THE UNIVERSITY OF WINCHESTER

Faculty of Business, Law and Sport

EXPECTANCIES, MARKING AND FEEDBACK IN UNDERGRADUATE STUDENT ASSESSMENT

Jo Batey

ORCID Number: 0000-0002-0111-4056

Doctor of Philosophy

November 2018

This Thesis has been completed as a requirement for a postgraduate research degree at the University of Winchester

DECLARATION AND COPYRIGHT STATEMENT

Declaration

No portion of the work referred to in the Thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

I confirm that this Thesis is entirely my own work

Copyright

DEDICATED TO MY MUM

1947-2009

You can shed tears that she is gone
Or you can smile because she has lived.
(David Harkins – She is Gone)

ACKNOWLEDGEMENTS

I would like to thank my supervisory team; Professor Eric Anderson, Professor Tansy Jessop and Doctor Janice de Sousa. In particular I would like to thank Eric for making this all seem possible at times when I definitely did not! For ALWAYS being there at the end of the phone, and for providing feedback on drafts so promptly. Special thanks also to Tansy who spent a lot of time with me towards the end of this marathon and who expertise in the pedagogical aspects of the project were invaluable. I enjoyed our time working through draft chapters. Your feedback and criticality have undoubtedly made my writing stronger.

Nonetheless, despite the contributions of the team above I would not have made it through this process without the love and support I have received at home from the indomitable Ruth Ashton. When I said, "I just have to do a bit more on my PhD", she never once complained or made me feel guilty, and has been my rock throughout. When my health deteriorated halfway through this journey and I had even less capacity to help at home, she just did more, holding everything together like the champion she is. I am truly grateful for her generous heart, her love and her understanding.

UNIVERSITY OF WINCHESTER

ABSTRACT

Expectancies, Marking and Feedback in Undergraduate Student Work

Jo Batey

ORCID Number: 0000-0002-0111-4056

Doctor of Philosophy

November 2018

This research explores how expectancies related to student gender and ethnic origin (as derived from the name) might impact on the feedback received on assignments. Calls for anonymous marking on the basis of biased marking practices according to gender and ethnicity have been long standing and widespread. However, research in this area has generally lacked methodological rigour, produced equivocal findings, and solely been concerned with grade bias. Therefore, using a mixed methods research methodology, this thesis focused on the feedback provided. Sixty sports academics agreed to mark and provide feedback on two first year undergraduate student essays. In-text feedback was text-to-text transcribed, coded utilising an existing analytical framework and content analysed. Summary feedback was hierarchically content analysed using established guidelines. In-text feedback provided more evidence for expectancy effects specifically in relation to White British females when compared to White British males and Asian females when compared to Asian males. There was also evidence that non-White British names were provided with less useful and educative feedback than their White counterparts. Summary feedback revealed less evidence of expectancy effects at work and sometimes contradicted the in-text feedback findings. Findings are discussed in relation to feedback quality, marker variability, marking as a social practice and the anonymous marking debate.

Keywords: Expectancies, Bias, Feedback, Gender, Ethnicity, Assessment, Marking

LIST OF CONTENTS

Page

DECLARATION AND COPYRIGHT STATEMENT.. 1
DEDICATION.. 2
ACKNOWLEDGEMENTS 3
ABSTRACT.. 4
LIST OF FIGURES 9
LIST OF TABLES 10
1 INTRODUCTION 12
1.1 Background to the Thesis 12
1.2 Introduction to Expectancy Effects 14
1.3 Introduction to Marking Processes 17
1.4 Introduction to Feedback.. 19
1.5 Purpose and Research Questions 20
2 REVIEW OF LITERATURE 22
2.1 Chapter Outlines 22
2.2 CHAPTER ONE: EXPECTANCIES.. 22
2.3 Olson, Roese and Zanna’s (1996) Model of Expectancy Processes 23
2.3.1 Expectancy Formation 25
2.3.2 Direct Experience 26
2.3.3 Indirect Experience 26
2.3.4 Other Beliefs 27
2.4 Warr and Knapper’s (1968) Schematic Model of Person Perception 27
2.5 Approaches to Information Processing.. 30
2.5.1 Schema-driven Processing 30
2.5.2 Data-driven processing 32
2.5.3 Dual processing 32
2.5.4 Moderators of schema and data-driven processing 35
2.6 The Automaticity Debate 38
2.7 PROPERTIES OF EXPECTANCIES.. 40
2.7.1 Certainty 41
2.7.2 Accessibility 42
2.7.3 Explicitness 44

