comedies split by the gender of the performer and the gender of the respondent.

---

[416] The responses to the seven programmes, included above, account for 53% of all Radio 4 comedy ratings in 2012.

[417] See from p105 and p108.

[418] See p223.

[419] See from p105 and p108.
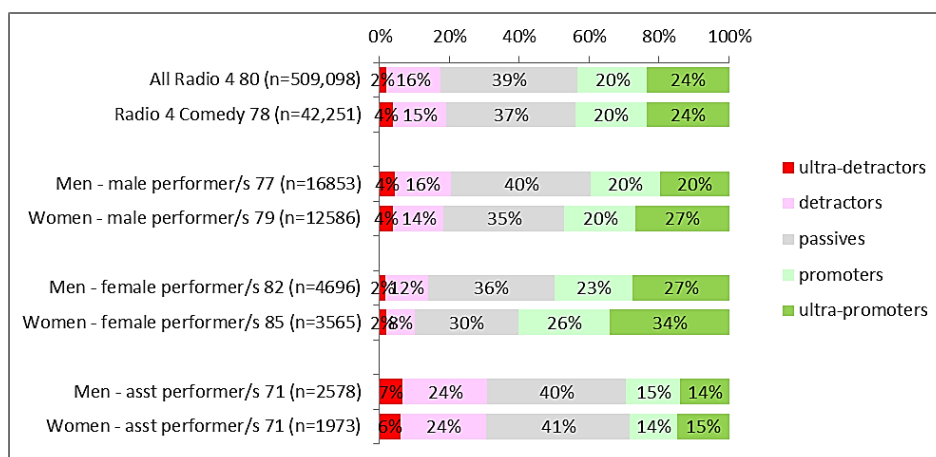
Contrary to expectations, programmes 'hosted' or dominated by female performers rate more highly with both men and women. Men rate these programmes 5 points higher (82 versus 77) and women 6 points higher (85 versus 79). This could seem surprising when we consider that the most popular show – *Clue* – is a male dominated show, only occasionally having a female guest as one of the five performers. Also of the comedies that attract the highest proportion of ultra-detractors, none are female dominated:[420]

Count Arthur	Male
Andrew Lawrence	Male
Mr Blue Sky	Male
Sketchorama	Assorted (but mostly male)
The Simon Day Show	Male

Thus, R4 comedies with mainly male performers tend to be loved or hated more than those with women when we look at programme-level. One could argue that the comedy coming from women tends to be safer, as it is drawing fewer extreme responses. Further analysis could indicate whether women tend to be involved in less divisive comedy genres or styles, but the samples size of this dataset precludes this further work.

There is little difference in how men and women rate programmes that do not have a particular male or female performer skew. Overall, however, the aggregate AI is much lower than that for programmes that have a preponderance of performers of one gender: 71 vs 78 for all Radio 4 comedy.

Overall, women respondents rate more highly than men. This can partly be explained by the fact that they tend to use a response pattern that has a higher aggregate AI. However, it is also driven by the content of the programmes as, while women tend to rate individual programmes more highly than men, some programmes can be particularly distasteful for women in particular. The use of a numerical scale rather than a continuous one could be a factor in the gap between the genders.

The gender of the performer appears to be an independent variable; programmes with female performers generally appear to fare better than those that are male dominated.

---

[420] See p223.

9.10.2 Age

Carrie (1997, p24) found that older television audiences 'tend to be more generous in their scoring patterns than younger viewers'. BBC audience research (2009b) also found that respondents over 65 tend to give higher mean aggregate ratings: 3 or 4 points higher than younger audiences, and the very youngest give the lowest scores. Menneer (1987, p254) also found that older respondents give higher ratings.[421]
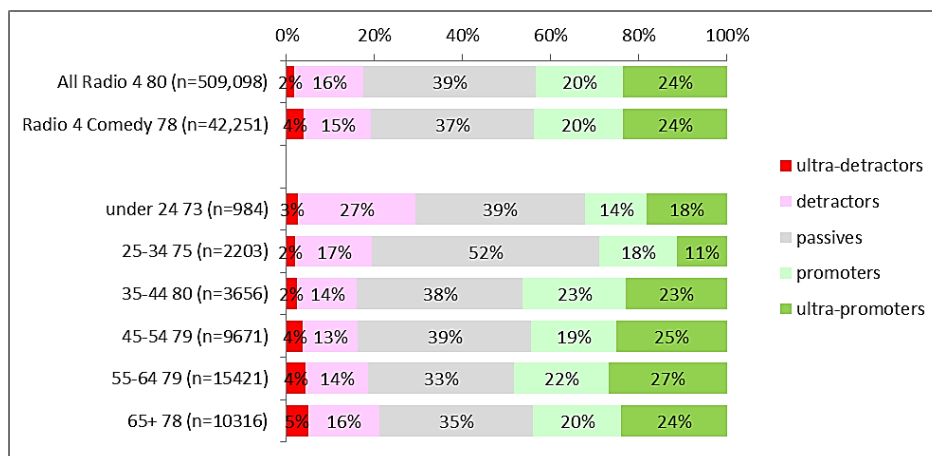
In the figure above, we see that for Radio 4 comedy there is no clear trajectory for increasing appreciation, contrary to the aforementioned findings. In fact it appears that appreciation peaks for the 35-44-year-old age group (AI of 80) and falls away either side. However, as discussed, it is an accepted aspect of comedy that new, more challenging formats appeal more to younger audiences than to the older ones.[422] Can this be illustrated with our data?
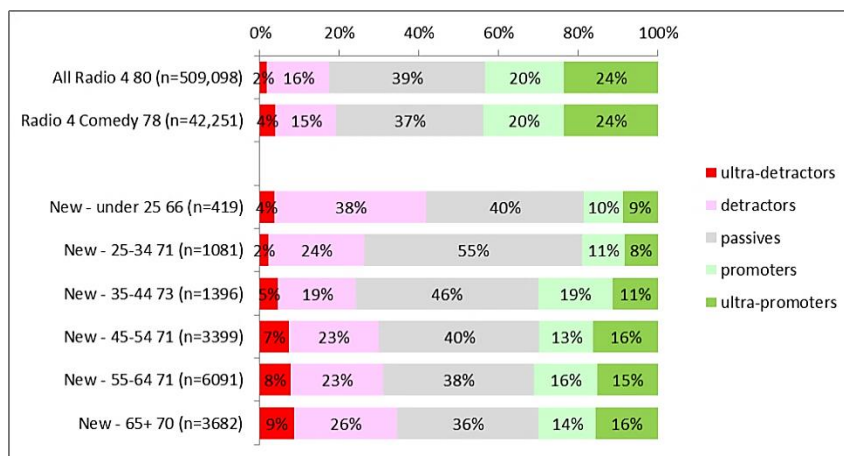
Figure 136, showing the split of ages for 'new' (fewer than 5) comedy series, clearly does not support the theory[423] that newer comedy appeals more to younger audiences. The most 'tolerant' group is the 35-44s, while the under 25s actually give the highest proportion of detractor scores – 42%. One factor that could be at work here relates to the nature of Radio 4 programming, what might be 'new and challenging' compared to other Radio 4 shows, may not be for these younger listeners. This may reflect a deliberate commissioning approach. Radio 4 often targets 35-54s specifically (BBC Executive, 2010, p17) so there might not be a

---

421 See from p105 and p108.
422 See p41.
423 See from p41.

requirement to be at the very sharpest cutting edge of comedy, as there is a fine balance between appealing to younger audiences and not alienating the older ones.

What does appear to be the case for all, including new Radio 4 comedies (i.e., the 2 figures above), is that from 35-44 onwards, the mix of low ratings (both ultra-detractors 1s-2s, and detractors 1s-6s) increases as the respondents get older. This indicates a growing intolerance for Radio 4 comedy generally as people age. This is contrary to finding that older listeners have a tendency to use the Silvey pattern for all Radio 4 and therefore are more likely to use the highest scores (see figure 72), indicating that comedy may be an exceptional genre in this respect.

> For Radio 4 comedy, older audiences do not necessarily give higher appreciation scores. However, there is a tendency for audiences to become more likely to give slightly poorer ratings as they move out of the 35-44 age bracket.

9.10.3 Social Grade

As with gender and age, research by Menneer (1987, p254), Carrie (1997, p24) and BBC researchers (BBC Audience Research, 2009b) drew a consensus that respondents of higher socio-economic grades tend to give lower appreciation scores.[424]

Figure 137 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows / originations – Gromic – Social grade comparison
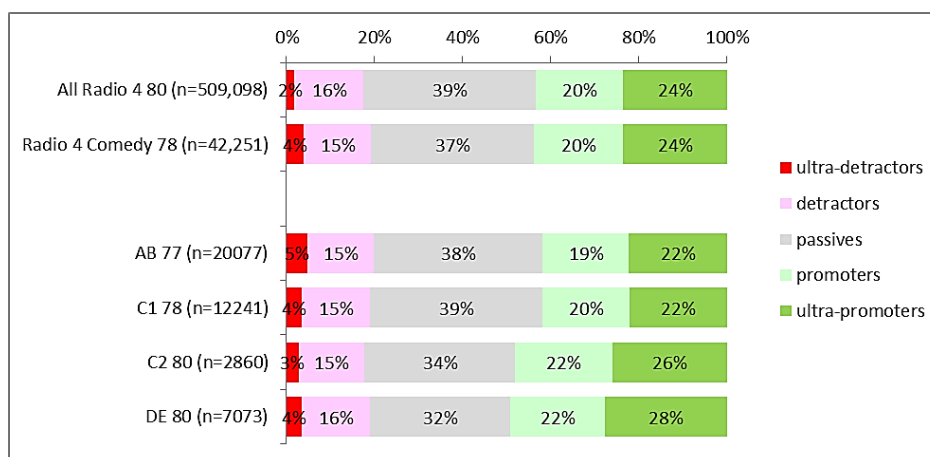
The chart above concurs with earlier findings as the lowest aggregate AI of 77 is given by the ABs compared to 80 from the C2DEs.

However, as when we considered gender, could we examine further factors here? In the earlier results section[425] we saw that ABs were less likely to use the 'Silvey' response pattern, a pattern that results in a far higher mean. Another consideration could be the mix of programmes. Might certain programmes specifically appeal to different groups? We can eliminate the 'mix effect' by looking only at specific programmes.

---

[424] See from p105 and p108.

[425] See p191.

The first example – **Count Arthur** – shows a very clear incremental progression of appreciation as the social grade lowers: fewer detractors and more promoters. This echoes the earlier research cited above. However, this is not true of every programme. For example, **Mark Steel's in Town** gets its highest AI from ABs and indeed the highest mix of 9s and 10s: 57% compared to just 38% for the DEs. Another example that doesn't follow the expected pattern is **The Now Show** where the C2s appear to be far more appreciative than the DEs: a very low level of detractors and a high mix of promoters.

Overall, DEs rate more highly than ABs. This can be partly explained by the fact that DEs tend to use a response pattern that has a higher aggregate AI. However, it is also driven by the content of the programmes. Some specific programmes can be rated more highly by ABs. Thus, we cannot simply say that lower social grades mark more highly as it does vary for some programmes.

9.10.4 With or Without Children

No literature has been found to indicate that presence of children in a household bears any impact on appreciation; however, the Pulse survey has sufficient detail for us to see if there is any effect:

Figure 139 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows– Gromic – Presence of children in the household
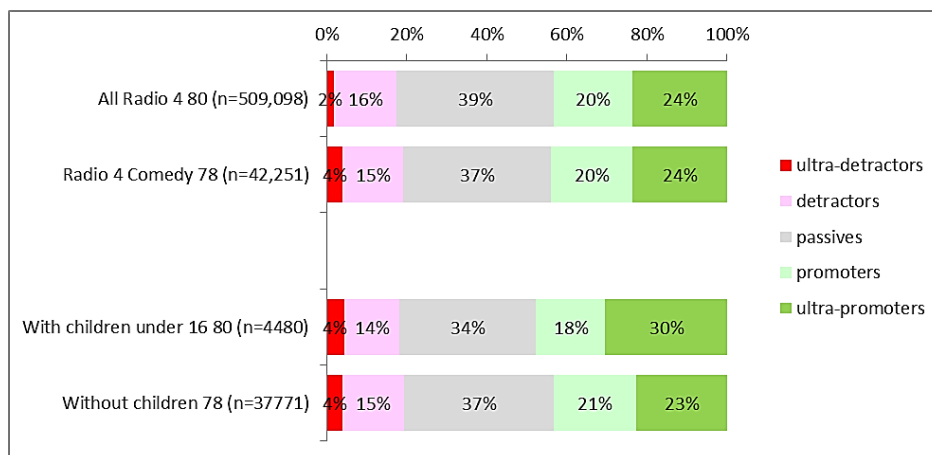
However, it is likely that presence of children is related to the age of the respondent and possibly the gender, so we can look at one age group specifically; the 45-54's have the highest mix of respondents with children in the house.

Figure 140 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows– Gromic – Presence of children in the household – Just 45-54s by gender
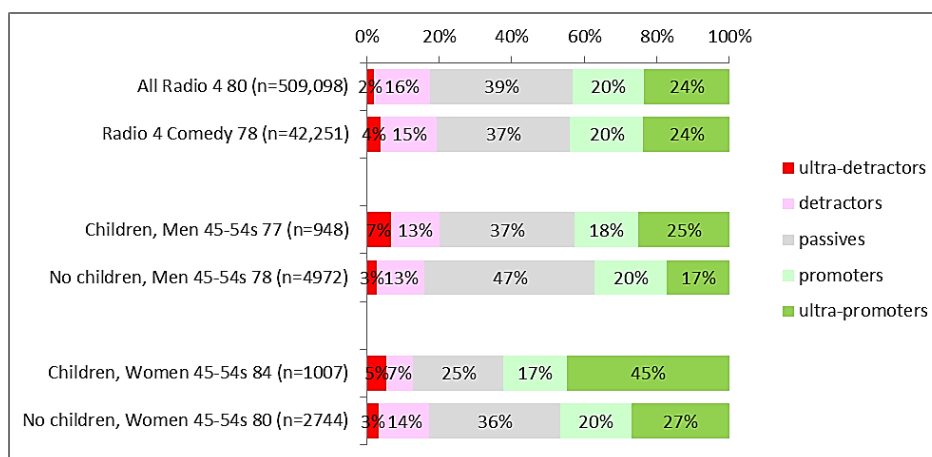
The chart above shows that differences can be seen for both genders if children are present in a household. Men aged 45-54 with children rate comedy programmes slightly lower as a mean (77 vs 78) but do so by giving more polarised scores. They are far more likely to use the extreme scores; 32% of their scores are 1, 2, 9 and 10, whereas men without children use these extremes only 20% of the time. Women aged 45-54 with children are also more prone to extreme scores, with 50% of their scores being 1, 2, 9 and 10, compared to those without at just 30%. However, women's extreme scoring is more generous and results in a higher AI, 84 versus 80.

Any analysis at lower levels, for example, if we were to further segment by series or sub-genre, would result in sample sizes too low to provide valid analysis. At this point we can say that the presence of children in a household appears to provoke more extreme responses to comedy. As discussed in the chapter on appreciation,[426] enjoyment of radio comedy may be enhanced by listening with other people. 'The presence of others, and their mirth behaviours, has long been acknowledged in the social and behavioural science literature as an important contextual factor' (Lieberman et al, 2009, p498). The Pulse data, unfortunately, doesn't include information about whether or not the respondent lives alone or has specifically

---

[426] See p100.

listened to any radio show with another person. But if the respondent has children in the household we do know, at least, that they are not living alone. If we assume that there may be occasions where the respondent is not listening alone and that the presence of the children does in fact enhance the listening, could this go some way to explaining the increased mix of high scores? But what of the lower scores? Perhaps parents could be more sensitive to certain types of content on behalf of their children. 1.9% of all those who answered the question regarding having heard anything offensive in the programme said yes. This increases to 2.7% for respondents who have children. We found in the earlier section[427] that when people find something offensive, their rating of the programme is, on average, much lower. Parents in this dataset were more likely to take offence at Radio 4 comedy shows and mark them lower. For example, might this respondent have given a score of 10 for this programme if she had not been with her children?

> I'm Sorry I Haven't a Clue   15-Jul-12   Female (respondent 40)
> Appreciation score = 9
> 'Jack Dee is very entertaining and a good replacement for Humphrey Lyttlelton. The whole show is very funny. Sexual innuendo – would be just as funny without this. Not comfortable listening to this programme with my children because of this.'

Presence of children in the household appears to provoke more extreme responses to Radio 4 comedy. Respondents are more likely to enjoy the programmes, possibly because they are not listening alone, but they are also more likely to consider content offensive.

9.10.5 Nationality / Regionality

Accents reflecting nationality and regionality could be a factor in comedy appreciation and there are also claims that there is regional variance in sense of humour.[428]

Figure 141 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows / originations – Gromic – Claimed Nationality
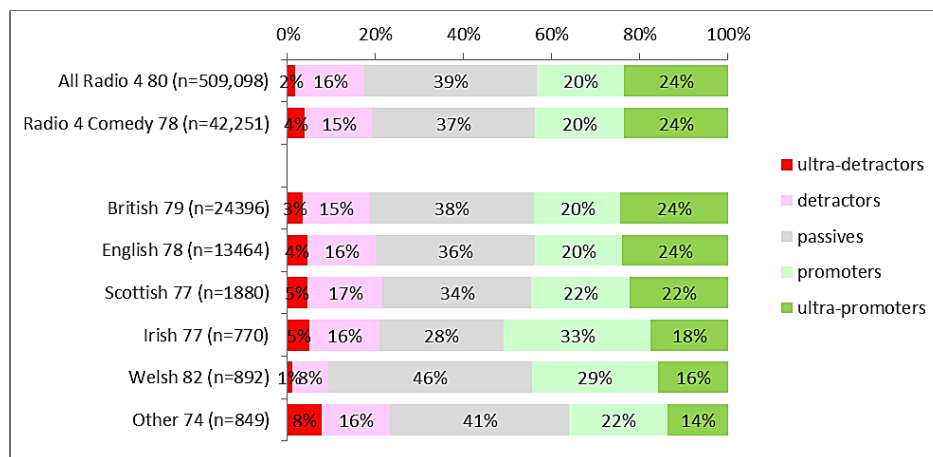
The chart above shows the claimed national mix of respondents. There is little difference between those who identify themselves to be English versus British. There appear to be differences for Irish and Welsh audiences particularly, but, despite reasonable response sizes, the numbers of actual individual respondents are relatively small: 36 and 67 respectively. Ideally, larger sample sizes would be needed to generalise about the response patterns, but there do superficially appear to be differences between the UK's national groups.