2.7.4 Importance 45
2.8 EXPECTANCY EFFECTS 46
2.8.1 Cognitive consequences 46
2.8.2 Impact on attention and encoding 46
2.8.3 Affective Consequences 52
2.8.4 Behavioural consequences 53
2.9 CHAPTER TWO: MARKING PROCESSES 58
2.9.1 Perceptions of Assessment 58
2.9.2 Reliability of Marking 58
2.9.3 Halo Effects 62
2.9.4 Gender bias 66
2.9.5 Ethnic Bias 71
2.9.6 Anonymous Marking 75
2.10 CHAPTER THREE: FEEDBACK 81
2.10.1 Classifying Feedback 82
2.10.2 Positive and Negative Feedback Practices 86
2.10.3 Student Perceptions of Feedback 88
2.10.4 Biased Feedback 92
2.10.5 Biased Feedback and Gender 93
2.10.6 Biased Feedback and Ethnicity 94
3 RESEARCH METHODS 100
3.1 Introduction 100
3.2 Introduction to Mixed Methods........ 100
3.3 Methodological Philosophy........ 103
3.3.1 Philosophical Background 103
3.4 Mixed Methods Research Designs 106
3.5 Research Design 107
3.5.1 Participant Recruitment 107
3.5.2 Participants 107
3.5.3 Development of the Assessment pack 108
3.5.4 Procedures 109
3.5.5 Process of Data Analysis 110
3.5.6 Analysis of In-Text Feedback 112
3.5.7 Hyatt's (2005) Feedback Classification System 115

3.5.8 Analysis 1: Gender 117
3.5.9 Analysis 2: Ethnicity 119
3.5.10 Analysis 3: Gender and Ethnicity 120
3.5.11 Analysis of Summary Feedback 122
4 RESULTS........ 124
4.1 In-Text Feedback 124
4.2 In-Text Feedback Analysis 1: Perceived Student Gender........ 126
4.2.1 Analysis 1.1: White British Male vs. White British Female 126
4.2.2 Analysis 1.2: Asian Male vs. Asian Female 130
4.2.3 Analysis 1.3: Chinese Male vs. Chinese Female 135
4.3 In-Text Feedback Analysis 2: Perceived Student Ethnicity........ 137
4.3.1 Analysis 2.1: Male Ethnicities 137
4.3.2 Analysis 2.2: Female Ethnicities 144
4.4 In-Text feedback Analysis 3: Perceived Student Gender and Ethnicity 147
4.4.1 Analysis 3.1: White British Male vs. Asian Female 148
4.4.2 Analysis 3.2: White British Male vs. Chinese Female 150
4.4.3 Analysis 3.3: Asian Male vs. Chinese Female 153
4.4.4 Analysis 3.4: Asian Male vs. White British Female 157
4.4.5 Analysis 3.5: Chinese Male vs. White British Female 161
4.4.6 Analysis 3.6: Chinese Male vs. Asian Female 165
4.5 Summary of In-text Feedback Results........ 170
4.5.1 Developmental Feedback 170
4.5.2 Structural Feedback 172
4.5.3 Stylistic Feedback 173
4.5.4 Stylistic Comments, Corrections and Emphasis 173
4.5.5 Content-Related Feedback 180
4.6 Summary Feedback 183
4.6.1 Summary Feedback Differences: Gender 186
4.6.2 Summary Feedback Differences: Ethnicity 189
4.6.3 Summary Feedback Differences: Gender and Ethnicity 191
4.6.4 Summary of Summary Feedback Results 192
4.7 Reflective Comments 193
5 GENERAL DISCUSSION........ 195
5.1 General reflections on the feedback........ 196

5.2 Marker Variability 198
5.3 Expectancy Effects 202
5.4 Key Results 205
5.4.1 Key Results: Gender 206
5.4.2 Key Results: Ethnicity 208
5.4.3 Key Results: Gender and Ethnicity 210
5.5 Key Results: Overview 211
5.6 Tackling Expectancies and Bias 215
6 LIMITATIONS 219
7 CONCLUSION 221
8 REFERENCES 223
9 APPENDICES 253
9.1 APPENDIX I: Brown, Gibbs, & Glover's (2003) Amended Feedback System 253
9.2 APPENDIX II: Kumar and Stracke's (2007) Feedback classification system 254
9.3 APPENDIX III: Hyatt's (2005) Feedback classification system 255
9.4 APPENDIX IV: Assessment Pack 258
9.5 APPENDIX V: Initial coding 273
9.6 APPENDIX VI: Hyatt's (2005) Amended Feedback classification system 275
9.7 APPENDIX VII: Steps for a hierarchical content analysis 278
9.8 APPENDIX VIII: In-text Feedback Results Tables 279
9.9 APPENDIX IX: Control Group Analyses for In-Text Feedback 303

LIST OF FIGURES Page

Figure 1 Olson, Roese and Zanna’s (1996) Model of Expectancy Processes…………… 26

Figure 2 Warr and Knapper’s (1968) Schematic Model of Person Perception………… 29

Figure 3 Participants Teaching Experience ……………………………………………………………. 108

Figure 4 Annual Marking Loads ……………………………………………………………….............. 108

LIST OF TABLES Page

Table 1 Basic characteristics of four worldviews used in research................................. 105

Table 2 Identification of the four worldviews used in this research 106

Table 3 Example of initial transcription process 114

Table 4 Example of researcher triangulation regarding coding 115

Table 5 Ambiguity of the coding process 117

Table 6 Analysis 1.1: In –text feedback: Gender (White British) 119

Table 7 Analysis 1.2: In-text feedback: Gender (Asian) 119

Table 8 Analysis 1.3: In-text feedback: Gender (Chinese) 119

Table 9 Analysis 2.1: In-text feedback: Ethnicity (Male) 120

Table 10 Analysis 2.2: In-text feedback: Ethnicity (Female) 120

Table 11 Analysis 3.1: In-text feedback: Gender and ethnicity (White British male and Asian female) 121

Table 12 Analysis 3.2: In-text feedback: Gender and ethnicity White British male and Chinese female) 121

Table 13 Analysis 3.3: In-text feedback: Gender and ethnicity (Asian male and Chinese female) 122

Table 14 Analysis 3.4: In-text feedback: Gender and ethnicity (Asian male and White British female) 122

Table 15 Analysis 3.5: In-text feedback: Gender and ethnicity (Chinese male and White British female) 122

Table 16 Analysis 3.6: In-text feedback: Gender and ethnicity (Chinese male and Asian female) 122

Table 17 Analysis 1.1: In-text feedback classifications for White British male versus White British female (Experimental) 127

Table 18 Analysis 1.2: In-text feedback classifications for Asian male versus Asian female (Experimental) 131

Table 19 Analysis 1.2: Developmental feedback differences across subcategories for Asian Male and Asian Female 135

Table 20 Analysis 1.3: In-text feedback classifications for Chinese male versus Chinese female (Experimental) 136

Table 21 Analysis 2.1: In-text feedback for male ethnicities (Experimental)....................139

Table 22 Analysis 2:1 -Stylistic feedback differences in punctuation across male ethnic groups ……140

Table 23 Analysis 2.1: Stylistic feedback differences in syntax/word order/grammar across male ethnicities ……140

Table 24 Analysis 2.1: Presentational feedback differences across male ethnicities ……142

Table 25 Analysis 2.1: Developmental feedback differences across male ethnicities …..143

Table 26 Analysis 2.2: In-text feedback classifications for female ethnicities ……145

Table 27 Analysis 2.2: Content-related feedback differences for female ethnicities…….147

Table 28 Analysis 3.1: In-text feedback classifications for White British male and Asian female (Experimental) ……149

Table 29 Analysis 3.1: Stylistic feedback differences between White British male and Asian female across conditions ……150

Table 30 Analysis 3.2: In-text feedback classifications for White British male and Chinese female (Experimental) ……152

Table 31 Analysis 3.3: In-text feedback classifications for Asian male and Chinese female (Experimental) ……154

Table 32 Analysis 3.4: In-text feedback classifications for Asian male and White British female (Experimental) ……158

Table 33 Analysis 3.4: Syntax/word order/grammar feedback differences between Asian male and White British female across conditions ……159

Table 34 Analysis 3.5: In-text feedback classifications for Chinese male and White British female (Experimental) ……162

Table 35 Analysis 3.5: Syntax/word order/grammar feedback differences between Chinese male and White British female ……163

Table 36 Analysis 3.6: In-text feedback classifications for Chinese male and Asian female (Experimental) ……166

Table 37 Analysis 3.6: Stylistic feedback differences between Chinese male and Asian female ……167

Table 38 Analysis 3.6: Syntax/word order/grammar feedback differences between Chinese male and Asian female across conditions ……169

Table 39 Hierarchical content analysis results for summary feedback ……184

Table 40 Number of comments for summary feedback themes (control) ……186

Table 41 Number of comments for summary feedback themes (experimental) ……187

1 INTRODUCTION

1.1 Background to the Thesis

> If we lived in a perfect world, students would be able to put their name on their coursework. Students would not have to fear that their work would be marked any differently based on their gender, sexuality or race. Unfortunately we don't live in that world (Wes Streeting, NUS Vice-President Education. Cited in Baty, 2007 p.1-2).

The National Union of Students (NUS) have campaigned for anonymous marking since 1999, claiming that it would safeguard Higher Education Institutions (HEIs) against claims of expectancy-based bias within the marking process. They have support from a number of high profile bodies. These include, the Quality Assurance Agency (QAA), who oversee the quality of higher education provision in the UK, The Equality and Human Rights Commission, and the Association of University Teachers (AUT). The initial claims of bias in marking practices made by the NUS were significant enough to make National news headlines. Newspaper coverage named a handful of HEIs who already anonymised students' work and subsequently suggested that others should do the same (Smithers, 1999). However, the initial call for anonymous marking only extended to examinations. The NUS drive only gained significant momentum in 2008 with the release of a paper entitled 'Mark my words not my name', where they called for anonymity across all written assignments. Within this paper the NUS claimed discrimination was in operation across HEIs on the basis of student gender and ethnicity. Specifically, they cited research which demonstrated females and non-white students of various ethnicities received lower grades. They further claimed that 44% of Students' Unions believed marking was biased. The paper sent shockwaves around the sector with many universities immediately employing anonymous marking in an attempt to be perceived as non-discriminatory and unbiased.