A further consideration is the regionality. We discussed how Radio 4 has been viewed as London-centric and as such we might expect London audiences to be more appreciative.[429]

---

[427] See p209.

[428] See from p108.
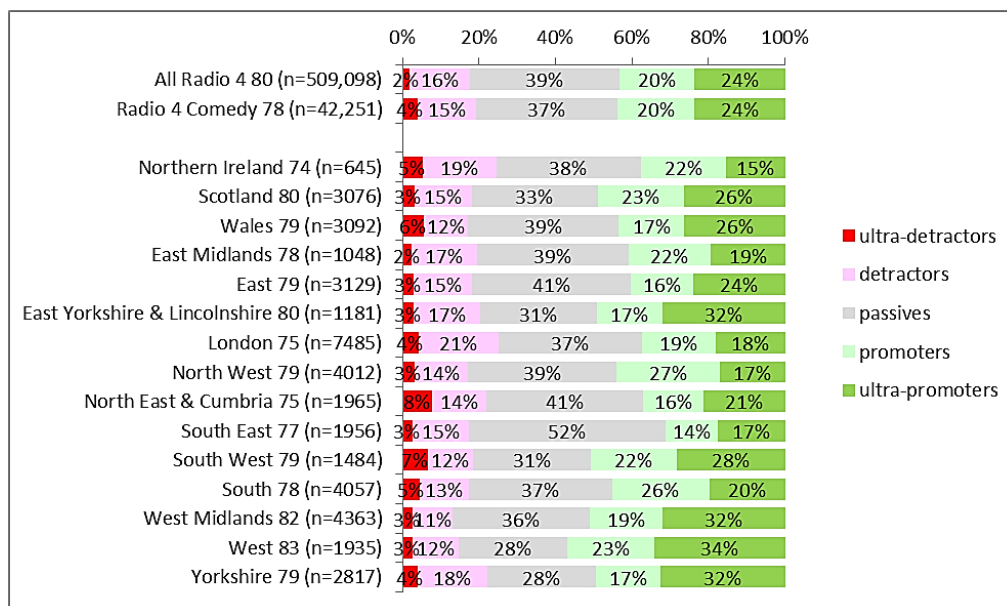
The chart above shows that, with the exception of Northern Ireland, the London region actually gives the lowest AI score for Radio 4 comedy: 75 compared to the aggregate of 78. 25% of the scores given by Londoners are 6 or lower, the biggest proportion of detractors by region and nearly double that of the West Midlands with just 13% of detractor level scores. If, as considered, there are regional variations in humour appreciation, it appears that Radio 4 humour is not London-skewed: quite the opposite. As discussed on *Feedback* (2012b), is it the case that southern listeners are particularly averse to northern voices such as **Count Arthur**? Not all Radio 4 comedies have distinct regional accents and not all regions are necessarily represented, but we can look at whether local accents have a particular appeal. Examples where samples sizes are sufficient are included in the chart below.
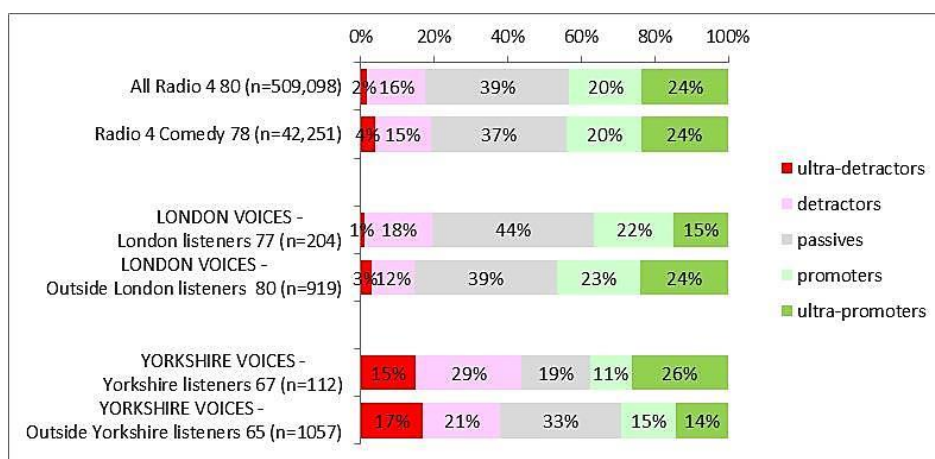
From the data in this chart above, it is not conclusive that regional accents appeal particularly to those from that region. For the shows with London accents, it appears that people outside London are more generous with their appreciation scores to the extent that the aggregate AI is 3 points higher, rather than being lower as

---

[429] See from p108.

[430] Programmes with London voices in this figure are ***Micky Flanagan: What Chance Change?***, ***Arthur Smith's Balham Bash***, ***Can't Tell Nathan Caton Nothing*** and ***Mark Steel's in Town***. Yorkshire voices are represented by ***Count Arthur Strong's Radio Show!*** and ***Tom Wrigglesworth's Open Letters***.

we might have expected. In contrast, Yorkshire accents produce a higher aggregate AI for those listening from Yorkshire at 2 points higher. However, while the AI is higher, it is derived from a very different distribution pattern. Yorkshire listeners appear to be more likely to give extreme scores. Those listening from Yorkshire are more likely to give scores of 6 or below than those outside the region (44% vs 38%). Again, the number of respondents is limited so further data would be needed for greater understanding. It should also be noted that the Yorkshire voices include *Count Arthur,* which has already been shown to be peculiarly polarising.

> Nationality and regionality appear to be independent variable in Radio 4 comedy appreciation but our data does not necessarily reflect the theory that people will be more highly appreciative of comedy that is closer to home.

9.10.6 Ethnicity

Ethnicity can be an area of particular interest for radio, often linked to tastes in music genres (McQuail, 1997, p95). Radio 4 performs poorly for Black and Asian audiences, even when accounting for skew in age and social mix (BBC Executive, 2010, p18-p20). 'Reach to white ABC1s aged 45-64 is around 37%; reach to similar Asian Britons is 26% and to similar Black Britons is 19%' (ibid). While ethnicity has been found to be an independent variable in comedy appreciation,[431] in terms of depiction there are no specific quotas for Radio 4 comedy that have to be fulfilled beyond ensuring a good general representation of the population (Berthoud, 2010).

Figure 144 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy shows – Gromic – Ethnicity
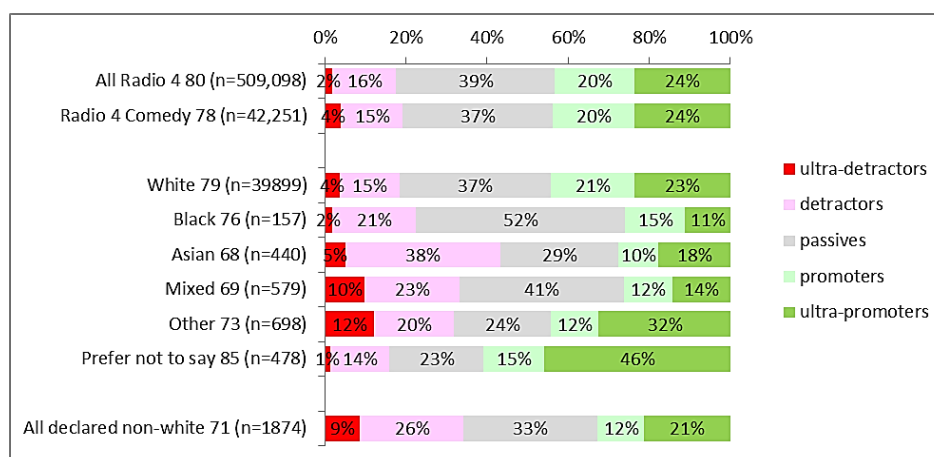
In the chart above, there appear to be variations across the different ethnic groups. White respondents rate the comedies more highly than Black, Asian, Mixed or Other. It is worth noting that Black and Asian Radio 4 audiences tend to be younger than typical Radio 4 listeners (BBC Executive, 2010, p20), and presented above suggested that the youngest audiences gave the lowest appreciation scores for comedy.[432]

Unfortunately, some of the groups are small. For example, 'Black' consists of just 49 respondents (giving 157 responses) so we cannot drill down to see if certain types of comedy are of more interest to the different groups.

---

431 See from p110.

432 See p244.

White respondents rate Radio 4 comedies more highly than non-white listeners, with an aggregate AI of 79 compared to 71. However, the demographic make-ups of the groups differ so it is not clear whether ethnicity alone accounts for this difference.

9.11 Answer to Research Question 3

Although AI's are generally used by the BBC as indicators of quality, and whilst genre/sub-genre are recognised as variables affecting appreciation, this analysis has identified a number of other factors which affect the appreciation scores. The demographic make-up of the audience appears to be a distinct factor in appreciation, for example, women tend to rate more highly than men, albeit not for every comedy programme.[433] The analysis also found that a number of factors relating to the broadcast (number of series, time of broadcast) can also affect appreciation scores.

For this dataset, an attempt to compare the sub-genre performances for all the demographic groups or all the broadcast features would result in sample sizes too small for valid analysis. Instead, we must accept that one of the limitations of the Pulse survey is that only those who listen to Radio 4 will participate in the survey, i.e., the demographic make-up of the respondent group is in itself part of the results. For example, if a particular genre has a particularly high aggregate appreciation, this could be a product of its 'quality' *and* its appeal to certain individuals and, in turn, the distribution patterns of appreciation scores used by those respondents. In this respect, our aggregate appreciation scores for each sub-genre can be compared directly.

Additionally, while there appeared to be effects due to these variables, the properties were not always clear or explainable. For example, when we considered whether a live audience was a factor of appreciation, the results were not conclusive.[434]

However, one broadcast feature did clearly emerge as influential. The analysis demonstrated a strong pattern of increasing appreciation scores as a programme had an increasing number of series. A better way of comparing performance, therefore, might be to examine scores with this element removed,

Figure 145 LEGITIMATE WEIGHTED RESPONSES Radio 4 comedy originations – Gromic – Sub-genres – just first series and pilots (n>99)
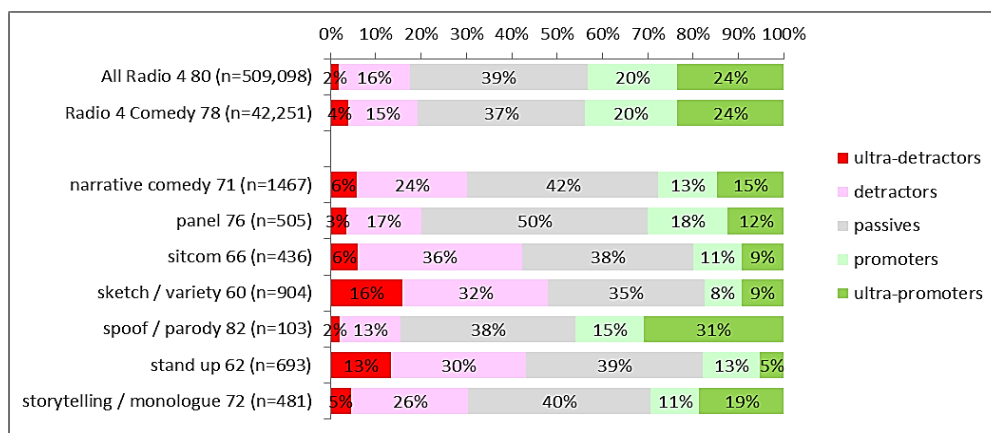
Figure 145 shows that, despite each of the sub-genres being aggregated from only pilots and first series,[435] there *are* observable differences between them. A Commissioning Editor wishing to benchmark a new series would need to do so against a comparable genre *and* series to assess its relative performance. The programmes that premiered in 2012 on Radio 4 did not include all sub-genres identified; impressionism,

---

[433] See p242.

[434] See p239.

[435] I.e., stripping-out the effect of increasing appreciation with increasing number of series.

investigation, and chat/interview are not seen in the chart above. Of those included, spoof/parody is the highest rated with 46% of scores being 9s and 10s and an aggregate AI of 82. However, this data comes from just one title. ***Believe It!***, which has only 103 responses. Taking this as a case study, can we apply some of our earlier findings to investigate why it might have been particularly appealing, attracting relatively high appreciation scores:

9.11.1 Case Study - Performance of *Believe It!*

i Familiarity – Talent

For new series, established talent can contribute to appreciation.[436] This series featured a well-known comedy actor, Richard Wilson.

ii Context – Expectations – Time of Day

This was an 11:30 comedy and it was found that the type of comedy commissioned for this slot tends to attract higher appreciation scores.[437]

iii Demographics

- Gender

64 of the 103 respondents giving appreciation scores for ***Believe It!*** were women, the weighted mix being 68%. The mix of legitimate weighted responses for all Radio 4 comedy for women is 46%, making this programme more female-skewed than is typical for Radio 4 comedy. Since women generally rate more highly than men,[438] this contributes to a higher appreciation score.

- Age

We have found that under-35s rate lower than other age groups.[439] For ***Believe It!***, only 7% of the legitimate responses were from the under-35s compared to 11% for all Radio 4 comedy. This skew will contribute to a higher appreciation.

- Social Grade

Believe It! attracts a weighted DE audience that makes up 22% of its total. This is greater than the 19% we see for all comedy. This is a higher than usual mix and we have found that DEs tend to rate more highly than the other social groups.[440]

---

436 See p233.
437 See p236.
438 See p241.
439 See p244.
440 See p245.

iv Offence

It has been found that where offence is caused, appreciation is lower.[441] Only 1 of the 103 responses said that they found the programme offensive; this is a weighted mix of 1.0% compared to all radio comedy at 1.4%, so a slightly lower rate of offence than is typical. The verbatim response to the offence question, given below, actually indicates that the offence might have been caused by a misunderstanding: the listener believed that the programme was a serious autobiographical programme, when in fact it was a spoof.

> ***Believe It!***   30-May-12   Female (respondent 28)
> Appreciation score = 1
> 'Thank heaven this is over. Why on eraht [*sic*] do we have to pay for the BBC to allow
> their favourite egotistical luvvies to promote their books. Shame on the BBC.'

v Verbatims

With a relatively small number of responses, it is viable to examine verbatim responses to inform our understanding. The show is conceived as a spoof diary, but a number of respondents appeared to believe that it is a true account of Wilson's life, or at least are confused about its veracity. This misapprehension did not detract from the appreciation scores for the show:

> ***Believe It!***   09-May-12   Female 68 (respondent 10638985)
> Appreciation score = 10
> 'A very amusing 'biography' of Richard Wilson. I have, of course, no idea just how much of it was true, but it didn't really matter. The actor is highly talented and the programme could have been a tissue of lies or 100% true – either way, he made me believe every word.'

> ***Believe It!***   09-May-12   Male 55 (respondent 10791185)
> Appreciation score = 9
> 'Very clever and funny. Difficult to believe (thus the title), of what was truth in the storyline, or what was not. Being mentor to George Best was hilarious though.'

> ***Believe It!***   16-May-12   Female 56 (respondent 10793788)
> Appreciation score = 10
> 'Enjoyed hearing Richard Wilson's life story being delivered with humour.'

> ***Believe It!***   16-May-12   Female 56 (respondent 13147965)
> Appreciation score = 10
> 'I love listening to auto biographical [*sic*] items and this was amusing and informative'.

> ***Believe It!***   30-May-12   Male 30 (respondent 10677073)
> Appreciation score = 10
> 'very funny entertaining memoirs like to hear more.'

A number of respondents claim that its appeal comes through good writing and performances:

> ***Believe It!***   09-May-12   Female 70 (respondent 943235)
> Appreciation score = 9
> 'what a treat. beautifully conceived and executed.'

> ***Believe It!***   16-May-12   Female 60 (respondent 2535718)
> Appreciation score = 10
> 'Love it… the actors, the story lines and it is funny.'

> ***Believe It!***   30-May-12   Female 62 (respondent 1164577)
> Appreciation score = 8
> 'very funny, well read and written; enjoyable.'

---

[441]

This example demonstrates the problems with the conventional usage of AIs discussed in Chapter 4. Such use of AI scores would suggest ***Believe It!*** was appreciated highly, based purely on its AI of 82. However an analysis of the distribution of appreciation tells us that, compared to other first series and, indeed, Radio 4 comedy as a whole, ***Believe It!*** received an unusually high proportion of high scores and a low proportion of low scores, and this was reflected in the mean score. The analysis suggests that this pattern of AI scores was influenced by a range of non-'quality' factors relating to the broadcast (the time of day, familiarity of the main performer, avoidance of offensive content) and the demographic composition of the respondents. Verbatim responses (used by the Commissioning Editor on occasion)[442] also contribute to the understanding of *why* people enjoyed the show.

9.11.2 Summary

In summary, taking into account other independent variables, there still appear to be observable differences between appreciation scores for radio comedy sub-genres. However, there are insufficient examples of early series to set benchmarks (some sub-genres are not even represented and some include only one programme). To really understand the performance of a programme, many aspects of appreciation need to be considered.

---

[442] See from p132.

The recommissioning process felt like it lacked momentum. That's not to say that they weren't keen;
I'm sure they can't all have been that thick. But they were dragging their heels in the way that only
a groaning bureaucracy populated by Oxbridge graduates can.