Discussion surrounding the adoption of anonymous marking has been far from harmonious, and despite growing pressure from the sector many have staunchly resisted (e.g., Whitelegg, 2002; Brennan, 2008; Owen, Stefaniak, & Corrigan, 2010). Arguments against its employment have been both pedagogical and practical. Pedagogically much of the focus has surrounded issues of feedback. Student anonymity is said to prevent tutors from being able to provide the individualised feedback which students prefer and find most helpful (Bols, 2013; Birch, Batten & Batey, 2015; Pitt & Winstone, 2018). Furthermore, Whitelegg (2002) suggested that it would interrupt the feedback loop and make it harder to identify weaker students and offer help. Practically, there were concerns about the additional administrative duties that the process of anonymising work would bring, but more substantive was the argument about the impossibility of anonymising specific forms of assessment (e.g., presentations, performance pieces).

Furthermore, even when the type of assignment allowed for anonymous marking (e.g., dissertation, project) it might still be possible to deduce the identity of the author. For example, a student might have been working closely with a tutor on a particular project or booked a tutorial to discuss the inclusion of a specific case-study in their assignment.

Alternatively, proponents of anonymous marking claim that it would safeguard both staff and students. For example, it can reassure students that concerns can be aired without the fear of reprisal (Brennan, 2008). Importantly, it might also stimulate students to be more actively involved in seeking feedback (Whitelegg, 2002). Finally, Owen at al. (2010) claimed that it might reduce the perception of bias, although recent research has shown no evidence of this perception among students (Pitt & Winstone, 2018).

Nonetheless, if changes to HEI policy regarding anonymous marking are to be enforced (at the moment the practice is strongly recommended), it is imperative they are evidence-driven and underpinned by a sound theoretical framework. The immediate decision taken by some HEI's to mark anonymously following the publication of the NUS report could be described as premature. Kowtowing to calls for anonymous marking without exploring the evidence upon which these calls were based demonstrated a lack of academic and critical engagement. This was surprising given that this report demonstrated a capacity to challenge HEI policy at the highest level. Had this engagement taken place it would have demonstrated the following:

1) Whilst the NUS cited several studies which demonstrated bias according to gender and ethnicity (e.g., Bradley, 1984; Goddard-Spear, 1984; Belsey, 1988) they omitted many others which did not show this bias (e.g., Kehle, Bramble, & Mason, 1974; Swim, Borgida, Maruyama, & Myers, 1989; Perry-Langdon, 1990; Newstead & Dennis, 1990, 1994). Therefore, rather than research unequivocally supporting the presence of bias, a more balanced appraisal would be that results are equivocal.
2) Much of the research that the NUS based their claims upon are derived from old archival data obtained from university marking records (e.g., University of Wales, University of Glasgow's Dental School). While these results provided a valuable insight into potential patterns of bias they also lacked experimental control.
3) The statistic that 44% of Student's Unions believed discrimination and bias existed made headlines within the HEI sector. However, a less publicised statistic revealed that in HEIs where assessment-wide anonymous marking had been employed only 66% of its students were convinced that there was little or no discrimination in assessment. Therefore marking anonymously did not appear to eliminate perceptions of bias.

Furthermore, when the area of marking bias is examined more thoroughly it becomes apparent that much research in the area has lacked experimental rigour and suffered from weak theoretical application. It has been suggested that teacher's expectancies might bias the marks awarded to students but has not detailed how and why those expectancy effects might prevail (e.g. Newstead & Dennis, 1990; Dennis Newstead & Wright, 1996). Additionally, and of critical importance to this PhD, research has been dominated by measuring bias solely on the basis of the mark awarded to the piece of work and has rarely considered the feedback provided. Feedback itself has been extensively researched in recent years, but an examination of the influence of gender and/or ethnicities on feedback has not been undertaken.

Consequently this PhD aims to explore whether knowledge of a student's gender and ethnicity as indicated by the name on the assignment can bias the feedback awarded on their work. This aim was addressed through the following research questions. Do expectancy effects as primed through knowledge of the students;

i) Gender impact upon feedback in a way that suggests biased practice?
ii) Ethnicity impact upon feedback in a way that suggests biased practice?
iii) Gender and ethnicity impact upon feedback in a way that suggests biased practice?

1.2 Introduction to Expectancy Effects

Although controversy permeates many areas of social psychology there are two fundamental truisms within the field. Firstly, that social influence is pervasive, and secondly, that people construct their own reality through the process of interpersonal interaction (Taylor, 1997; Eysenck, 2009). Interestingly, even when such interactions are fleeting, the initial judgements people make when meeting others can have long-lasting effects, and thus affect future interactions between both the perceiver (the observer) and the target (the observed) (Miller and Turnbull, 1986; Fiske & Taylor, 1991; Jussim & Harber, 2005). It is therefore not surprising that the process of perceiving those with whom we interact (i.e., person perception), combined with a) the expectancies that such perception creates and, b) the subsequent impact of these expectancies, has been the focus of research attention since the 1950's (Jones, 1986).