(Partridge, 2011, p162)

This final chapter summarises the findings of this dissertation and then discusses the significance of the results. The limitations of the research are then given, before recommendations are suggested for further research and changes to current BBC practices.

10.1 Summary of Key Findings

This dissertation has investigated in detail the BBC's Pulse appreciation ratings in relation to Radio 4 comedy programmes. The fundamental findings of the research have been that presenting the data as a mean score is problematic, confirming hypothesis 1,[443] and that more detailed analysis of AI data can give a much clearer understanding of the performance of programmes.

In addition to problems with representing the data as mean scores, the analysis also identified problems with the current method of compiling AI data, relating to respondent response validity.[444] The volume of invalid responses was shown to be substantial enough to have an impact on the Radio 4 reported aggregate appreciation score,[445] the measure used and published as a synonym for programme quality.

Removal of the invalid responses from the dataset left only valid responses, which were then analysed to address the issue of the hypothesised non-unimodal distribution of AI scores. The research found evidence that respondents have particular response patterns or *types*,[446] and this finding, if replicable in other datasets, has implications beyond BBC AI scores and could have a bearing on any survey where respondents are required to give a number of subjective responses to any scaled question over a period of time. Further testing would be required to establish that the observed response patterns are not exclusive to the BBC's Pulse survey. The identification of these response types is important as, similarly to response *styles*, they could affect any survey that has scaled responses, and therefore, should be accounted for:

> RS potentially affect all empirical studies that use rating scales and are alternative explanations for the results. However, despite its importance, many researchers do not control for this source of bias. (Van Vaerenbergh and Thomas, 2013, p196)

This research provides original contributions both in terms of the data chosen for analysis and of the conclusions drawn. Considering the data analysis first, previously published academic research on appreciation scores has been limited to television responses. This thesis is the first detailed study of radio appreciation data. Secondly, previous research on appreciation scores for media has not addressed the question of whether the distribution of this data is appropriately represented as a mean score. Thirdly, no previous study has been based on the granularity of data analysed in this BBC radio study. Thus, this is the first study to be able to analyse the effects of variables such as sub-genre, series number and verbatim responses on AI scores. The conclusions drawn from this study suggest a range of alternative ways of using AI data that can inform academic and public discussion of the performance of BBC comedy and also the

---

[443] See p160

[444] See p167.

[445] See p170.

[446] See p178.

professional commissioning, scheduling and production of radio comedy. Evidence of response types found could potentially have implications for a wide range of studies.

Evidence of redundancy in the Pulse survey was found, and if duplicate questions could be eliminated, this might lead to improved survey completion rates[447] and/or widening of the survey's scope.

10.2 Summary of the Findings in Relation to the Hypothesis and Research Questions

Returning to the hypothesis posed on p160, it is now possible to state that the mean appreciation rating is a poor representation of the quality of Radio 4 comedy programmes. The answers to the research questions illustrate this:

RQ1 focused on whether the mean is a good representation of the data. The analysis demonstrated that, after sanitising the data, the appreciation responses were distributed in a fashion that rendered the mean an invalid indicator of central tendency. Firstly, the scores are not distributed in a unimodal, hump-backed distribution[448] and, secondly, the distribution shapes of appreciation responses were shown to vary (for example by programme, by sub-genre[449] and by respondent response pattern type)[450] and therefore their mean scores were not comparable. However, while the mean may not be statistically valid, it was found to be a proxy indication of 'goodness' and is therefore potentially still useful *if* presented alongside information regarding the distribution.[451]

A further finding is that rather than being a simple mean, the aggregate appreciation score is actually the product of a 'mixture distribution'.[452] Analysis of Radio 4 and 4 Extra respondents provided evidence of different patterns of response.[453] For example, the aggregate distribution was bimodal despite 51% of the respondents distributing their scores in a unimodal, humpbacked distribution.

The research demonstrated a key problem with AIs, common to all subjective ratings scales, which is that an assumption is made that the information is given on an interval or ratio scale. The research demonstrated the lack of evidence to support this view. As an attempt to address this problem, a semantic analysis of AI data was conducted to try to indicate the true relationships between the scale points. Analysing the relationship between the use of the word 'love', for example, and the appreciation scores[454] revealed a positive, but not linear, relationship between appreciation and increased likelihood of the word being used in a verbatim. This finding opens up opportunities for further research into whether semantic analysis can indicate true relationships between the scale points, and thus allow more rigorous statistical analysis of this kind of data.

Agreement to a number of other questions from the survey was analysed in relation to appreciation ratings and many were found to correlate extremely highly to the general rating score. This indicates that findings for the basic ratings are also applicable to other Pulse questions. Furthermore, this indicates that a number of the questions may be virtually interchangeable in their results.

---

[447] See from p221.

[448] See p177.

[449] See p251.

[450] See from p180.

[451] See from p201.

[452] See from p179.

[453] See p178.

[454] See p216.

RQ2 inquired whether the genre of comedy might generate unusually polarising responses which would therefore be *particularly* poorly represented by a mean score. The analysis demonstrated that while the genre's aggregate scores did not show it to be the most polarising genre, comedies did constitute both the most hated and most loved shows[455] (when individual programmes were analysed) based on their proportion of extreme low and high scores. This indicates that not only do AIs obscure useful information from those attempting to discern a comedy programme's performance, but also that genre aggregates can conceal what is in fact a wide range of programme responses.

RQ3 went on to ask whether AI's might be affected by factors other than 'quality'. The analysis found that AI scores for Radio 4 comedy do appear to be influenced by a number of factors in addition to the judgement about its actual 'quality'.[456] The most clearly observable independent variables related to programming appear to be:

- The sub-genre of comedy.[457]
- The number of series for which a programme has been running.[458]
- Whether a show includes well-known talent.[459]
- The schedule slot into which it is commissioned.[460]

Furthermore, this research also found evidence of a relationship between response pattern and demographic effects on appreciation response.[461] The skew of the demographic variations supports the findings of other studies that have identified differences due to gender, age and social grade.

10.3 Implications of the Findings

The results of this research have implications for the way the BBC, the BBC Trust, policy makers and academics approach the compilation, processing, presentation and interpretation of programme appreciation ratings.

Firstly, there is the finding concerning illegitimate responses[462] which produce an inflated aggregate AI.[463] While there is evidence that the proportion of illegitimacy was in slight decline,[464] this problem had not been resolved by the end of 2012. A BBC researcher claimed, during 2013, to be aware of the overstatement,[465] but at that stage no research had been undertaken to quantify it, and there had been no communication to programmers to highlight the issue, nor any amendment to the published figures. For example, the published 2012 Radio 4 AI figure of 80.8[466] would have been 79.7 if corrected for illegitimate responses. While this reported *aggregate* was overstated by around 1 point, the analysis showed that this was not consistent at programme-level with AIs affected by up to 5.5 points.[467] Whilst these may seem small differences, such small differences could be cited in support of improvements or decline in BBC performance (for example, Chapman, 2014). Furthermore, there is no reason to believe that this problem is isolated to

---

[455] See p223-224.

[456] See from p251.

[457] See from p226.

[458] See from p227.

[459] See from p233.

[460] See from p236.

[461] See from p190.

[462] See from p167.

[463] See p170.

[464] See p169.

[465] Private correspondence.

[466] This is the aggregate figure across 2012, published across four quarters. See figures 41 and 42 from p169.

[467] See p171.

Radio 4 and 4 Extra, nor to 2012, meaning that all BBC published aggregate AIs *could* be overstated. If the error were consistent it might not pose too large a problem as figures may still allow comparison, but the analysis suggested that the problem appeared to vary in size, at topline level, depending on time of the year (quarter) measured and, in size and direction at programme level. Incorrect representation of programme appreciation could contribute to an incorrect evaluation of a programme's 'quality' and contribute to sub-optimal decisions in commissioning, scheduling and production.

Secondly, individuals appeared to give appreciation ratings couched in their own response patterns.[468] These patterns have an effect on aggregate scores[469] and are skewed by demographic groups.[470] This implies that a respondent make-up, skewed toward a certain demographic, could have an effect upon the mix of response patterns and, consequently an impact upon the resulting appreciation ratings. There is no reason to think that these patterns are limited to radio appreciation ratings; this phenomenon may occur in any panel study with a numerical scale of subjective evaluation, implying that research of this kind should attempt to identify response patterns and their effect before attempting to evaluate the responses on a broader scale.

Thirdly, the data indicated that appreciation score distributions are not unimodal and humpbacked, and an AI distribution is actually derived from a variety of distribution shapes. This supports H1 in suggesting that the expression (as well as interpretation and use) of AIs as a central tendency *alone* – as they are derived from a mean score – is a flawed representation of the data. One implication of this is that programme level AI scores obscure valuable detail about programmes that are particularly liked or disliked. This may be especially significant in using AIs to make statements about the comedy genre because results of RQ2 found that Radio 4 comedy programmes can indeed polarise audience response.[471] The consequence of this is that mean AIs alone are a particularly poor primary indicator for evaluating the performance of comedy programming.

There was additionally a (possibly) minor issue with Pulse responses not accurately reflecting audience demographic proportions as measured by RAJAR, as Pulse weightings are not done at station level.[472]

10.4 Significance of the Findings

This study has contributed substantially to understanding of the audience appreciation data for Radio 4 comedy. In finding evidence of respondent response patterns for subjective evaluation on a numerical scale[473] it prompts the need for a suggested a programme of further research to apply this analysis to similar data sets beyond radio appreciation.[474] If the findings of this study are repeated in other data sets, this could have wide-ranging implications for understanding survey response data across a range of audiences industries and subject areas.

While television appreciation has been the subject of some, relatively limited, academic and industry research, no equivalent study on radio appreciation has previously been conducted. One of the reasons for the neglect of such a study may relate to the difficulties of obtaining access to the data. Another has been the general tendency for television to dominate the attention of academics, policy-makers and commentators.

---

[468] See p178.

[469] See p190.

[470] See from p190.

[471] See p225.

[472] See p175.

[473] See p179.

[474] See from p268.

However, Radio 4 comedy is an area worthy of study; it is listened to by over five and a half million people every week in the UK alone, at a cost of over £5 million per year to the licence fee payer. Radio has also functioned as a pilot platform for television comedy; this is important as it allows a relatively cheap route to test new talent and new formats. If comedy programme evaluation can be improved as a consequence of this research, and in turn, decision-making improved, there could be an enhancement for many Radio 4 listeners.

Finding that the published measure of quality for the station was overstated for the measured period is of noteworthy importance.[475]

The following sections are segmented to highlight the research's significance relating to: 1) station level quality indication 2) programme level performance and 3) wider implications.

10.4.1 Significance 1 – Using the Aggregate AI as a Measure of the Quality of Radio 4 as a Station

In 2010 the BBC announced that it would begin to publish 'quality' information at station/channel level.[476] The aggregate AI has been the primary 'quality measure' in these publications since quarter 1, 2011. This research has found that the aggregate score for Radio 4 had been compromised by illegitimate appreciation ratings, leading to an overstatement of Radio 4's aggregate by around 1 point.[477] As this published figure is being used as the BBC's principal indicator of station quality, the data illegitimacy problems identified in this study are clearly issues that the BBC and its stakeholders should wish to correct.

There is no reason to think that the Radio 4 data is unique in this respect, suggesting that similar errors may affect the BBC's key quality measures for all of its radio stations and television channels. If the bias in the data was consistent then, one might argue, the AI scores, while not an absolute measure, might still be useful as a *comparative* measure of appreciation. However, the analysis suggests that the proportion of illegitimate scores decreased throughout 2012. This means that the degree of error in the aggregate AIs varies across time and that AI figures are not comparable across different quarters, even for the same station or channel. This affects interpretation of the data. For example, if we were to see a change in aggregate AI from one quarter to another, it would not be possible to say whether this was due to a change in overall 'quality' or a change in the number of illegitimate scores. We can see the published aggregate AIs for Radio 4 in Q2 2012 and Q3 2012 are both at 80.7 (see table on p169). This could be interpreted to suggest that the overall programme quality across the quarters was therefore static. However, stripping out illegitimate responses reduces the aggregates to 79.5 and 79.9 respectively; legitimate responses actually show an increase from one quarter to the next of 0.4. Were this change transparent, it could have inspired further investigation as to the source of the increase, leading to a better understanding of the audience's preferences.

Even when illegitimate scores are eliminated, the analysis in this study calls into question the usefulness of the AIs as a simple measure of the quality of individual stations. The BBC document (figure below) that publishes these figures invites comparison between stations, the implication being that one score can be compared to another. This table suggests that Radio 4 has a higher quality performance than Radio 5 live and a lower performance than Radio 2 and Radio 3, but performs better than the average for all BBC radio. However, despite the implications of this table, the BBC claims that comparisons across genres are not meaningful.[478]

---

[475] See from p171.

[476] See from p88.

[477] See p170.

[478] See from p126.

BBC Radio Quality Measures	
Appreciation Index (AI)	
Service 16+	**AI out of 100**
All BBC Radio	80.3
BBC Radio 1	76.9
BBC Radio 2	81.5
BBC Radio 3	80.9
BBC Radio 4	80.7
Radio 5 Live	77.7
Radio 5 Live Sports Extra	81.3
BBC 1Xtra	78.6
BBC 6 Music	84.0
BBC Radio 4 Extra	80.0
BBC Asian Network	73.7
BBC Local Radio in England (England TSA)	81.3
BBC Radio Scotland (Scotland TSA)	79.0
BBC Radio Wales (Wales TSA)	81.6
BBC Radop Cymru (Cymru TSA)	78.3
BBC Radio Ulster (Ulster TSA)	81.6

*Source: Pulse, panel of 20,000 UK Adults, 16+ by GfK for the BBC, October - December 2012. * BBC Radio Ulster includes Radio Foyle and BBC Local Radio in England Excludes Radio Jersey & Radio Guernsey*

Radio 4 (compared to the other BBC national stations) comprises a particularly wide range of different genres and sub-genres within its programming. This research has indicated that even at sub-genre level there were observable differences in appreciation. The aggregate AI for the station represents the average across these different, arguably incomparable types of programme. Thus, the aggregate AI for Radio 4 is attempting to give the average of apples, oranges and bananas, meaning that the resultant aggregate is, to some extent, meaningless. It follows that instead of expressing overall quality, the aggregate AI may merely be measuring the proportional makeup of different genres broadcast on the station.[479] As we found that different programme genres garnered varying aggregate means,[480] this could be manipulated to increase the station average. This has implications for how the BBC or policy makers might use AIs to try to improve quality. It is theoretically possible for example, that the Controller of Radio 4 could decide to replace the channel's religious programmes (aggregate AI 76) with consumer programmes (aggregate AI 82). Superficially, this would result in an increase in the total score and the inference could be that 'the station has increased its quality'. However, it is likely that the very range of programming on Radio 4 is part of what appeals to its audience.[481] To take an extreme example - if Radio 4 broadcast solely consumer programmes, the individual shows may still be as good and attract a high AI, but this would not necessarily reflect an improvement in station quality as a whole.

Similarly, it would be possible for Radio 4 to improve its AI scores by attracting a higher proportion of a specific audience demographic. While the strategy for the station is to attract younger audiences, this study has demonstrated that, for comedy at least, older women of a lower socio-economic status tend to rate programmes more highly.[482] An audience including a higher proportion of this demographic group could therefore increase the station's mean appreciation score. To a certain extent Radio 4's AI scores may actually reflect this tendency, since Radio 4's share of 35-54 ABC1s (the so-called replenisher audience) has dropped from 19.7% at the beginning of 1999 to 11.8% at the end of 2013 (RAJAR, Q4 2013).

We have found at station level and lower levels alike that appreciation scores are not suitable for parametric analysis.

---

[479] See from p126.
[480] See p222.
[481] See from p33.
[482] See from p190.

10.4.2 Significance 2 – Using the AI as the Measure of Radio Comedy Programme Quality

The inclusion of illegitimate scores in the programme-level detail has resulted in poor representation of Radio 4 comedies. Examples have been examined that illustrate how a Radio 4 comedy's appreciation score could be inflated by as much as 5.5 points, or even decrease, due to illegitimate scores.[483] Because the magnitude and direction of the errors are not consistent, a simple correction of the data is not possible. These errors have practical consequences, as AIs are the primary source of objective programme evaluation for Radio 4's Comedy Commissioning Editor.[484] A popular programme might not be recommissioned if its AI was understated. The historical sections of this study have demonstrated that such decisions could have consequences for the genre. The first ever episodes of Radio 4's most long-lived comedy formats – *Just a Minute* and *I'm Sorry I Haven't a Clue* – both attracted relatively low AI scores (57 and 49 respectively),[485] and had they had been subject to an error reducing them further then, they might never have been recommissioned.

A consideration of legitimate scores found that appreciation ratings are not distributed in a unimodal humpback fashion.[486] This means that, strictly, programme AIs are not statistically correct. More robust measurements, suitable for non-parametric analysis, should be used. Most significantly, visibility of the extreme appreciation scores is required, as we have seen that comedy programmes can be particularly polarising.[487] It is accepted that comedy needs to attract dissatisfaction from some factions of the listenership in order to appeal to those who may consider themselves comedy connoisseurs (the 'cult' audience).[488] This aspect of the data should be made apparent. Where a mean *alone* is used, there is no transparency as to whether the ratings are constituted from similar scores or are the average of extremes. For example, take the first two programmes in the figure on p223, which shows the Radio 4 comedies with the greatest proportion of scores of 1s and 2s; the AIs for *Count Arthur Strong's Radio Show* and *Andrew Lawrence: How Did We End Up Like This?* are similar at 61 and 59 respectively. What the means do not tell us is that *Count Arthur* garners scores of 10 from 15% of the responses versus just 3% for *Andrew Lawrence*. For programmers to understand truly both the strength of the audience's feeling and its direction, access to this aspect of the data is paramount.