Expectancies are most simply defined as "…beliefs about a future state of affairs" (Olson, Roese, & Zanna, 1996, p.210). This fits well with the general consensus among social psychologists (e.g., Jones, 1986; Fiske & Neuberg, 1990), that when people enter social interactions they actively seek to make sense of them and thus predict how they are likely to progress and conclude.

People's beliefs and expectancies are seen as being central to this sense-making process, since they house existing knowledge structures (or schemas) and are influential in determining the cognitions (i.e., thoughts, knowledge, memories and judgements), affect, and behaviour of the perceiver, and the process and outcome of social interactions (e.g., Darley & Fazio, 1980; Fiske & Taylor, 1984; Snyder, 1984; Jones, 1986). Jussim (1991) also notes that these schemas guide the processing of new information. Therefore once a perceiver has formed an impression, whether accurate or not, it will in all likelihood, guide their future decisions and social interactions.

The work of Olson et al., (1996) emanates from social psychology and provides the most comprehensive model to explore expectancy effects. Their Model of Expectancy Processes is central to this thesis as it examines both expectancy formation (their sources and properties) and effects (their consequences). Much research concerns itself solely with expectancy effects, but expectancy formation is critical to understand as an antecedent to this process. An alternative and much-used model by Warr and Knapper (1968) – the Schematic Model of Person Perception - is also discussed. This model derives from the domain of social cognition within social psychology. It focuses more exclusively on how perceivers process information during interpersonal interactions. Given the advances in the area of social cognition in recent years, and in recognition of essay marking being a complex cognitive task which challenges information processing (Hamp-Lyons & Henning, 1991; Sadler, 2009), its inclusion was considered important.

Expectancy formation and expectancy effects are generally underpinned by two broad approaches to information processing; schema-driven and data-driven (Greenlees, 2007). The approach a perceiver chooses to use is partly guided by the cognitive resources at their disposal. A schema-driven approach to information processing is less resource-heavy for the perceiver and therefore more cognitively efficient (Allport, 1954; Tversky & Kahneman, 1974). Thus under high cognitive load (such as marking a batch of assignments) this might be the default strategy for many perceivers. However, because this processing activates schemas (or categories) about the characteristics of certain *types* of people as opposed to focusing on the *individual* it is prone to cognitive biases in judgment and subsequent expectancy effects. Schema-driven information processing and its reliance upon categories to inform expectancies has led to parallels being drawn between this type of processing and stereotypes.

Stereotypes, expectancies, and bias are distinct, but interrelated concepts. A stereotype can be defined as,

> …a cognitive structure containing the perceiver's knowledge and beliefs about a social group and its members…[and] an important source of expectancies about what the

> group as a whole is like as well as about attributes that individual group members are likely to possess (Hamilton, Sherman, & Ruvolo, 1990, p.36).

Similarly a bias is an inclination or prejudice for or against a person or group (OED [Online]). Expectancies are heavily influenced by stereotypes and biases since beliefs already held by the perceiver can influence the formation of later expectancies about a target (Darley & Fazio, 1980; Hamilton et al., 1990; Macrae & Bodenhausen, 2000). Nonetheless, it is important to recognise that not all expectancies are stereotypes, not all stereotypes and biases influence expectancies, and therefore not all expectancies lead to biased practice. This notwithstanding it is also true that, category activation does create the potential for stereotypic-based expectancies and cognitive biases to arise.

Data-processing is a more resource-intensive type of information processing, since it involves the perceiver integrating each new piece of information about a target in a systematic fashion. Instead of activating preconceptions about category membership, the perceiver forms a stand-alone impression of the target using only the information available to them in the present interaction (Fiske & Neuberg, 1990; Pendry and Macrae, 1994). As such it is considered that cognitive biases and expectancy effects are attenuated when data-driven processing is used (Neuberg & Fiske, 1987). However, the discussion surrounding information processing and the perceiver's choice to engage in one type of processing over another is blurred somewhat by another contentious argument surrounding the concept of automaticity. This debate began more than 60 years ago when Allport (1954) suggested that the world was too complex for the human mind to process all available information and that in order to survive the human brain automatically used schematic processing. Contemporary researchers have considered that things might not be that straightforward (e.g., Bargh, 1994; Wegner & Bargh, 1998; Schwartz, 1998; Bargh, 1999; Macrae & Bodenhausen, 2000), and this argument is examined in more detail later. Furthermore, in addition to cognitive capacity there are numerous alternative moderators of schema and data-driven processing.