We have found that appreciation can vary due to a number of variables for slot,[489] for example sub-genre[490] and series number.[491] To further understand performance at programme-level, benchmarking of the expected performances, taking these variables into account, needs to be developed. The first series of a late-night sketch show with all-new talent should not be expected to achieve appreciation ratings in the range of those garnered by an 18:30 panel show that has been running for 40 years.

Likewise, as some types of comedy may appeal to different demographic groups to varying extents,[492] an understanding of the audience composition is important when evaluating programme performance. The AIs of two equivalent series might not be comparable if one of the series was to have a

---

[483] See p171.

[484] See p88.

[485] See from p126 and p73.

[486] E.g., see from p178.

[487] See p225.

[488] See from p48.

[489] See p237.

[490] See p226.

[491] See p228.

[492] See from p108.

particular appeal to women for example, as women tend to give higher appreciation scores.[493] Examining comparisons of programme performance across the genders,[494] it was found that weighted proportions of male and female respondents varied across programmes. Thus, an example comparison can be made between *Museum of Curiosity*, 49% of whose respondents were women, and *Count Arthur*, whose respondent audience mix was 36% women. This difference in gender skew can affect overall programme appreciation.

A principal objective of this research is related to the evaluation of quality. While the BBC researchers have found that appreciation ratings are analogous to quality ratings,[495] we have discussed how this may not always be the case[496] and that respondents can find it hard to identify precisely what they mean when they respond that they 'like' a programme.[497] For a news or current affairs programme, for example, they might consider it to be high quality because they agree with it. In contrast, for a Radio 4 comedy, they might consider it high quality only if they found it funny. Respondents' ratings for programmes as a whole correlated extremely highly with their quality rating (0.98).[498] This indicates that the measures could be analogous for Radio 4 comedies and that the respondents do indeed find it hard (or unnecessary) to separate the two.

However, examination of the data can reveal that higher appreciation ratings *do not* necessarily relate purely to programme quality. Regarding, in particular, the observed positive correlation between appreciation ratings and the number of series for which a programme has been running, a trend may be apparent. *Clue*, for example, achieved an AI of 49[499] (albeit measured on a different scale) in its first ever episode, but in 2012 its legitimate AI score was 88. It seems unlikely that this increase is a measure of increased *quality*, since the format of the show, the performers and the style of the jokes are largely unchanged over its lifecycle. Rather, it would seem that the respondents are interchanging appreciation and quality for aspects such as familiarity and enjoyment.

10.4.3 Significance 3 – Wider Implications

The methods used for this research may be applied to any panel study where respondents give subjective evaluations on a scale over a period of time.

i Appreciation Data is From an Ordinal Scale

We have found that BBC practitioners[500] and academic researchers[501] treat appreciation ratings as if they were on an interval or ratio scale by presenting the figures as a mean without reference to the distribution shape of the data. It is therefore reasonable to assume that researchers in other areas are treating similar data sets as if the data is more robust than it actually is.[502] If so, the implication of this research is that they too may be erroneously subjecting their data to parametric analysis. This is implicitly recognised by some new media platforms. Amazon and IMDB, while they present their consumer ratings as a mean, also allow users to see the distribution of the scores.[503] YouTube has moved away from using a scale altogether, using simply

---

[493] See from p190.

[494] See p242.

[495] See from p127.

[496] See from p127.

[497] See from p132.

[498] See p203.

[499] See p73.

[500] See p116 and p260.

[501] See from p143 and from p159.

[502] See from p142.

[503] See from p206.

a dichotomous thumbs-up, thumbs-down system. In this context, the BBC's continued use of the mean as the sole numerical presentation of the appreciation data risks at best losing valuable information about the nature of the audience's appreciation, and at worst misinforming decision-makers.

ii Response Patterns Affect Results

This research identified varying response patterns from Pulse panel members.[504] Respondents appear to develop habits – self-prescribed methods of response – which appear to be comparative rather than absolute. This behaviour could relate to Tourangeau and Rasinski's theory that posits that respondents go through a series of processes when attempting to answer an attitude question (1988, p299)[505] – one of the processes is comparing the subject to be rated to others that have previously been considered. We could plausibly hypothesise that the *primary* driver for any individual score they give is not their absolute response to the subject, but rather their habitual scoring pattern. We might further argue that a researcher should control for these habitual response patterns in the way that they interpret the data. For example, this study found respondents who would never rate any programme higher than a self-imposed limit, despite submitting glowing verbatim responses for programmes at the top of that limit.[506] So if a respondent with a 'Silvey' pattern and an upper limit of 8 gives a programme a rating of '8', it is arguable that that score should be interpreted as being equal to that of another respondent who gives the same programme a score of 10, if that respondent also gives 95% of their responses as 10s.

It is reasonable to assume that this phenomenon is not restricted to Radio 4 and 4 Extra programme appreciation data, and that respondents would carry these response patterns across the whole of the Pulse survey covering BBC radio and television stations. As the appreciation scores were found to be analogous with a number of other variables, such as 'quality', 'original and different' and 'the kind of programme I would talk to other people about',[507] we can hypothesise that these questions may also be subject to response patterns.

The implications of this finding go beyond BBC audience data to the broad area of qualitative research (both academic and industrial) since, based on the patterns identified in this study, we might hypothesise that this phenomenon applies to *any* panel survey with scaled responses. This is clearly an important area for further research. While this study observed these patterns on a 10-point numerical scale, further research is needed to know if similar patterns could be found with other scale dimensions. It is possible that analysis of Likert-type scales might demonstrate this effect, although this seems less likely as the scale points may be less open to interpretation by the respondents if there are verbal descriptions applicable to each point. However, as online resources become increasingly widespread and online surveys become easier to administer and use over an extended time period, it may be that many areas of academic and industrial research are interpreting data without due recognition of the implications of panel member response patterns.

iii Response Patterns Relate to Demographic Differences

Research into television and radio appreciation has generally found that certain groups give higher ratings than others:[508] older people, women and those of a lower socio-economic background. This study found that this also holds true for BBC Radio 4 comedy.[509] While an explanation of this phenomenon may partly be

---

[504] See p179.

[505] See from p95.

[506] See p183.

[507] E.g., see p207.

[508] See from p105.

[509] See from p241.

attributable to the 'mix effect',[510] this research found that these tendencies also related to response patterns. For example, women rate more highly on average because they are more likely to have a response pattern that results in a higher average rating.[511] There appears to be no previous literature that has identified this link. If this connection were to be found in other types of panel study, it could go some way to explain demographic differences in results. For example, if we were to conduct a political panel survey that finds that women give higher 'appreciation' ratings for a political party, politician or policy, we might interpret this data as evidence that the party, politician or policy was successful in appealing to women. However, this study suggests that such data could simply reflect a greater propensity among women to use the 'Silvey' response pattern (as in the Radio 4 data set), which gives a higher mean score on average.

10.5 Limitations of the Study

A number of caveats must be noted regarding this study:

10.5.1 Assumptions

- While the study identified a number of illegitimate responses in the data,[512] the broad assumption was that the respondents were generally attempting to be truthful in their answers. The Pulse survey allows participants to select the programmes that they wish to rate, but there is no guarantee that they have even heard those shows. The analysis has therefore had to assume that, on the whole, respondents are acting in good faith and have no ulterior agenda.
- The data was supplied by GfK via the BBC research department. The assumption was made that there were no errors in the data files. A check of our topline figure compared to the BBC's reported figure indicated that, while there was a variation, it was minor and acceptable.

10.5.2 Delimitations

- The data used for the detailed analysis pertained purely to BBC Radio 4. Radio 4 was chosen in particular as it is listened to widely in the UK and consists of a number of speech genres. It is also of a relatively high budgetary value for a radio station and therefore of particular significance to licence payers. Earlier studies had focused on television, and it was felt that radio was also worthy of consideration as it remains a highly penetrative medium.
- The author worked full time for Radio 4 throughout the duration of the research, and in doing so was able to access people and information that would have been unavailable or difficult to access for outside agencies.
- As previous research had indicated that it is not possible to compare different genres,[513] one genre was chosen for the study. There is a recognition that comedy may be particularly polarising in terms of response[514] so it was selected for this research. Additionally, comedy is of particular interest from a financial point of view as it is as it is relatively expensive to produce,[515] has the potential to deliver commercial gains, and can be utilised as an economical form of piloting for television.[516]
- The data was limited to the calendar year 2012. This was the most recent full calendar year at the time the analysis was conducted.

---

[510] See from p105.
[511] See p191.
[512] See p167.
[513] See from p126.
[514] See from p48.
[515] See from p44.
[516] See from p45.

-   While the data included a number of variables, appreciation ratings were the main focus as this is the primary measure used at the BBC[517] and the one most comparable with earlier academic research.[518]

10.5.3 Limitations

-   The data for Radio 4 comedy was based on a survey from a sample group and as such is limited in the same ways as any survey data. While the number of responses was high at around 42,000, once analysis began to drill down to specific demographics or programme titles, for some groups the sample sizes became too small for valid analysis.[519]

-   Arguably, the dataset gave the least amount of information about the programmes that were of the greatest interest. Higher risk, newer shows tended to be placed in late night slots with smaller listenership. This meant fewer Pulse responses and less data to analyse for these examples.[520]

-   As the detailed analysis was limited to Radio 4 comedy for a specific year, there are limitations on the generalisability of the findings. Comedy was found to be divisive in some respects,[521] and may not therefore be a typical genre to consider. However, the distribution shape across all of the responses for Radio 4 and 4 Extra were very similar to those solely for comedy[522] (a bimodal distribution peaking at 8 and 10), indicating that there may be commonalities. This data does not allow consideration of whether distribution for other radio stations and television appreciation scores are also composed of mixture distributions.

-   As with any survey, there could be issues with the panel insofar as those who participate in such research may not be representative of the population as a whole.[523] This study did not seek to identify whether this was the case. However, it was found that the demographic mix of panel members responding to Radio 4 did not exactly reflect the proportion that we might expect if compared to the RAJAR figures.[524] As the Pulse panel is balanced for the BBC as a whole, rather than specifically for Radio 4, it seems that this may be a minor source of error in the data, slightly limiting the extent to which it is representative of Radio 4 listeners.

-   A further factor in the validity of the sample sizes of responses is that of under-reporting bias.[525] Each respondent has to fill in the survey for 10 days each month to earn their incentive;[526] however, the chosen days are at the respondent's whim. It was found that not all programmes were equally represented by responses.[527] This bias could be due to a conscious effort to give a rating for a programme that respondents were particularly stimulated by – either positively or negatively – or it could be a subconscious influence. As respondents don't have to rate everything that they hear, the aggregate scores could be skewed toward polarised responses if we are to consider the theory of 'extreme response bias'.[528]

---

[517] See from p87.

[518] See from p10.

[519] E.g., see p250.

[520] E.g., see p236.

[521] See p225.

[522] See from p177.

[523] See from p119.

[524] See p174.

[525] See p151.

[526] See from p115.

[527] See p217.

[528] See p154.

- There is additionally the aspect of 'consumption bias'.[529] Only the people who listen to Radio 4 comedy are rating Radio 4 comedy. This means that our results are not absolute as they do not include evaluations from non-listeners. Since the recruitment of new listeners is an ongoing challenge for Radio 4,[530] understanding what non-listeners think about its content could be an important factor for the BBC.

- The data was limited to a one-year timeframe partly due to the limitations of the analytical hardware and software. The 665,000 rows of data made Excel extremely slow and prone to crashing, so any additional data would have been unworkable with the systems available at the time.

- Difficulties negotiating access to the data meant that requesting data covering a time period longer than one year may have compromised the chance of getting any access at all. However, it turned out that the option to look over a *significant* time period in order to do a meaningful longitudinal study was never available: a request for data for specific programmes from earlier years elicited the response that respondent-level data was not available from earlier than August 2009.

- The analysis was carried out on the Pulse rating scale of 1-10. Whether the same patterns might have been seen on a different numerical scale, or even a Likert-style scale, is not covered in this analysis.

- While the majority of the data analysis was, to some extent, automated in Excel – such as summarising data using pivot tables, using chart templates to present the results and using the vlookup function for amalgamating different data sources – there were many opportunities for human error. For example, each respondent's responses were allocated a distribution shape. Beyond viewing each on a chart that was automated in its set-up, this was done manually. The process of allotting a category to each was partly subjective. While back-checks were made, it is also possible that this process could have given rise to errors.

- Furthermore, the consideration of which respondents were giving illegitimate scores involved an element of subjectivity. While initial criteria was based on those claimed to be used by GfK (i.e., highlighting respondents rating programmes same score 5 or more times in a row),[531] further information in the dataset allowed more subtle, albeit subjective, interpretations. For example, even when a respondent appeared to give an unusual response pattern, if the verbatim responses reflected the scoring, the ratings were considered to be valid. In some cases, however, the identification of illegitimacy may be considered to have erred on the conservative side; when every score given by a respondent was exactly the same, they were *all* considered illegitimate. However, it could be the case that that the respondent really did feel that every one of the thousands of programmes that they had heard *was* worthy of a score of 10. The factor of illegitimacy could be interpreted in a number of ways.

- A further area of subjectivity was the interpretation of each Radio 4 comedy programme's sub-genre. In determining this aspect, people in senior positions in BBC radio comedy were consulted in order to include expert opinion.[532] However, there are a number of ways that the sub-genres could be categorised[533] so these findings are specific to this particular segmentation.

- There are a number of variables that were not included in the Pulse survey but may have been of particular use in this study:

---

[529] See p152.

[530] See p35.

[531] Van Meurs, 2009b, slide 6.

[532] See from p164.

[533] See from p39.

- Participants are not asked whether they are listening in groups or alone. This is a key factor in humour studies[534] and we might have expected to find differences in this respect if information on this independent variable had been available.

- The mood of the listener at the point of hearing the programme could have been useful, were it even possible to effectively measure this aspect effectively.[535]

- Mood could in turn be affected by the content immediately ahead of the show being rated as 'excitation transfer theory' has been found to affect appreciation.[536] However, we did not have the scope to evaluate all of the interstitial Radio 4 content adjacent to the comedy programmes in order to understand whether this was a factor.

- A further desirable factor could have been knowing what the listener was doing at the time of hearing the programme. As radio can be seen as a secondary or even tertiary medium, knowing the primary activity of the respondent may have been informative.[537]

- The Pulse survey asks participants to 'rate' programmes but we have discussed how appreciation can be a nebulous concept.[538] If, as has been mooted, appreciation is related to utility,[539] perhaps questions such as 'how funny' or 'how entertaining was this programme' would have been useful inclusions in comparison.

- The dataset did not include any information about how long respondents had been participating in the survey. It may have been useful to compare this to the response patterns, as panel conditioning could be a factor.[540]

- Respondents' ratings are a subjective measure, but it would have been useful to compare these scores against an objective measure. Measuring physiological responses would have been outside the scope of this research and, even if it were possible to obtain a large enough sample size, it is recognised as a limitation of humour research that experimental evaluations can give results that are not typical of real life.[541] However, more realistic to acquire may have been behavioural measures which can also be interpreted as objective measures. For example, since people choose what they consume, the more people who listen to a Radio 4 comedy the *better* the programme can be deemed to be. Television BARB figures allow researchers to compare appreciation ratings with viewing figures, enabling correlation analysis. This type of information would have been useful for our research, but RAJAR does not release data on a level of granularity that would have allowed this.

- The key finding of this research was that many respondents gave ratings in distinct response patterns. 94% of those analysed could be categorised into one of three patterns: unimodal, 'Silvey', or bimodal.[542] What this research does not illustrate is why these patterns occur; without further information and analysis we cannot even speculate as to whether these patterns are conscious or subconscious phenomena.

---

[534] See from p100.

[535] See from p33 and p102.

[536] See from p103.

[537] See from p78.

[538] See from p132.

[539] See from p127.

[540] See from p119.

[541] See from p100.

[542] See p179.

10.6 Recommendations for Further Research

This research has revealed many questions in need of further investigation. Supplementary studies might explore some of the following areas – segmented into three sections: 1) BBC research using just existing data, 2) BBC research considering expanding the data collection and 3) considering wider implications.

10.6.1 Further Research 1 – Related to BBC Quality Measurement – Using Existing Data

- It has been seen that throughout 2012 the aggregate of both Radio 4 and Radio 4 comedy appreciation responses are bimodally distributed.[543] Is this the same for all BBC radio stations, and indeed BBC television stations? If the data could be obtained, this would be a simple study to carry out. If it is the case that, as we have found with Radio 4 comedy, the distributions are not unimodal and humpbacked, then they should not be expressed as solely a mean.

- Another observation is an extremely high correlation between the proportion of agreement to a number of Pulse questions for Radio 4 comedies (such as: 'This programme felt fresh and new') and the high appreciation-ratings.[544] Further analysis should be done to ascertain whether this is the case across all genres and stations. There is no reason to think that comedy would be unique in this respect.

- Data analysis could be expanded to a wider time period than one year, as we found that some areas returned sample sizes that were too small to analyse when we drilled down. Being able to obtain enough data to establish benchmarks for new series and build a view of changes over a number of series would be of great help to programmers.

- Response patterns were observed from the amalgamation of appreciation responses from Radio 4 and Radio 4 Extra.[545] Do these patterns extend to other stations and to television? Again, the data is collected as part of the ongoing survey, so it would be necessary to obtain response data at the lowest level, then code and analyse the information.

- Finding that respondents had a variety of response patterns[546] implies that what constitutes a 'very good' score for one response pattern might not be so for another.[547] One approach to this problem could be to determine where values such as 'good/average/bad' would fall on the 1-10 scale for each identified distribution pattern. In the example discussed earlier, a score of 8 from a respondent with an upper limit of 8 could be considered comparable to a score of 10 from a respondent whose upper limit is 10; in each of these respective cases, 8 and 10 represent each respondent's maximum appreciation and could therefore be given the maximum *transformed* rating. This might yield more accurate results, as it would exclude any affect from demographic skews in response pattern. Further consideration would be needed to understand if any other effects might be seen that could mitigate the efficacy of this methodology.