Expectancy effects can manifest themselves in three ways cognitively, affectively, and behaviourally (Olson et al., 1996). There are five key elements to the impact of expectancies on cognitive functioning; attention and encoding (perceivers often attend to and recall information that is consistent with their expectancy above that which is inconsistent), memory (perceivers often falsely recall information that matches their expectancies), interpretation (perceivers often interpret information in expectancy-consistent ways), attributions (causes of events are often interpreted in line with expectancies), and counterfactual thinking (perceivers engage in this type of 'what-if' thinking when expectancies are disconfirmed). Affective manifestations of expectancies are less well acknowledged by Olson et al, (1996) but have been considered to be mood state; perceivers in a good mood are prone to more cognitive biases than those in a bad mood (Edwards & Weary, 1993; Bodenhausen, Mussweiler, Gabriel, & Moreno, 2001; Forgas, 2007; Forgas & Laham, 2009), and liking; how much warmth a perceiver feels towards a target (Bradley, 1983; Cardy & Dobbins, 1986). Finally, the most researched expectancy effect has been in the behavioural domain. However, this research attention is not reflected in a wide range of behavioural effects. Instead, most of the focus has been on the self-fulfilling prophecy (SFP) in schools. In this context a SFP refers to how a teachers expectations about a pupil can evoke academic performances from the student which matches the teacher's expectations. It has been examined since the late 1960's (e.g., Rosenthal & Jacobsen, 1968; Rist, 1970; Jussim, 1986; Schultz & Oskamp, 2000; Jussim & Harber, 2005; Jussim, Robustelli, & Crane, 2009; Jussim, 2012).

1.3 Introduction to Marking Processes

As Brennan (2008) notes, although social psychology theories that explain how expectancies impact upon social judgement did not originate within the context of marking and assessment, their relevance to this field has been long acknowledged. Psychologists who have examined expectancy effects within marking processes in university settings have also noted that, "…it would be strange if markers were immune from these effects" (Dennis et al., 1996, p.516). The recognition that social psychology has a role to play in marking and assessment has been underlined by a growing number of researchers (Shay, 2005; Shay, 2008; Tuck, 2012) who believe that all forms of assessment are, 'social techniques which have social consequences' (Connell, Johnson, & White, 1992, p. 23).

While it is true that assessment underpins academic standards (Price 2005), it is also true that confidence in assessment within HEIs is low (Newstead 1996; Race 2003; Knight, 2002; Yorke, 2008; Bloxham, Boyd & Orr, 2011). This has been reflected in the results of the National Student Survey (NSS) where more than 40% of students have expressed some dissatisfaction with assessment and feedback (Bols, 2013). Whilst it is a criticism of the NSS that these results cannot be sufficiently unpacked to disaggregate where the dissatisfaction lies, it still paints an unflattering picture of assessment in HEIs. Part of the lack of confidence in assessment derives from the knowledge that although it is of central importance for students, the marking process often remains highly subjective (Elwood, 1999; Orr, 2007; Shay, 2008; Tuck, 2012). The calls for anonymous marking (NUS, 1999, 2008, 2012), and the concomitant flurry of research examining whether the removal of student names from assignments reduces the ability for lecturer expectancies to bias marks awarded is evidence of this concern.

Many of the concepts discussed later relate to the challenges of marking and the potential these have to impact on the mark awarded to students. For example, the ability of lecturers to reliably mark subjective assessments such as essays has been debated for many years (Hartog & Rhodes, 1935). Therefore the contentious issue of marker stringency, or variability will be addressed. Additionally, the impact of halo effects (Thorndike, 1920) on marking student work is also examined. Halo effects can manifest themselves positively or negatively and are a form of cognitive bias where knowledge about an irrelevant trait of a target (he is handsome), means a perceiver makes other judgements about their personality (he is kind). In the case of marking a students work this might take the form of, she is Asian, she will struggle to write well. Here the perceiver's cognitive bias about the target's ability to write well has been influenced by knowledge of her ethnicity.

The most researched form of marking bias however relates to gender bias. In general the contention has been that males are marked more favourably than females, especially in certain subjects. However, research findings in this area are equivocal, and much of the research has been school as opposed to HEI based (e.g., Bradley, 1984; Goddard-Spear, 1984; Newstead & Dennis, 1990; Breda & Ly, 2014; Enzi, 2015; Krawcyzk, 2017).Ethnic bias is also relevant here, but has been under-researched. This lack of attention is surprising given that research has maintained that ethnicity is linked to academic attainment (Dee, 2004; Lindahl, 2007; Ouazad, 2008; Burgess & Greaves, 2009) and that prejudices against ethnic minority students have been found in related areas (e.g., teachers judged Black students more negatively on personality traits [Coates, 1972]; teachers paid Black students less attention [Rubovits & Maehr, 1973]; Black students created more negative expectancies on the part of the teachers [Baron, Tom, & Cooper, 1985]). However, once again the focus has been on bias within schools rather than within HEIs, and with the exception of a paper by Burgess and Greaves (2009), the studies have been non-UK based. Nonetheless, most of these studies do indicate bias which favours White students (Ouazad, 2009; Spietsma, 2009; Burgess & Greaves, 2009; Kiss, 2013). University research into ethnic bias has not examined marking bias, but instead appears to have focused primarily upon whether admission processes are biased (Shiner & Madood, 2002; Gittoes & Thompson, 2005; Higher Education Funding Council for England [HEFCE] 2005; Madood, 2006; Shiner & Madood, 2010; HEFCE, 2010; Noden, Shiner & Madood, 2014); tutor expectations of different ethnicities (Tenenbaum & Ruck, 2007); and whether tutor gender and ethnicity impact on student perceptions of tutor competence (e.g., Dee, 2004; Dee, 2005; Bavishi, Madera, & Hebl, 2010; Reid, 2010). Thus there is currently very little research examining marking bias according to ethnicity at HEI level.