- GfK possess information about how long each respondent had been involved with the Pulse survey. Comparing the response pattern to the duration of participation might indicate a relationship between these two factors, possibly related to panel conditioning.[548]

- This research found the Pulse verbatim responses to be of use in informing aspects of the quantitative analysis.[549] However, the responses themselves could be further utilised. The sheer volume of these

---

[543] See p178.

[544] See p206.

[545] E.g., see p204, p205, p206 and p207.

[546] See p179.

[547] See p263.

[548] See p119.

open-ended responses makes it difficult to imagine examining them one by one, but research areas such as sentiment analysis could further inform researchers as to programme quality. The basic analysis of the word 'love' indicated that there is a positive, but not linear, relationship between this word and higher appreciation scores.[550] More in-depth analysis of the responses – widened beyond Radio 4 comedy – may begin to tell us the exact nature of how respondents interpret the scale points and whether the points are more than merely ordinal. This data could be used to create a weighting to allocate to the scores, allowing parametric analysis[551] and possibly giving a 'construct validity' (Gunter, 2012, p240) to the appreciation scores.

10.6.2 Further Research 2 – Related to BBC Quality Measurement – Expanding Data Collection

The Pulse survey poses a limited number of questions that relate to appreciation. To develop our ability to evaluate Radio 4 comedy, in particular, further questions could be added to the Pulse survey and/or additional behavioural measures could be taken into account:

- While the questions in the Pulse survey are generally fixed, there is scope to include additional questions by exception. Such questions could be useful in further identifying comedy appreciation. In particular, queries such as 'How funny did you find this programme?' and 'Were you alone or listening with others?'[552]

- Trialling the collection of appreciation data through a continuous (VAS) scale, rather than the current numerical scale, could contribute to an increased chance of appreciation scores being allocated on an interval level.[553]

- In theory, the number of Pulse responses for a programme should correlate to the audience size,[554] however, this was not found to be true across different genres,[555] and lack of granularity in RAJAR data prevented analysis determining whether this correlation was present within comedy sub-genres.[556] If further analysis could be carried out to validate the relationship between response rate and audience size, the response rate could then be used as an indicator of listening and hence audience behaviour. Research for television appreciation found that for entertainment programmes in particular there was a positive correlation between audience size and appreciation.[557] This data would allow would allow radio programmers to garner additional behavioural information at episode level – not possible with current RAJAR reporting.

- Rate of on-demand online listening has the potential to be a further behavioural measure, and could feed into appreciation analysis once this manner of consumption is better understood. This research did not find a direct relationship between online listening and appreciation for Radio 4 comedies.[558] However, as non-linear consumption continues to grow, further research in this area will be important and an understanding of the relationship should be established. Currently, the BBC possesses no data on the demographics or specifics of the situation in which online listening occurs, but it is conceivable

---

[549] See from p214.
[550] See p216.
[551] See from p147.
[552] See from p265.
[553] See from p146.
[554] See from p93.
[555] See from p217.
[556] See p219.
[557] See from p91.
[558] See p219.

that in the future BARB and RAJAR measurement systems could be enhanced or even replaced if a measurable sample of online consumers could be obtained or, better still, a measurable population.[559]

- As noise around social media continues to grow, analysis routes, such as 'content power rating', could further inform audience researchers.[560] Where particular programmes have low sample sizes, social media responses developed to give quantitative date could be utilised as a supplement to appreciation ratings; they could possibly, in the future, replace them.

- It is possible to survey those who do not tend to listen to Radio 4 comedy, as was done in the experiment examining Amazon ratings.[561] Comparing their responses to existing listeners could identify the effect of consumption bias[562] and inform researchers of the attitudes of potential listeners.

10.6.3 Further Research 3 – On a Broader Scale

- On the most basic level, any research that is measuring subjective responses using a scale should be monitoring the distribution of those responses.[563] It is all too easy to present data as if it were measured on an interval or ratio scale when there is no justification for doing so.[564] Further research could seek to explain and highlight the circumstances where researchers might expect to be limited to using non-parametric analysis.

- We have no reason to believe that the response patterns seen in this panel analysis[565] are exclusive to our data set. Further research could investigate whether these patterns are observable in other areas and with other types of scale, such as VAS and Likert-style.

- If response patterns are indeed found elsewhere, it would be extremely interesting to try to understand why respondents behave in this manner. Is it a conscious decision, or are they unaware that they are doing it? Could it be related to respondents attempting to provide consistency across their responses (Bertrand and Mullainathan, 2001, p67)?[566]

10.7 Implications and Recommendations for Practice of BBC Audience Research

The findings of this study have a number of important implications for future practice, particularly for BBC researchers as it is important that they provide accurate information regarding programme evaluation:[567]

10.7.1 – Using Just Existing Data Collection Methods

- Fundamentally, the Pulse survey needs to be more carefully monitored for invalid responses as it was found that these illegitimate ratings affected both corporate-level published figures and programme-level results.[568] Furthermore, we found that the mix of Pulse responses did not reflect the proportion of listening as measured by RAJAR, the industry standard.[569] Work should be undertaken to ensure that the Pulse data reflects actual listening.

---

[559] 'Population' in this case would be all users of the BBC; perhaps feasible with a subscription model and all usage via online routes.

[560] See from p63.

[561] See p155.

[562] See p152.

[563] See from p135.

[564] See from p142.

[565] See p179.

[566] See from p122.

[567] See p66.

[568] See p169 and p170.

[569] See from p175.

- This research identified a number of factors that have an effect on the audience's evaluation of a ratio programme;[570] these aspects should be considered when using appreciation scores to evaluate programmes. An example of how this might be approached is shown from p252.

- Currently, while the AI, as a mean of a non-parametric dataset, is not strictly statistically valid,[571] it does correlate highly with a non-parametric measure.[572] As such, this establishes that it is a valid *proxy* measure, and directly relates to appreciation. That is, the higher the AI, the more a programme is appreciated. However, if the AI continues to be the primary indicator of quality, it should be recognised that it should only be done so when supplemented with additional data relating to the shape of the distribution, as we have found the distribution shape to be bimodal.[573] As a direct result of enquiries for this study, David Raybould, COO of TRP Research,[574] has developed a database allowing easier access to respondent-level data (as of 16th May 2013). This level of data should be utilised to allow rating distribution to be made more transparent and, ideally, the summary measure should be presented using non-parametric data analysis rather than parametric. The ideal presentation should additionally provide the reader with an idea of the extreme responses and allow comparison with either typical responses or, better still, benchmarks for equivalent programmes at sub-genre level.

- Aggregate AI as a measure of station quality should be updated to another style of representation,[575] as it suffers from the same lack of validity as using the mean as a sole indicator at programme level. This is also an issue at the higher level. To fully understand the performance of the station, transparency of the number of people who have more extreme responses is important. Varying genres should be represented individually and presented over a longer time period to highlight trends.

- This analysis found that agreement to a number of additional questions on the Pulse survey were analogous to the appreciation ratings.[576] If this is the case for other genres and stations, these additional queries are redundant and should be removed from the survey. This would mean a shorter survey and could result in a better quality of response and a lower drop-out rate.

- If panel retention could be improved, for example, by making the survey shorter, there could be scope to increase sample sizes. Many of the newest Radio 4 comedies are higher risk, and are scheduled in slots that attract the lowest audience size and, in turn, the lowest number of appreciation responses.[577] Having higher response rates for these shows in particular would aid programme evaluation for radio programmers.

- If a relationship between response rates and audience size could be established, it could be considered that this behavioural indicator was a key performance indicator with regard to appreciation. That is, for entertainment shows (including comedy) in television, it has been observed that audience size is related to appreciation.[578] If this is true for radio, we do not need to ask listeners what they think of a Radio 4 comedy, we simply need to count how many of them are listening to measure its appreciation compared to other shows of the same genre. This, in turn, could be used as a proxy measure for quality. Were the Pulse survey to be more prescriptive about the programmes that the respondents are expected to rate, i.e., more akin to the RAJAR sweep, then responses would be more likely to

---

[570] See from p226.

[571] See from p141.

[572] See p201.

[573] See p178.

[574] TRP is the company that takes the GfK raw Pulse survey data and processes it for the BBC.

[575] See p260 for current presentation.

[576] See from p220.

[577] See p35.

[578] See p91.

represent the true audience size. For example, currently the respondent is asked to respond ten times in any month and there could be external influences which affect the response on certain days: such as weekdays versus weekends, or whether the listeners heard a programme that they particularly liked or hated. If the respondents were just told to, for example, rate everything that they heard in a specific day or week, this could eliminate problems such as extreme response bias[579] and ensure that response rate related directly to audience size and could become a useful behavioural measure.

10.7.2 – Allowing for Expansion of Data Collection

- If research indicated that VAS scales could deliver appreciation ratings on an interval scale,[580] they would allow more robust analysis and if the data were normally distributed, a mean score (the AI) *would* be a valid representation of the central tendency.

- The theory that appreciation is related to utility[581] means that, perhaps, a 'cover-all' rating question regarding the programme's quality may not be the best indicator. For example, for comedy, we might prefer to ask the respondents, 'How funny did you find this programme?' to get the best indication of the show's performance. Another question that could be revealing is whether the respondent is listening alone or with others. As comedy is considered to be a genre that is best experienced when not alone,[582] understanding of how solo listening affects appreciation could be of interest. While it may be the case that adding additional questions to a survey could result in lower completion rates (Best & Krueger, 2008, p223), evidence of redundancy elsewhere in the survey has been found[583] and reduction in the number of these duplicate questions would allow for these additional question.

- There is an issue with sample sizes, particularly for newer Radio 4 comedies, and it would be problematic to increase sample sizes just for these particular slots. One solution could be to examine the style of research undertaken back in 1938. A group of variety 'enthusiasts' were sought, and of the 47,000 volunteers, a panel of 2,000 was selected.[584] Considering the relative ease with which such a panel could be maintained through social media today, this could allow for more in-depth responses than might be obtained from the current Pulse panel. Self-professed enthusiasts may be keen to aid the network directly with its decision making. Maintaining a number of genre-specific panels could be an alternative or supplement to the Pulse survey.

- Longer term, it may be possible to dispense with the Pulse panel all together. If the BBC moved to a subscription model, behavioural analysis could be the primary indicator of programme quality, i.e., the more people who listen to a Radio 4 comedy, the better it is. Further measures from social media could provide additional information as to the reasons behind the behaviour by investigating areas such as sentiment analysis and content power ratings.

---

[579] See p154.

[580] See p146.

[581] See from p126.

[582] See from p100.

[583] See from p219.

[584] See from p68.

Bibliography

Private Interviews

Anonymous BBC Radio Comedy Producer, (2010). Interview. [Details witheld].

Anonymous BBC Radio Comedy Producer, (2011). Interview. [Details witheld].

Anonymous BBC Radio Comedy Writer, (2012). Interview. [Details witheld].

Barwise, P. (2013). Interview. 5th February 2013. [Patrick Barwise: Emeritus Professor of Management and Marketing at The London Business School].

Baxter, S. (2014). Interview. 29th May 2014. [Stanley Baxter: Radio comedy performer since 1940, most recently *The Stanley Baxter Playhouse*].

Bennett, S. (2011). Interview. 29th March 2011. [Steve Bennett: Editor and main reviewer for *Chortle.co.uk* (website for comedy listings and reviews with circa 10,000 unique users each day)].

Berthould, J. (2010). Interview. 8th September 2010. [Jane Berthoud: Head of BBC Radio Comedy].

Black, M.I. (2011). Interview. 10th November 2011. [Michael Ian Black: US Stand-up, actor, writer, director, podcaster with circa 2 million *Twitter* followers].

Canny, S. (2011). Interview. 7th September 2011. [Steve Canny: then Executive Producer for BBC Radio Comedy].

Collins, P. (2010a). Interview. 6th May 2010. [Patrick Collins: BBC Senior Research Executive: Statistical Analysis].

Collins, P. (2010b). Interview. 7th October 2010. [Patrick Collins: BBC Senior Research Executive: Statistical Analysis].

Cook, D. (2010). Interview. 15th July 2010. [Daniel Cook: then BBC Researcher: Head of Special Projects].

Dare, B. (2012). Interview. 16th January 2012. [Bill Dare: Radio Comedy producer, writer and creator of *Dead Ringers*, *I've Never Seen Star Wars* and *Brian Gulliver's Travels*].

Davie, T. (2009). Interview. 23rd April 2009. [Tim Davie: then Director of BBC Audio and Music].

Doody, N. (2011). Interview. 22nd Nov 2011. [Nick Doody: Stand-up and creator of *Bigipedia* and writer/performer for *The Now Show*, *Armando Iannucci's Charm Offensive* and *Tonight*].

Galton, R. and Simpson, A. (2012). Interview. 5th March 2012. [Ray Galton and Alan Simpson: Writers and Creators of *Hancock's Half Hour* and *Steptoe and Son*].

Hubbard, K. (2012). Interview. 21st March 2012. [Katy Hubbard: Presentation Editor BBC Radio 4 and BBC Radio 4Extra].

Lock, S. (2011). Interview. 17th March 2011. [Sean Lock: Writer and star of *15 Storeys High* on R4 and BBC3].

Mayhew-Archer, P. (2011). Interview. 17th Jan 2011. [Paul Mayhew-Archer: Radio 4 Acting Comedy Commissioning Editor during 2010/11, radio comedy producer and co-writer of *The Vicar of Dibley*].

McConway, K. (2013). Interview. 11th Oct 2013. [Kevin McConway: Professor of Applied Statistics in the Department of Mathematics and Statistics at The Open University].

Mitchell, D. (2010). Interview. 5th November 2010. [David Mitchell: Writer and performer on *That Mitchell and Webb Sound* and *That Mitchell and Webb Look*, and Chairperson on *The Unbelievable Truth*].

Naismith, J. (2010). Interview. 24th September 2010. [Jon Naismith: Producer of *I'm Sorry I Haven't A Clue* as a freelance BBC producer and *The Unbelievable Truth* as an independent producer].

North, N. (2010). Interview. 9th June 2010. [Nick North: Global Lead, Media and Entertainment, GfK].

Raphael, C. (2012a). Interview. 27th March 2012. [Caroline Raphael: Commissioning Editor for Radio Comedy, BBC Radio 4 and 4 Extra].

Reynolds, G. (2011). Interview. 23rd Nov 2011. [Gillian Reynolds: Radio Critic for *The Telegraph*].

Rumble, J. (2010). Interview. 7th October 2010. [Jane Rumble: Head of Media Research at Ofcom].

Schlesinger, P. (2010). Interview. 11th May 2011. [Paul Schlesinger: BBC TV Comedy Executive Producer and former Head of BBC Radio Comedy 2006-2009 (Light Entertainment as it was called)].

Private Correspondence

Herring, R. (2010). Email correspondence. 25th September 2010. [Richard Herring: Radio 4 comedy writer and performer].

McConway, K. (2010). Email correspondence. 26th October 2010. [Kevin McConway: Professor of Applied Statistics in the Department of Mathematics and Statistics at The Open University].

Menneer, P. (2014). Email correspondence. 15th-20th May 2014. [Peter Menneer: Former Head of BBC Research].

Raphael, C. (2012b). Email correspondence. 11th Aug 2012. [Caroline Raphael: Commissioning Editor for Radio Comedy, BBC Radio 4 and 4 Extra].

Tingay, J. (2013). Email correspondence. 28th August 2013. [Jo Tingay: Senior Research Manager TRP].

Audio Recorded at Conferences and Presentations

Cryer, B. (2009). Interviewed as part of: *Radio Academy Celebration of Radio Comedy.* London. 18th March 2009. [Barry Cryer: Radio and Television Comedy writer and star of *I'm Sorry I Haven't A Clue*].

Holden, J. (2010). Engaging with UK Audiences via Digital Media: A Public Service Perspective. *Ripe@2010 Conference: Public Service Media After the Recession.* 8th-11th Sept 2010. At BBC Television Centre: Hosted by The University of Westminster. [James Holden: BBC Director of Audiences].

Punt, S. (2009). Interviewed as part of: *Radio Academy Celebration of Radio Comedy*. London. 18th March 2009. [Steve Punt: Radio comedy writer and star of *The Now Show*].

Thickett, J. (2010). The UK PSB Landscape in 2010. *Ripe@2010 Conference: Public Service Media After the Recession.* 8th-11th Sept 2010. At BBC Television Centre: Hosted by The University of Westminster. (Presentation available online http://ripeat.org/wp-content/uploads/tdomf/1995/Thickett.pdf [accessed 06/08/2012]). [James Thickett: Director of Research and Market Intelligence, Ofcom].

Broadcasts

Atkinson, G. (2008). Talking on *Pilots That Never Flew*, episode 2. BBC Radio 4 FM, 15:45, 30th December 2008. [Geoff Atkinson: Comedy writer and producer].

Auntie's War on Smut (2008). BBC4 television. 18th March, 2008, 21:00.

Carr, J. (2011). Talking on *The Museum of Curiosity.* BBC Radio 4 FM, 18:30, 3rd October 2011. [Jimmy Carr: Comedy writer and performer].

Feedback (2012a). BBC Radio 4 FM, 16:30, 23rd March 2012.

Feedback (2012b), BBC Radio 4 FM, 16:30, 30th March 2012.

Lloyd, J. (2009). Talking on *Pilots That Never Flew,* episode 4. BBC Radio 4 FM, 15:45. 1st January 2009. [John Lloyd: Comedy writer and producer].

McGowan, A. (2008). Talking on *Pilots That Never Flew*, episode 1. BBC Radio 4 FM, 15:45. 29th December 2008. [Alistair McGowan: Comedy writer and performer].

Satiwrist (2010). *Radio 4 Comedy Commissioning.* YouTube broadcast. Added to YouTube: February 26, 2010. http://www.youtube.com/watch?v=BLPvAFlVPqQ. [Accessed 2nd Oct 2010].

BBC Reports, Meetings and Blogs

BBC A&M Report (2008). *The Future of Speech Radio*. Presented at The Radio Festival, July 2008.

BBC Archives - Caversham (accessed 2012). A selection of *Daily Listening Barometers* and *Audience Research Reports* viewed in the BBC archives at Caversham.

BBC Audience Research (2005). *What the Audience Wants from Comedy*. November 2005.

BBC Audience Research (2009). *Representing the UK on BBC Network Radio*. May 2009.

BBC Audience Research (2009). *Understanding AIs and the Pulse panel.* March 2009.

BBC Audience Research (2009). *Share of Ear*. August 2009.

BBC Audience Research (2010). *Comedy at the Weekend*. October 2010.

BBC Executive (2009). *Taste, Standards and the BBC. Public attitudes to morality, values and behaviour in UK broadcasting.* http://downloads.bbc.co.uk/guidelines/editorialguidelines/research/taste-standards.pdf [Accessed 11/04/2012].

BBC Executive (2010). *BBC Executive Submission to the Trust's Review of BBC Radio 4*.

BBC Gateway: Audience Portal (2011). W*hat exactly is an AI score?* http://www.audiencesportal.com/article_list/audiences_articles/ai_video.aspx [Accessed 25/11/11].

BBC Radio 4 (2006 - 2012). *Radio 4 Commissioning Guidelines*.

Counterpoint Research (1997a). Radio 4 Scheduling. Report 1: Background to the Dayparts, and Typology of Listeners. *Counterpoint Research.* July 1997.

Counterpoint Research (1997b). Radio 4 Scheduling. Report 2: Weekday Mornings. *Counterpoint Research.* July 1997.

Counterpoint Research (1997c). Radio 4 Scheduling. Report 4: Weekday Evenings. *Counterpoint Research.* July 1997.

Counterpoint Research (1997d). Radio 4 Scheduling. Report 6: Sunday. *Counterpoint Research.* July 1997.

Damazer, M. (2010). Interviewed by Libby Purves at the BBC Audio and Music staff meeting in the BBC Radio Theatre, 14th September 2010. Video was available on the BBC intranet: http://explore.gateway.bbc.co.uk/mcanda/audio__music/am_news/allstaffmeeting140910.aspx [accessed 3rd Oct 2010]. [Mark Damazer: Controller of BBC Radio 4 at the time].

Davie, T. (2010). Presentation given for BBC Audio and Music staff meeting in the BBC Radio Theatre, 14th September 2010. Video was available on the BBC intranet: http://explore.gateway.bbc.co.uk/mcanda/audio__music/am_news/allstaffmeeting140910.aspx [accessed 3rd Oct 2010]. [Tim Davie: Director of BBC Audio & Music at the time].

Mayhew-Archer, P. (2010). Q&A lunchtime session with [then] BBC7 Network Team. 24th Nov 2010. [Paul Mayhew-Archer: BBC Radio 4 Commissioning Editor for Comedy at the time].

Mytton, G. (1997). Handbook on Radio and Television Audience Research. *BBC World Service Training Trust.* http://unesdoc.unesco.org/images/0012/001242/124231eo.pdf [Accessed 25/5/10].

National Audit Office (2009). The Efficiency of Radio Production at the BBC: Review by the Comptroller and Auditor General presented to the BBC Trust's Finance and Strategy Committee. *National Audit Office.* 5th February 2009.

Pilgrim, T. (2009). *Scheduling Radio 4*. BBC Radio 4 Blog: 22nd April 2009. http://www.bbc.co.uk/blogs/radio4/2009/04/scheduling_radio_4.html [Accessed 14th April 2012].

Raphael, C. (2011). Presentation to Suppliers at the Radio 4 Commissioning Launch for 2012/13. London, Broadcasting House. 18th March 2011. [Caroline Raphael: Commissioning Editor for Radio Comedy, BBC Radio 4 and 4 Extra]

Stone, J. (2012). Real Time Measurement Could Change the Future of Radio. *BBC Audio and Music Interactive Conference*. 23rd Nov 2012. [Jem Stone: BBC Executive Producer, SM Syndication, Social Media Team]

Williams, G. (2011). *Guarding Radio's Sacred Flame: Gwyneth Williams in conversation with Gillian Reynolds*. Council Chamber, BBC Broadcasting House, 16th November 2011. [Gwyneth Williams: Controller of BBC Radio 4].

Winter, A. (2011). RAJAR. *Audio and Music Learning Lunch*. 16th February 2011. [Alison Winter: BBC Head of Audiences, Radio].

Books

Aitkin, R. (2007). *Can We Trust the BBC?* London: Continuum.

Ang, I. (1991). *Desperately Seeking The Audience*. London: Routledge.

Barnard, S. (2000). *Studying Radio*. London: Arnold.

Barnes, B.E. and Thomson, L.M. (1994). Power to the People (Meter): Audience Measurement Technology and Media Specialization. In: Ettema, J.S., and Whitney, D.C. (eds.) *Audiencemaking: How the Media Create the Audience.* London: Sage.

Barwise, T.P. and Ehrenberg, A.S.C. (1988). *Television and its Audience.* London: Sage.

Barwise, P. and Meehan, S. (2011). *Beyond The Familiar. Long-term Growth Through Customer Focus And Innovation.* Chichester: John Wiley and Sons.

BBC Audience Research Department (1953). *BBC Handbook for Interviewers*. BBC publication.

BBC Audience Research Department (1976). *Annual Review of BBC Audience Research Findings 1974/75, Volume 2*. London: British Broadcasting Corporation.

BBC Audience Research Department (1978). *Annual Review of BBC Audience Research Findings 1976/77, Volume 4*. London: British Broadcasting Corporation.

BBC Broadcasting Research Department (1980). *Annual Review of BBC Broadcasting Research Findings 1980, Volume 7*. London: British Broadcasting Corporation.

BBC Broadcasting Research Department (1984). *Annual Review of BBC Broadcasting Research Findings 1983, Volume 9*. London: BBC Data Publications.

BBC Broadcasting Research Department (1985). *Annual Review of BBC Broadcasting Research Findings 1984, Volume 10*. London: BBC Data Publications.

BBC Broadcasting Research Department (1986). *Annual Review of BBC Broadcasting Research Findings 1985, Volume 11.* London: BBC Data Publications.

BBC Broadcasting Research Department (1990). *Annual Review of BBC Broadcasting Research Findings 1989, Volume 15.* London: John Libbey and Company Ltd.

BBC Broadcasting Research Department (1993). *Annual Review of BBC Broadcasting Research Findings 1992, Volume 18.* London: John Libbey and Company Ltd.

The BBC Year Book (1932). BBC publication.

The BBC Year Book (1934). BBC publication.

The BBC Handbook (1939). BBC publication.

The BBC Handbook (1941). BBC publication.

The BBC Handbook (1942). BBC publication.

The BBC Year Book (1945). BBC publication.

The BBC Year Book (1947). BBC publication.

The BBC Year Book (1948). BBC publication.

The BBC Year Book (1949). BBC publication.

The BBC Year Book (1951). BBC publication.

The BBC Handbook (1956). BBC publication.

The BBC Handbook (1958). BBC publication.

Beck, A. (1997). *Radio Acting.* London: A and C Black.

Beck, A. (2006). You've Got to Hide Your Love Away: Gay Radio. In: Crisell, A. (ed.) *More Than A Music Box. Radio Cultures and Communities in a Multi-Media World.* USA, New York: Berghahn Books.

Best, S.J. and Krueger, B.S. (2008). Internet Survey Design. In: Fielding, N., Lee, R.M. and Blank, G. (eds.) *The Sage Handbook of Online Research Methods.* London: Sage.

Blastland, M. and Dilnot, A. (2008). *The Tiger That Isn't. Seeing Through A World Of Numbers*. Expanded edition. London: Profile Books.

Bradbury, D. and McGrath, J. (1998). *Now That's Funny! Writers on Writing Comedy*. London: Methuen.

Briggs, A. (1961). *The History of Broadcasting in the United Kingdom: Volume I: The Birth of Broadcasting.* London: Oxford University Press.

Briggs, A. (1965). *The History of Broadcasting in the United Kingdom, Volume II: The Golden Age of Wireless.* London: Oxford University Press.

Briggs, A. (1970). *The History of Broadcasting in the United Kingdom, vol. III: The War of Words.* London: Oxford University Press.

Briggs, A. (1979). *The History of Broadcasting in the United Kingdom: Volume IV: Sound and Vision.* London: Oxford University Press.

Briggs, A. (1995). *The History of Broadcasting in the United Kingdom, Volume V: Competition.* London: Oxford University Press.

Brooke-Taylor, T., Cryer, B., Rushton, W., Garden, G., Lyttelton, H. and Pattinson, I. (1999). *The Almost Totally Complete I'm Sorry I Haven't A Clue*. London: Orion Books Ltd.

Bryman, A. and Cramer, D. (1994). *Quantitative Data Analysis For Social Scientists* (revised edition). London: Routledge.

Cantril, H. and Allport, G.W. (1935). *The Psychology of Radio*. New York: Harper and Brothers.

Caulfield, A. (2009). *Writing For Radio.* Malborough: The Crowood Press Ltd.

Chapman, A.J. (1976). Social Aspects of Humorous Laughter. In: Chapman, T. and Foot, H. (eds.) *Humour and Laughter: Theory. Research and Applications.* Bath: John Wiley and Sons, Ltd.

Chignell, H. (2009). *Key Concepts in Radio Studies*. London: Sage.

Coward, M. (2003). *Classic Radio Comedy.* Harpenden: Pocket Essentials.

Crisell, A. (ed.) (1994). *Understanding Radio.* 2nd ed. London: Routledge.

Crisell, A. (2002a). *An Introductory History of British Broadcasting.* 2nd ed. London: Routledge.

Crisell, A. (2002b). Radio: public service, commercialism and the paradox of choice. In: Briggs, A. and Cobley, P. (eds.) *The Media: An Introduction.* 2nd ed. Harlow: Pearson Education Limited.

Crisell, A. (ed.) (2006). *More Than A Music Box. Radio Cultures and Communities in a Multi-Media World*. New York: Berghahn Books.

Crook, T. (1999). *Radio Drama. Theory and Practice*. Abingdon: Routledge.

De Mooij, M. (1998). *Global Marketing and Advertising.* London: Sage.

De Vaus, D. (2002). *Surveys in Social Research.* 5th ed. London: Routledge.

Douglas, S. J. (2004). *Listening In. Radio and the American Imagination*. Minneapolis, USA: University of Minnesota Press.

Ehrenberg, A.S.C. (1986). *A Primer In Data Reduction.* Published in 1982, reprinted with corrections. Chichester: John Wiley and Sons.

Elmes, S. (2007). *And Now On Radio 4.* London: Random House.

Evans, E. (1977). *Radio: A Guide to Broadcasting Techniques.* London: Barrie and Jenkins Ltd.

Farnes, N. (1997). *The Goons. The Story.* London: Virgin Publishing limited.

Foster, A. and Furst, S. (1996). *Radio Comedy, 1938-1968.* London: Virgin.

Fricker, R.D. (2008). Sampling Methods for Web and E-mail Surveys. In: Fielding, N., Lee, R.M. and Blank, G. (eds.) *The Sage Handbook of Online Research Methods*. London: Sage.

Fry, S. (2010). *The Fry Chronicles.* London: Penguin.

Galton, R. and Simpson, A. (2011a). *The Masters of Sitcom. From Hancock to Steptoe.* London: Michael O'Mara Books Limited.

Goldman, W. (1985). *Adventures in the Screentrade.* London: Futura.

Goodale, G. (2011). *Sonic Persuasion.* Illinois, USA: University of Illinois Press.

Goodwin, C. (1999). *When the Wind Changed: The Life and Death of Tony Hancock.* London: Century.

Gorham, M. (1948). *Sound and Fury.* London: Percival Marshall.

Gray, F. (2006). Fireside Issues: Audience, Listener, Soundscape. In: Crisell, A. (ed.) M*ore Than A Music Box. Radio Cultures and Communities in a Multi-Media World*. USA, New York: Berghahn Books.

Gross, B., Gross, L.S. and Perebinossoff, P. (2005). *Programming for TV, Radio and The Internet: Strategy, Development and Evaluation.* Oxford: Focal Press.

Gruner, C.R. (1976). Wit and Humour in Mass Communication. In: Chapman, T. and Foot, H. (eds.) *Humour and Laughter: Theory. Research and Applications.* Bath: John Wiley and Sons, Ltd.

Gunter, B. and Wober, M. (1992). *The Reactive Viewer*. London: John Libbey and Son.

Gunter, B. (2000). *Media Research Methods.* London, Sage.

Gunter, B. (2012). The Quatitative Research Process. In: Jensen, K.B. (ed.) *A Handbook of Media and Communication Research*. London: Routledge

Hawes, C. (2004). The Ballad of Charles Parker. In: Boardman-Jacobs, S. (ed.) *Radio Scriptwriting.* Bridgend: Seren.

Helitzer, M. (2005). *Comedy Writing Secrets.* Ohio, USA: Writer's Digest Books.

Hendy, D. (2000). *Radio in the Global Age.* Cambridge: Polity Press.

Hendy, D. (2006). 'Reality Radio': The Documentary. In: Crisell, A. (ed.) *More Than A Music Box. Radio Cultures and Communities in a Multi-Media World*. USA, New York: Berghahn Books.

Hendy, D. (2007). *Life on Air. A History of Radio 4*. Oxford: Oxford University Press.

Hepp, A. (2006). Radio and Popular Culture in Germany: Radio Culture Between Comedy and 'Event-isation'. In: Crisell, A. (ed.) *More Than A Music Box. Radio Cultures and Communities in a Multi-Media World.* USA, New York: Berghahn Books.

Kahneman, D. (2011). *Thinking Fast and Slow*. London: Penguin.

Keith, M.C. (2004). *The Radio Station.* 6th ed. Oxford: Focal Press

Kent, R. (1994). Measuring Media Audiences: An Overview. In: Kent, R. (ed.) *Measuring Media Audiences.* London: Routledge.

Kent, R. A. (2002). Audience Research: Administrative Research of Audiences. In: Briggs, A. and Cobley, P. (eds.) *The Media: An Introduction.* 2nd ed. Harlow: Pearson Education Limited.

La France, M. (1983). Felt versus Feigned Funniness: Issues in Coding Smiling and Laughing. In: McGhee, P.E. and Goldstein, H. (eds.) *Handbook of Humor Research.* USA, New York: Springer-Verlag.

Lewis, P. (ed.) (1981). *Radio Drama*. Essex: Longman Group Limited.

Lewis, P.M. (1991). Referable words in radio drama. In: Scannell, P. (ed.) *Broadcast Talk*. London: Sage.

Lynch, R. (2006). *Corporate Strategy.* 4th ed. Harlow: Prentice Hall.

Lynn, J. (2011). *Comedy Rules: From the Cambridge Footlights to Yes Prime Minister.* London: Faber and Faber.

Lyttelton, H. (2006). *It Just Occurred To Me…* London: Anova Books.

MacFarland, D.T. (1997). *Future Radio Programming Strategies.* 2nd ed. New Jersey: Lawrence Erlbaum Associates.

McCann, G. (2002). *Dad's Army.* London: Fourth Estate.

McLeish, R. (1999). *Radio Production*. Oxford: Focal Press.

McQuail, D. (1997). *Audience Analysis.* London: Sage.

McWhinnie, D. (1959). *The Art of Radio.* London: Faber and Faber.

Miller, P.V. (1994). Made-to-order and Standardized Audiences: Forms of Reality in Audience Measurement. In: Ettema, J.S. and Whitney, D.C. (eds.) *Audiencemaking: How the Media Create the Audience.* London: Sage.

Mitchell, D. (2012). *Back Story. A Memoir.* London: Harper Collins.

Moser, C.A. and Kalton, G. (1971). *Survey Methods in Social Investigation.* 2nd ed. London: Heinemann Educational Books.

Nathan, D. (1971). *The Laughtermakers*. London: Peter Owen Limited.

O'Sullivan, T. (2000). Series Editor's Preface. In Barnard, S., *Studying Radio.* London: Arnold.

Osborne, J. (2008). *Radio Head: Up and Down the Dial of British Radio.* London: Simon and Schuster.

Pagano, R.R. (2001). *Understanding Statistics in the Behavioral Sciences.* 6th ed. USA, California: Wadsworth.

Parsons, N. (2010). *With Just a Touch of Hesitation, Repetition and Deviation.* Edinburgh: Mainstream Publishing.

Partridge, A. (2011). *I, Partridge: We need to talk about Alan.* London: Harper Collins.

Provenza, P. and Dion, D. (2010). *¡SATARISTAS! Comedians, Contrarians, Raconteurs and Vulgarians.* New York: Harper Collins.

Raban, J. (1981). Icon or symbol: the writer and the 'medium". In: Lewis, P. (ed.) *Radio Drama.* Essex: Longman Group Limited

Scannell, P. (ed.) (1991). *Broadcast Talk.* London: Sage.

Sharot, K. (1994). Measuring Television Audiences in the UK. In: Kent, R. (ed.) *Measuring Media Audiences*. London: Routledge.

Shingler, M. and Wieringa, C. (1998). *On Air: Methods and Meanings of Radio.* London: Arnold.

Silvey, R. (1974). *Who's Listening? The Story of BBC Audience Research.* London: George Allen and Unwin.

Starkey, G. (2004). *Radio in Context.* Hampshire: Palgrave MacMillan.

Street, S. (2002). *A Concise History of British Radio.* Devon: Kelly Publications.

Street, S. (2006). *Historical Dictionary of British Radio.* USA: Scarecrow Press.

Sturgis, P., Allum, N. and Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In: Lynn, P. (ed.) *Methodology of Longitudinal Surveys.* London, Wiley-Blackwell.