It is important however to caveat definitive claims about bias in marking with a recognition that much of the research in this area has lacked experimental rigour (including not using real teachers to mark work, not using assignments submitted for authentic assessments), has not controlled for confounding variables such as marker variability, and has not used marking criteria. It has also suffered from weak theoretical underpinning and application, often suggesting that teacher's expectancies might bias the marks awarded to students but not detailing how and why those expectancy effects might prevail.

1.4 Introduction to Feedback

Feedback is the cornerstone of learning, and its importance is espoused at all levels of education (Hattie & Timperley, 2007). It is also an emotive topic within HEIs and synonymous with frustration for students and academics alike (Tuck, 2012; Shields, 2015; O'Donovan, Rust & Price 2016; Pitt & Norton, 2016; Winstone, Nash, Rowntree & Parker, 2017). The frustrations are multiple, but for students seem to relate to quality, quantity and timeliness (Orrell, 2006; Lizzio & Wilson, 2008; Poulos & Mahoney, 2009; Ferguson, 2011; Small & Attree, 2015; Pitt & Norton, 2016). From a lecturers standpoint, many work hard under increasing amounts of pressure to provide feedback they hope will be helpful (Mutch, 2003; Tuck, 2012). However, there is a perception that students fail to engage with the feedback provided and that they are more interested in the grade or mark awarded (Weaver, 2006). Arguments to explain why this might be the case include that students do not understand the feedback, are unsure how to use it, and that the content is largely negative (Lea & Street, 2000; Mutch, 2003; Lizzio & Wilson, 2008; Poulos & Mahoney, 2009; Ferguson, 2011; Pitt & Norton, 2016). Therefore engagement with feedback is something that students often wish to avoid. Given that many have argued that the meaning of assessment is only achieved through the feedback provided (e.g., Orrell, 2006), this lack of engagement is worrying.

However, while student attitudes and beliefs about feedback have been explored, there has been no substantial research into student perceptions of bias regarding feedback. As part of a wider study, Pitt and Winstone (2018) have recently examined anonymous marking and perceptions of fairness, although perceptions of fairness were not tied to feedback per se, but to the issue of anonymity. Aside from the fact that perceptions of bias might impact upon student's motivation to engage with feedback, it is a concern in its own right. Feedback is a social practice (Shay, 2008; Tuck, 2012), and as such is a window through which cognitive biases and expectancies are made visible. There is evidence that students show bias when engaging with feedback, since their responses to it are influenced by their perception of the lecturer providing it (Orsmond, Merry, and Reiling, 2005). Therefore lecturers' provision of feedback might also be influenced according to who they are providing the feedback for.

Given that research examining ethnic bias in schools has demonstrated that the oral feedback minority students receive has been of poorer quality than that provided to dominant ethnicities (Coates, 1972; Rubovits & Maehr, 1973; Tenenbaum & Ruck, 2007), an examination of written feedback appears overdue. However, although research in social psychology has examined feedback bias across different ethnicities (e.g., Harber, 1998, Harber, Stafford, & Kennedy, 2010) the only educationally based research to explore feedback differences among ethnic groups was conducted by Richardson, Alden Rivers, and Whitelock (2014). However, their focus was to identify whether different feedback practices could explain the attainment gap rather than exploring whether feedback itself was biased. Furthermore, the only research that explicitly claimed to explore feedback bias according to gender at HEI level is by Read, Francis, and Robson (2005) although it was not the sole focus of their research. They identified no differences in feedback provided to each gender although they were cognisant of a number of limitations to their study which may have impacted on these findings.

Although there has been much research on feedback very little has explored expectancy effects in relation to gender and ethnicity. Given that feedback has a big impact on student self-esteem and motivation to study (Thorpe, 2000; Shields, 2015) this would seem an area ripe for examination.

1.5 Purpose and Research Questions

The purpose of this thesis was to explore whether expectancy effects, as primed by knowledge of the student name on their work, would bias the feedback awarded by lecturers.

These problems were addressed through the following research questions:

iv) Do expectancy effects as primed through knowledge of student gender impact upon feedback in a way that suggests biased practice?

v) Do expectancy effects as primed through knowledge of student ethnicity impact upon feedback in a way that suggests biased practice?

vi) Do expectancy effects as primed through the combined knowledge of both student gender and ethnicity impact upon feedback in a way that suggests biased practice?