Took, B. (1981). *Laughter in the Air. An Informal History of British Radio Comedy.* London: Robson Books and the BBC.

Townsend, S. (2004). *Adrian Mole and the Weapons of Mass Destruction.* London: Penguin.

Twyman, T. (1994). Measuring Audiences to Radio. In: Kent, R. (ed.) *Measuring Media Audiences.* London: Routledge.

Tydeman, J. (1981). The producer and radio drama: a personal view. In: Lewis, P. (ed.) *Radio Drama.* Essex: Longman Group Limited

Vehovar, V. and Manfrede, K.L. (2008). Overview: Online Surveys. In: Fielding, N., Lee, R.M. and Blank, G. (eds.) *The Sage Handbook of Online Reseach Methods.* London: Sage.

Walliman, N. (2005). *Your Research Project.* 2nd Ed. London: Sage.

Webb, N. (2003). *Wish You Were Here.* London: Headline.

Webster, J.G. Phalen, P.F. and Lichty, L.W., (2006). *Ratings Analysis.* 3rd ed. USA, New Jersey: Lawrence Erlbaum Associates Ltd.

Webster, J.G. (2001). Audience Measurement. In: Smelser, N.J. and Bates, P.B. (eds.) *International Encyclopedia of the Social and Behavioral Sciences.* Oxford: Pergamon.

Wertheim, A.F. (1979). *Radio Comedy.* USA: Oxford University Press.

Zillman, D. and Cantor, J.R. (1976). A Disposition Theory of Humour and Mirth. In: Chapman, T. and Foot, H. (eds.) Humour and Laughter: Theory. Research and Applications. Bath: John Wiley and Sons, Ltd.

Academic Articles

Allen, I.E. and Seaman, C.A. (2007). Likert scales and data analysis. *Quality Progress.* **40** (7), 64-66.

Barwise, T.P. and Ehrenberg, A.S.C. (1987). The Liking and Viewing of Regular TV Series. *Journal of Consumer Research.* 63-70.

Barwise, T. P., Ehrenberg, A.S.C. and Goodhardt, G. J. (1979). Audience Appreciation and Audience Size. *Journal of the Market Research Society.* **21**, 269-289.

Benson, A. and Perry, S. (2006). The influence of Humor in Radio Advertising on Program Enjoyment and Future Intent to listen. *Journal of Radio Studies.* **13** (2), 169-186.

Bertrand, M. and Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *The American Economic Review.* **91** (2), 67-72.

Cantor, J., Bryant, J. and Zillman, D. (1974). Enhancement of humor appreciation by transferred excitation. *Journal of Personality and Social Psychology.* **30**, 812-821.

Castleberry, S.B., Barnard, N.R., Barwise, T.P., Ehrenberg, A.S.C. and Dall'Olmo Riley, F. (1994). Individual Attitude Variations Over Time. *Journal of Marketing Management*. **10**, 153-162.

Coelho, P.S. and Esteves, S.P. (2007). The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement. *International Journal Of Market Research.* **49** (3), 313-339.

Danaher, P.J. and Lawrie, J.M. (1998). Behavioral Measures of Television Audience Appreciation. *Journal of Advertising Research.* **38** (1), 54-65.

Dawes, J.G. (2008). Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5 Point, 7 Point and 10 Point Scales. *International Journal of Market Research*. **50**, 61–104.

Greenleaf, E.A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly.* **56**, 238-351.

Grondin, S. (2010). Timing and time perception: 'A review of recent behavioural and neuroscience findings and theoretical directions'. *Attention, Perception and Psychophysics*. **72** (3), 561-582.

Hu, N., Pavlou, P. A. and Zhang, J. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*. **52** (10), 144-147.

Jackson, D.E. and Jackson, K.A. (1997). Gender, object and humor. *Current Psychology*. **15** (4). 286-291.

Jamieson, S. (2004). Likert Scales: How to (ab) use Them. *Medical Education.* 38 (12) 1,217-1,218.

Knapp, T.R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing Research*. **39**, 121–123.

Koh, N. S., Hu, N. and Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications.* **9** (5), 374-385.

Lacey, K. (2008). Ten years of radio studies: The very idea. *The Radio Journal.* **6** (1), 21-31.

Lieberman, E.A., Neuendorf, K.A., Denny, J., Skalski, P.D. and Wang, J. (2009). The Language of Laughter: A Quantitative/Qualitative Fusion Examining Television Narrative and Humor. *Journal of Broadcasting and Electronic Media*. **53** (4), 497-514.

Madden, T.J. and Weinberger, M.G. (1982). The Effects of Humor on attention in Magazine Advertising. *Journal of Advertising*. **11** (3), 8-14.

Martin, G.N. and Gray, C.D. (1996). The Effects of Audience Laughter on Men's and Women's Responses to Humor. *The Journal of Social Psychology*. **136** (2), 221-231.

Menneer, P. (1987). Audience appreciation: A different story from audience numbers. *Journal of the Market Research Society.* **29** (3), 241-264.

Perry, S. D., Jenzowwski, S. A., King, C. M., Hester, J. B. and Yi, H. (1997a). The influence of commercial humor on program enjoyment and evaluation. *Journalism and Mass Communication Quarterly*. **74**, 388-399.

Reips, U-D. and Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator'. *Behavior Research Methods.* **40** (3). 699-704.

Robinson, L. (2000). Radio Research in Transition. *International Journal of Market Research*. **42,** (4), 381-396.

Rocereto, J. F., Puzakova, M., Anderson, R. E. and Kwak, H. (2011). The Role of Response Formats on Extreme Response Style: A Case of Likert-Type vs. Semantic Differential Scales. *Advances in International Marketing*. **22**, 53-71.

Starkey, G. (2002). Radio Audience Research: Challenging the 'gold standard'. *Cultural Trends*. **45**, 43-68.

Stevens, S.S. (1946). On the Theory of Scales of Measurement. *American Association for the Advancement of Science*. **103** (2684), 677-680.

Studer, R. (2011). Does it Matter How Happiness is Measured? Evidence From a Randomized Controlled Experiment. *Working Paper Series: University of Zurich.* [Subsequently published: Studer, R. (2012). Does it matter how happiness is measured? Evidence from a randomized controlled experiment. *Journal of Economic and Social Measurement*. **37** (4), 317-336.]

Sun, M. (2012). How Does the Variance of Product Ratings Matter? *Management Science*. **58** (4), 696-707.

Svensson E. (2001). Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of Rehabilitation Medicine.* **33**, 47-48.

Tourangeau, R. and Rasinski, K.A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*. **103** (3), 299-314.

Van Vaerenbergh, Y. and Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research.* **25** (2), 195-217.

Weinberger, M.G. and Spotts, H.E., (1989). Humor in U.S. versus U.K. TV Commercials: A Comparison. *Journal of Advertising*, **8** (2), 39-44.

Unpublished Doctoral Thesis

Carrie, D.G. (1997). Patterns of Audience Appreciation Ratings for Television Programmes. University of London [Unpublished Doctoral Thesis].

Conference Papers

Bjorna, H., Karsai, F., Vicary, R., Wagner, R. and Perry, S. D. (2001). *The impact of commercial excitation on program appreciation.* Broadcast Education Association, Las Vegas, NV. April 2010.

Engel, B. (2009). Improving appreciation with accompanying programme information. *GfK Appreciation Panel System User Group.* Dublin. 12th-13th June 2009.

Hunter, S. (2010). Head of UK Public Policy Google Europe: The User-First Approach. *Ripe@2010 Conference: Public Service Media After the Recession*. At BBC TVC: Hosted by The University of Westminster. 8th-11th September 2010.

Perry, S.D. (1998). Commercial humor influence on sitcom liking: A re-evaluation. T*he Mass Communication and Society division of the Broadcast Education Association (BEA)*. Las Vegas, NV. April 1998.

Preunkert, T. (2009). Appreciation scores minute per minute. *GfK Appreciation Panel System User Group.* Dublin. 12th-13th June 2009.

Sanner, R. E., Grant, A. J., Collins, T. D. and Perry, S. D. (2001). Gender Differences in Humor Appreciation: Effects on Program Liking. *The NCA Mass Communication Division.* Atlanta, GA. November, 2001.

Treiblmaier, H. and Filzmoser, P. (2009). Benefits from using continuous rating scales in online survey research. *Technische Universitt Wien, Forschungsbericht*. November 2009.

Van Meurs, L. (2009a). Comparison of the Appreciation Panels. *GfK Appreciation Panel System User Group.* Dublin. 12th-13th June 2009.

Van Meurs, L. (2009b). Checks on the Quality of Online Responses. *GfK Appreciation Panel System User Group.* Dublin. 12th-13th June 2009.

Press

Armstrong, S. (2007). Why radio comedy is a joke. *The Guardian:* Media Guardian. 15th January 2007. 1.

Brown, M. (2012). Recession sparks 'golden age' in comedy as TV chiefs double demand for new scripts: Audiences are yearning for shows packed with jokes - and it can be ratings gold dust if the channels get it right. *The Observer* (home news pages). 13th May 2012. 11.

Chapman, A. (2014). Appreciation for all BBC TV channels falls. *Broadcast*. http://www.broadcastnow.co.uk/appreciation-for-all-bbc-tv-channels-falls/5075430.article. 21st July 2014. [Accessed 18/09/2014].

Chortle (2011a). Dark days are over: TV wants big, 'laugh-out-loud' comedies. *Chortle*. 2nd November 2011. http://www.chortle.co.uk/news/2011/11/02/14234/dark_days_are_over. [Accessed 17/10/11].

Chortle (2011b). Cutting out the middle men: Louis CK releases stand-up special direct. *Chortle*. 4th November 2011. http://www.chortle.co.uk/news/2011/11/04/14246/cutting_out_the_middle_men. [Accessed 17/10/11].

Clarke, S. (2011). A Quick Guide to Cheap Laughs. *The Journal of the Royal Television Society*. 2nd July 2011. 21.

Clinton, J. (2012). Is comedy cruelty a laughing matter? *The Sunday Express*. 15th January 2012. 56.

Dowell, B. (2008). Sound Judgement: Television Comedy has long borrowed from Radio. But should commissioners be casting their nets wider? *The Guardian*, The Media Guardian, 28th April 2008. 9.

Foster. P. (2011). BBC scores highest audience appreciation figures ever. *The Guardian.* 19th September 2011. http://www.guardian.co.uk/media/2011/sep/19/bbc-high-audience-appreciation-score [Accessed 19/09/2011].

Frean, A. (2011). The signal fades for TV ratings. *The Times*. 26th May 2011. 59.

Galton, R. and Simpson, A. (2011b). Galton and Simpson: Our view. *The Daily Mail.* 17th December, 2011. 15.

Greer, G. (2009). Beaten to the punchline. *The Guardian,* Guardian 2. 2nd March 2009, 22.

Kirby, J. and Samson, A. (2008). Customer advocacy metrics: the NPS theory in practice. *Admap.* February 2008. 17-19. http://www.tmiaust.com.au/downloads/The%20NPS%20Theory%20in%20Practice.pdf [accessed 08/08/2012].

Lawson, M. (2012). Just a Minute needs a format deviation to succeed on TV. *The Guardian Online, TV and Radio Blog.* 28/03/2012. http://www.guardian.co.uk/tv-and-radio/tvandradioblog/2012/mar/28/just-a-minute-tv-radio [Accessed 27/04/2012].

Lawson, M. (2013). The Wrong Way. *The Guardian,* Guardian 2. 11th July 2013. 19.

Logan, B. (2009). Transfer Window. *The Guardian*, Media Guardian, 6th April 2009, 2.

MacKenzie, K. (2000). Let's lose the tea leaves. *The Guardian,* The Media Guardian. 4th September 2000. 6.

McLean, G. (2009). How do I rate thee? Let me count the ways…. *The Guardian,* Media Guardian. 1st June 2009. 3.

Morinan, G. (2010a). Are men funnier? Here are the vital statistics. *Chortle.* http://www.chortle.co.uk/correspondents/2010/09/30/11838/are_men_funnier%3F_here_are_the_vital_statistics [Accessed 05/04/12].

Morinan, G. (2010b). Are Comedy Reviewers Sexist? *Chortle.* http://www.chortle.co.uk/steveimages/dataanalysis.pdf [Accessed 05/04/2012].

Moss, S. (2006). Gotcha!. *The Guardian:* Guardian 2. 12th May 2006.

Mullone, L. (2013). Why isn't there any decent comedy on Radio 4? *The Spectator*, 21 September 2013. http://www.spectator.co.uk/spectator-life/spectator-life-culture/9022741/gagging-orders/ [Accessed 15th August 2014].

Neilan, C. (2012). Thompson: BBC has "lessons to learn" over Jubilee. *Broadcast.* 20th June 2012. http://www.broadcastnow.co.uk/news/broadcasters/thompson-bbc-has-lessons-to-learn-over-jubilee/5043532.article [Accessed, 23/06/2012].

Oakley, G. (2009). An appreciation of AI. *Broadcast.* 8th June 2009. 5.

Pidgeon, J. (2008). How we developed great comedy, despite Radio 4. *The Guardian*, Media Guardian, 12th May 2008. 6.

Reynolds, G. (2008). Radio Comedy: the joke's wearing thin. *The Daily Telegraph,* The Arts. 17th June 2008, 24.

Sweney, M. (2014). Netflix gathers detailed viewer data to guide its search for the next hit. *The Guardian.* 24th February 2014. 29.

Online Publications

BBC Press Office (2011). BBC 'Quality' measures at highest levels in recent years. *BBC Press Release.* 8th July 2011. http://www.bbc.co.uk/pressoffice/pressreleases/stories/2011/07_july/08/quality.shtml [Accessed 27/04/2012].

BBC Press Office (2012). BBC announces shortlist for first ever BBC Audio Drama Awards. *BBC.* 9th January 2012. http://www.bbc.co.uk/mediacentre/latestnews/2012/audio-drama-awards-shortlist.html [Accessed 30/12/2012]

BBC Trust (2010). Radio Service Review - Radio 3, Radio 4 and Radio 7. *BBC.* http://www.bbc.co.uk/bbctrust/assets/files/pdf/regulatory_framework/service_licences/service_reviews/radio_347/audience_research/audience_research.pdf [Accessed 04/01/2012].

BBC, TV Commissioning (2012). BBC Tariff Prices for Independents. *BBC.*
http://www.bbc.co.uk/commissioning/tv/how-we-work/business-requirements/tariff-ranges.shtml
[Accessed 30/08/2012]

BBC (2011a). BBC Audience Information January to March 2011. *BBC.*
http://downloads.bbc.co.uk/aboutthebbc/reports/pdf/audience_0711.pdf [Accessed 10/5/2012].

BBC (2011b). BBC Full Financial Statements 2010/11. July 2011. *BBC.*
http://downloads.bbc.co.uk/annualreport/pdf/bbc_ar_online_2010_11.pdf [Accessed 16/4/2012].

BBC (2012a). BBC Audience Information October-December 2011. *BBC.*
http://downloads.bbc.co.uk/aboutthebbc/insidethebbc/howwework/accountability/pdf/audience_infor
mation_october_december_2011.pdf [Accessed 10/05/2012].

BBC (2012b). Context - BBC Audience Information October-December 2011. *BBC.*
http://downloads.bbc.co.uk/aboutthebbc/insidethebbc/howwework/accountability/pdf/context_docume
nt_october_december_2011.pdf [Accessed 10/05/2012].

BBC (2013a). BBC Full Financial Statements 2012/13. July 2013. *BBC.*
http://downloads.bbc.co.uk/annualreport/pdf/2012-13/bbc-full-financial-statements-2012-13.pdf
[Accessed 04/04/2014].

BBC (2013b). BBC Audience Information October-December 2012. *BBC.*
http://downloads.bbc.co.uk/aboutthebbc/insidethebbc/howwework/accountability/pdf/summary_audie
nce_information_october_december_2012.pdf [Accessed 19/8/2013].

Goddard, G. (2010). BBC Trust: Independent Radio Productions Commissioned by the BBC. *BBC.*
http://www.bbc.co.uk/bbctrust/assets/files/pdf/our_work/radio_supply/independent_report.pdf
[Accessed 24/11/11].

Hall, S., Pettigrew, N., Skinner, G., Stockley, R., Latter, J. and Cornick, P. (2008). The Audience's View on
the Future of Public Service Broadcasting. *Ofcom.*
http://stakeholders.ofcom.org.uk/binaries/consultations/psb2_1/annexes/annex5.pdf [Accessed
24/11/11]

RAJAR (2011). Making sense of the differences between audio streaming stats and RAJAR listeners.
RAJAR. http://www.rajar.co.uk/docs/news/2011_12_Streaming.pdf. Published Wednesday 21st
December 2011 [Accessed 09/03/2012].

RAJAR (2012a). Key Facts. *RAJAR.* www.rajar.co.uk/docs/about/RAJAR_key_facts.pdf [Accessed
14/04/2012].

RAJAR (2012b). RAJAR diary example page. *RAJAR.*
www.rajar.co.uk/docs/about/RAJAR_diary_example_page.jpg [Accessed 14/04/2012].

Sony, Radio Academy Awards (2012). Production Awards: Categories: Best Comedy. *Radio Academy
Award*s. http://www.radioacademyawards.org/categories. [Accessed 07/06/2012].

Van Meurs, L. (2008). Television Appreciation Research. *GfK.*
http://www.gfkaudimetrie.be/GfK%20TV%20Day%20-%20Appreciation%20Research.pdf [Accessed
26/04/2012].

Appendix

JON HOLMES: *Here at Listen Against, we very much welcome your comments about anything you've heard on BBC radio in the past seven days. But while you're not forgetting that, also don't forget that you can now send in your 'tuts' while we're on air. If you've been irritated or mildly vexed by something you've heard all you have to do is press the red button on your DAB digital radio, – or if you've got an old one, the FM button twice – and you can register a tut on our tut line.*

Listen Against, Series 1, episode 1

Contained in the Appendix are further sections that, although not essential reading for the dissertation, may be of interest to the reader. The first part gives information regarding a survey undertaken alongside this research, which, though it proved to be of no direct relevance to the hypothesis, still provides insight as to the continuing importance of BBC radio comedy to comedy talent development. Next is a brief section discussing how terminology may be artificially inflating the perceived robustness of AI figures. Then, a list of all variables held on Pulse respondents is listed for reference. Next presented is a section that transcribes a satirical comedy piece shared amongst independent radio comedy producers that illustrates the distaste for the Radio 4 commissioning process felt by some. Finally, a list of original TX dates of radio programmes mentioned in this document is listed.

Appendix 1 – 2012 Survey of Comedians

In a tweedy way, there's something momentous and sacred about BBC Radio Comedy. It's a world in which new performers and writers, aspirant and broke, desperate to turn promising Edinburgh Fringes into ways to pay the rent, are thrust together with some of the grand old men of comedy… This is the place where pipe-smoking and ale-drinking sit alongside surrealism and satire – a perfect environment for a conservative who thinks the world needs to change. I love working in radio: it's quick, you don't have to learn your lines, they always give you sandwiches and you're encouraged to go to the pub afterwards. I feel I belong there, and I'm proud of that feeling.

(Mitchell, 2012, p146)

Throughout 2012, over 100 comedians were surveyed as part of this research project, in order to ascertain whether they considered BBC radio to be an important aspect of their careers. With the increasing ease of self-broadcasting, there may be a risk that radio comedy could become a moribund medium in terms of development, as the money available is a fraction of that available for of television.

The survey was basically a convenience sample but with deliberate upweighting of female respondents, there being fewer women on the circuit than men. A wide range of ages and experience was sought and included established names such as Stewart Lee, Andy Zaltzman, Shappi Khorsandi and Simon Evans, alongside up-and-coming names such as Matt Richardson, Humphrey Ker, Chris Stokes and Suzy Bennett. The years in the business of the respondents ranged from a couple of months to over 30 years, and ages ranged from those in their early 20s to over 65s. Respondents were found mainly from London clubs, the Leicester Comedy Festival and the Edinburgh Fringe Festival.

The survey included the question 'Which routes do you think will be important in developing your career?' A number of options were offered and this allows us to see how BBC radio fares against other media and platforms:

Figure 147 2012 Survey of UK Comedians (n=105) – 'Which routes do you think will be important in developing your career?'

% of respondents who selected each of the groups:	club gigs	comedy fests	touring	corporate	podcasts	social networking	PR / marketing	Online	BBC radio	commercial radio	BBC television	Commercial television	film	other
All respondents	78%	84%	65%	30%	41%	60%	52%	69%	70%	29%	73%	54%	41%	0%
Men	74%	81%	60%	26%	38%	54%	44%	65%	67%	18%	74%	51%	29%	0%
Women	88%	91%	76%	39%	48%	73%	70%	76%	76%	52%	73%	61%	67%	0%
Older (born 1980 or earlier)	78%	80%	67%	29%	36%	55%	44%	64%	65%	25%	65%	51%	33%	0%
Younger (born since 1981)	78%	88%	62%	32%	46%	66%	62%	74%	74%	32%	82%	58%	50%	0%

Across all of the respondents, 70% claimed that BBC radio is still an important route for their career progression. This is slightly higher for women (n=33) at 76% versus 67% of the men, and higher for the younger respondents (n=50) versus the older at 74% versus 65%. However, without a temporal view, we cannot see if this is increasing or decreasing; undertaking the survey again in the future might indicate the trajectory of the importance of BBC radio to the comics. What we can do from the above figure is compare to other media. BBC radio is only rated lower (as a proportion) than club gigs, festivals and BBC television. It is still rated more highly than online and social networking. However, a further question asked which route was thought most important to them and only 3 of the 105 respondents said that radio was the platform of primary importance to them. This may indicate that, while comics do rate radio highly, most might not worry too much if it were not an option.

Appendix 2 – The 'Out Of Ten' Scale Misnomer

MICHAEL IAN BLACK: *Last week we started a rating system. It's a very simple rating system; it's one to ten. One being the worst, ten being the best.*
TOM CAVANAGH: *Should we have zero on the list? Is anything going to get a nought? I think we should go zero to ten.*
MICHAEL: *OK, that's fine. I'll change the system.*
TOM: *If you had parsnips and gravy as a snack you might not even want to give it a zero.*
MICHAEL: *Depends on the gravy.*

(Mike and Tom Eat Snacks Podcast no 2 - released 01/12/2011)

It is worth noting that the term 'out of ten' is a term often used for both 0-10 and 1-10 scales. This is carried over to the calculated AI scores. Since 1992 BBC appreciation has been measured on a 1-10 scale and, as there is no zero on the initial scale, it is impossible to give a programme a score of under '1' and as such describing the score of 'out of ten' is incorrect.[585] Likewise, describing AIs as 'out of one-hundred', or even being described as a percentage is clearly wrong. 'Out of ten' and 'out of one-hundred' are used within the BBC, by experts, in the press and even in academic papers; for example:

Danaher and Lawrie, 1998, p55:

While the survey methodology has continued to be diary-based, the measurement scale has changed to **marks out of 10**.

Kent, 2002, p250-253:

The ROM collects Reaction Indices to each programme based on a **marks out of ten** evaluation.

North, 2010:

*So what we did there was to take into account that a 10 out of 10 is something that was much more significantly better than an 8 **out of 10** or 9 **out of 10**.*

BBC Gateway Audience Portal, 2011:

We ask people to rate all the programmes they have seen on a ten-point scale – we then scale the score up to be **out of 100**.

BBC Press Office, 2011:

The audience Appreciation Index (AI) – a score **out of 100**.

Neilan, 2012:

'But please see that in the context of people saying this was a show that deserved 8.2 **out of 10**' [Mark] Thompson said.

BBC, 2012b, p4:

A sample of the UK public is asked to give the BBC a score **out of 10**.

Chapman, 2014:

BBC1's AI score dropped to 81.5 **out of 100** in 2013/14.

This may seem to be just pedantry, but the incorrect terminology may imply a level of statistical robustness that does not exist. Really, when one considers the semantics of the phrase, 'out of' has the implication that there could be a zero point. De Vaus (2002, p103) explains it correctly when defining an '*out of 10*' scale, clarifying that the respondent can 'give any score between 0 and 10'.

One can argue that when saying 'out of ten' for appreciation scores, the implication is that not only is the data interval, but also *ratio*. Also, in expressing the AIs as up to 100, there is a layman's view that it *feels like* a percentage. AIs can even be completely misrepresented, as in Hendy (2007, p71) who claims that appreciation indices (RIs as they were in the example) were measured on a 1-100 scale which, in fact, they

---

[585] The range of AI scores in the current system is between 10 and 100 as it is calculated by multiplying the mean appreciation scores by 10.

have never been: either 10-100 or 0-100 being correct depending on the point in BBC history. This error illustrates that care needs to be taken when explaining the scales behind appreciation scores and AIs lest the interpreter be misled: '...care should be taken in attributing to the categories of an ordinal scale an arithmetic quality that the scoring seems to imply' (Bryman and Cramer, 1994, p65).

Appendix 3 – All Variables Held on Pulse Respondents

Tommy Handley – MURGATROYD:	*Well, it ought to be a good act tonight.*
Ronald Frankau WINTERBOTTOM:	*It won't be appreciated by the audience in the studio.*
MURGATROYD:	*Why not?*
WINTERBOTTOM:	*Well, look at them - there's a fellow there eating a bit of cheese.*
MURGATROYD:	*Yes. He got in on the strength of that. They certainly don't look a very intelligent lot.*
WINTERBOTTOM:	*No. Obviously friends of the producer.*

(***ITMA***: Took, 1981, p31-32)

Email correspondence, Joanne Tingay jtingay@trponline.co.uk 23/5/13

RespondentID
Weighting
ABGenreRadio_News_01
ABGenreRadio_News_Factual _01
ABGenreRadio_Regional_01
EnglishRegion
people living in house at least 9 months a year
Country
Social class
TV ownership type
Gender respondent
Age respondent
marital status
speaking Welsh
respondent is main income earner
Number of television sets
Children 15 or under
Number of Children in household (age group: 6-12)
Number of Children in household (age group: 13-15)
Nationality
Ethnicity
Disability
Difficulty in hearing
Difficulty in seeing
Other physical difficulties
Other health problems
Sexuality
Number of children 0-15
BBC Region by post code
BBC sub areas by post code
Shorter Ethnicity band
HD/SD platform type
Mosaic 2011
Number of Children in household (agegroup: 0-3)
Number of Children in household (agegroup: 4-5)
Frequency internet access
Number of adults in household
TV platform: Main set
TV platform: 2nd set
TV platform: 3rd set
TV platform: 4th set
PVR: Main set
PVR: 2nd set
PVR: 3rd set
PVR: 4th set
HD TV: Main set
HD TV: 2nd set
HD TV: 3rd set
HD TV: 4th set
HD receiver: Main set
HD receiver: 2nd set

HD receiver: 3rd set
HD receiver: 4th set
Smart TV: Main set
Smart TV: 2nd set
Smart TV: 3rd set
Smart TV: 4th set
Digital Households
Satellite Households
Cable Households
Freeview Households
Freeview ONLY Households
Other television Households
PVR Households
Sky PVR Households
Virgin PVR Households
Smart television Households
Religion
Device owned: a home computer/PC/laptop
Device owned: a mobile phone
Device owned: a tablet device
Device owned: a television set that can connect to the Internet
Device owned: a Blu-ray DVD player
Device owned: a games console connected to your TV
Device owned: an e-book reader e.g., Kindle, Sony e-book, Kobo etc.
Device owned: a portable games console
Device owned: another mobile device
Device owned: other
Device owned: Don't know
Device owned: No devices owned
Own personal mobile phone
Type of mobile
Mobile OS: iOS
Mobile OS: RIM
Mobile OS: Android
Mobile OS: Windows
Mobile OS: Other operating system
Mobile OS: Smartphone
Mobile OS: Featurephone
Mobile OS: Internet enabled phone
Mobile OS: Other/Don't know
How access internet last month: a home computer/PC/laptop
How access internet last month: a work computer/PC/laptop
How access internet last month: a mobile phone
How access internet last month: a tablet device, e.g., iPad
How access internet last month: a television set that can connect to the Internet
How access internet last month: a Bluray DVD player
How access internet last month: a games console connected to your TV
How access internet last month: an e-book reader e.g., Kindle, Sony e-book, Kobo etc.
How access internet last month: a portable games console
How access internet last month: another mobile device
How access internet last month: Other
How access internet last month: Don't know
How access internet last month: None
Qualify for Mobile Survey
TV licence good Value
Sky Subscription Value
Virgin Media Value

Appendix 4 – Satire about Radio Comedy Commissioning

*Your chances of actually getting your radio sitcom made
are slimmer than an anorexic stick-insect's slim-fitting drainpipe trousers.*

(Hawes, 2004, p73)

http://www.youtube.com/watch?v=BLPvAFlVPqQ. Added to YouTube: February 26, 2010

PRODUCER: I have an idea for a new comedy for BBC Radio 4.
COMMISSIONING EDITOR: It doesn't really leap off the page.
P: Of course it doesn't. It's a radio piece you talentless bint.
C: To be considered, it had to be submitted nine months ago.
P: It was. You didn't reply.
C: I was busy spearheading a new initiative to encourage people to write for Radio 4. Who is in it?
P: Eric Morecambe, Tommy Cooper, Lenny Bruce, Bill Hicks and Aristophanes.
C: I don't know them. They aren't really Radio 4 names.
P: That's because they're good.
C: The script isn't very Radio 4.
P: That's because it's good. It has things like actual jokes in it.
C: The script doesn't work for me.
P: Surely the point is not what makes you laugh, surely the point is that you are here to commission what makes other people laugh.
C: What is that word you keep using? Laugh?
P: Your job is to read the market, not inflict your personal taste on the nation.
C: I don't have any taste.
P: Now we're making progress. Look, the point is, we're professional comedy writers and performers. We can go out and make audiences laugh. What makes you think you know better?
C: Look, we are not looking for anything new at the moment.
P: Name me one other medium or creative outlet that actively discourages new stuff. Any art form that doesn't innovate will die.
C: Your series is set in the middle ages. We are not commissioning period stuff.
P: Like Monty Python and the Holy Grail? Ok, I can change the setting to the Renaissance.
C: That doesn't really work.
P: Like Blackadder? Ok, what about in a prison?
C: We're not commissioning prison stuff at the moment.
P: Like Porridge? How about in the future?
C: Future stuff doesn't really work.
P: You mean like Red Dwarf? The Hitchhiker's Guide to the Galaxy and those other failures? Don't you realise that a good comedy has great characters and jokes in it? The setting isn't important.
C: We have no sense of humour and no feel for the medium. That's why we have erected ludicrously arbitrary frameworks by which to judge programmes.
P: What do you want?
C: I'll know it when I see it.
P: You are destroying radio and comedy. You have been here for decades. When will you and the other morons like Caspari fuck off?
C: Never ever ever. I am like a banker. No matter how shit we are you can't sack us. Ha ha ha ha ha.
P: Fucking shitty shitty bollocks.

Appendix 5 – List of Radio Programmes and Dates of Broadcast

Spike Milligan: Radio, where the pictures are better because they happen on the other side of your eyes.
(Farnes, 1997, p118)

The list below gives the original TX dates for all BBC radio programmes mentioned throughout the document. It is in alphabetical order.

Title	Radio TX dates
15 Minute Musical	2004-
A Charles Paris Mystery	1984-
A Comedy of Danger	1924
Absolute Power	2000-06
After Henry	1985-89
Andrew Lawrence: How Did We End Up Like This?	2012
Another Case of Milton Jones	2005-11
Armando Iannucci's Charm Offensive	2005-08
Ayres on the Air	2004-14
Band Waggon	1938-39
Believe It!	2012-14
Beyond Our Ken	1958-64
Bigipedia	2009-11
Black-outs for the Black-out	1940-?
Brian Gulliver's Travels	2001-12
Bridget Christie Minds the Gap	2013-
Cabin Pressure	2008-14
Chain Reaction	1991-
Clare in the Community	2004-
Clitheroe Kid	1958-72
Count Arthur Strong's Radio Show!	2005-
Dead Ringers	2000-14
Down the Line	2006-
Ed Reardon's Week	2005-
ENSA Half Hour	1940-44
Everyone Quite Likes Justin	2010-12
Flight of the Conchords	2004
Garrison Theatre	1939-41
Genius	2005-08
Goodness Gracious Me	1996-98
Hancock's Half Hour	1954-59
Happy Mondays: Pappy's Fun Club	2008
Hey Rrradio!!!	1988
I'm Sorry I Haven't A Clue	1972-
I'm Sorry, I'll Read That Again (ISIRTA)	1964-68
I've Never Seen Star Wars	2008-
In All Directions	1952-55
In and Out of the Kitchen	2011-
It's That Man Again (ITMA)	1939-1949
Jeremy Hardy Speaks to the Nation	1993-
John Finnemore's Souvenir Programme	2011-
Just a Minute	1967-
Kevin Eldon Will See You Now	2012-
King Pins of Comedy	1940-?
Knowing Me, Knowing You	1992-93
Les Kelly's Britain	2010-11
Life with the Lyons	1950-61

Listen To This Space	1967
Little Britain	2000-02
Look Away Now	2007-10
Mark Steel's in Town	2009-
Mark Thomas: The Manifesto	2009-
Mark Watson's Live Address to the Nation	2007-11
Meanwhile, It's Will and Greg	2012
Meet David Sedaris	2010-
Meet the Huggetts	1953-61
Mr and Mrs Smith	2010-12
Mr Blue Sky	2011-12
My Teenage Diary	2009-
Nurse	2014
On the Hour	1991-92
On the Town with the League of Gentlemen	1997
People Like Us	1995-97
Radio Active	1980-88
Radio Radiance	1925-?
Room 101	1992-94
Round the Horne	1965-68
Sean Lock's 15 Storeys High	1998-2000
Sketchorama	2012-
Son of Cliché	1983-84
Sparkhill Sound	2011
Steptoe and Son	1966-1972
Susan Calman is Convicted	2013-
Take It From Here	1948-60
Take It From Here	1948-60
That Mitchell and Webb Sound	2003-
The 7:30 Revue	1925-?
The Adam and Joe Show	2007-11
The Bleak Old Shop of Stuff	2007-12
The Bob Hope Show	1942-?
The Boosh	2001
The Burkiss Way	1976-80
The Cabaret Upstairs	1986-88
The Cowards	2007-08
The Goon Show	1951-60
The Guns of Adam Riches	2012-14
The Hitchhiker's Guide to the Galaxy	1978-2005
The Jack Benny Show	1942-?
The Mary Whitehouse Experience	1989-90
The Men from the Ministry	1962-77
The Million Pound Radio Show	1985-1989
The Museum of Curiosity	2008-
The National Theatre of Brent's Iconic Icons	2009-11
The Navy Lark	1959-77
The News Huddlines	1975-2001
The News Quiz	1977-
The Now Show	1998
The Pig and Whistle	1938-44
The Secret World	2008-14
The Simon Day Show	2011-12
The Spam Fritter Man	1976
The Stanley Baxter Playhouse	2006-
The Sunday Format	1996-2004
The Unbelievable Truth	2006-
They Think It's All Over	1992-93
Tim Key's Late Night Poetry Programme	2012-
Tina C's Global Depression Tour	2012

To The Manor Born	1997
Tom Wrigglesworth's Open Letters	2011-12
Tonight	2011-12
Up The Garden Path	1987-93
Variety Bandbox	1944-53
Warhorses of Letters	2011-14
Week Ending	1970-88
Whack-O	1961-63
When the Dog Dies	2010-14
Whose Line Is It Anyway?	1988
Wordaholics	2012
You'll Have Had Your Tea	2002-07