This PhD aims to explore expectancy effects within assessment and feedback. Its originality can be expressed in the following ways:

1) It explores whether expectancy effects related to gender, ethnicity, and gender and ethnicity bias the provision of feedback in undergraduate student work.
2) It explores the interactive impact of expectancy effects on feedback (i.e., gender and ethnicity).
3) It introduces a level of experimental rigour not previously found in research in this domain. Specifically, this included the use of a control essay within the experimental protocol in an effort to determine marker severity. Furthermore, it used clear marking criteria for participants to make reference to whilst marking and providing feedback on the assignments.
4) Previous research in the educational domain has been largely a-theoretical. This thesis is underpinned by substantial social and cognitive psychological models, concepts and research as well as pedagogical research.

2 REVIEW OF LITERATURE

2.1 Chapter Outlines

Chapter one will discuss expectancies and implicit biases and introduce their relevance to marking and feedback. Specifically, Olson et al.'s., (1996) Model of Expectancy Processes will be introduced and discussed, alongside complimentary and alterative models. Styles of information processing will then be examined as well as the numerous moderators of such processing. Consideration is then given to the contentious role of automaticity within the expectancy formation process before examining the cognitive, affective and behavioural outcomes of expectancies.

Chapter two examines marking processes. It begins with an introduction to cognitive and perceptual biases and their relevance to the area of assessment. The importance of assessment is then introduced before providing an overview of some of the current challenges damaging confidence in university-level assessment. These include, marker variability, marking criteria and the subjective nature of assessment. Halo effects are introduced as one explanation for how bias might operate when marking student work. Research examining gender bias and ethnic bias specifically within marking processes is explored and critiqued, before examining research that has specifically searched for biases that might emerge when student work was anonymised versus de-anonymised.

Chapter three explores the role of feedback. It identifies how research has grown in recent years partly due to growing levels of accountability within the sector. It then provides an overview and critique of the attempts that have been made to define and then classify what constitutes good feedback. Research pertaining to positive and negative practices are discussed and criticised along with a discussion of student perceptions of useful feedback. The chapter concludes with an examination of the potential for feedback to be biased according to gender and ethnicity; an area that has been largely ignored in the educational literature to date.

2.2 CHAPTER ONE: EXPECTANCIES

The most succinct definition of expectancies is provided by Olson et al., who consider them to be 'beliefs about a future state of affairs', (1996, p.211). Expectancies have been classified in two ways; interpersonal and intrapersonal. Social psychologists have predominantly been interested in interpersonal expectancies since these refer to the expectancies individuals have about other people. Conversely, intrapersonal expectancies refer only to expectancies about the self (Jussim, 1990). Throughout this thesis the term expectancies will refer to interpersonal expectancies.

In a university setting, expectancies between a lecturer (perceiver) and a student (target) are inevitable. Specifically, in the context of marking student work, expectancy effects may be visible in the mark awarded, the feedback, the approach the student adopts towards subsequent work, and the behavioural interactions between perceiver and target. Indeed, much of the early work on expectancies and their impact was conducted in educational settings. Rosenthal and Jacobson's (1968) ground breaking research on the Pygmalion in the classroom demonstrated how perceivers' expectations influenced their behavioural response to a target. Rosenthal and Jacobson (1968) manipulated teachers' expectations by informing them that some students had been identified as 'late bloomers,' and therefore had the greatest potential to improve academically. Over the year 'late bloomers' demonstrated a significantly greater increase in IQ scores compared to other students. The researchers claimed that the teachers' expectancies influenced their behaviour towards the students, and were instrumental in the results.

Whilst methodological and interpretive criticisms have been levelled at this research (Thorndike, 1968; Snow, 1995; Jussim & Harber, 2005), it was influential in spearheading further investigations into expectancy effects. Subsequently their influence has been documented in a range of domains including job interviews (Biesanz, Neuberg, Smith, Asher, & Judice, 2001; Ridge & Reber, 2002), sport (Horn & Lox, 1993; Buscombe, Greenlees, Holder, Thelwell, & Rimmer, 2005), psychotherapy (Miller and Turnbull, 1986), race related weapon bias (Payne, 2006), and other organisational settings (Kierein & Gold, 2002).

Jones and McGillis (1976) identified that expectancies come in two types. Category-based expectancies, derived from knowledge about the groups or categories that individuals belong to (e.g., Asian student, female student), and target-based expectancies, derived from prior knowledge or previous experience with a specific individual (e.g., this student seldom attends lectures and has a poor academic profile). One of the key models providing an holistic overview of expectancy formation, processes, and effects/consequences was designed by Olson et al. (1996). Presented as a general framework to help study the concept of expectancies, the model is therefore applicable to any domain.

2.3 Olson, Roese and Zanna's (1996) Model of Expectancy Processes

Olson et al.'s (1996) model was chosen as the central model for this theses since it examines expectancies in their entirety. It explains the sources and properties of expectancies, as well as their cognitive, affective, and behavioural consequences or effects (see Figure 1). Alternative models have been guilty of simply focusing on expectancy effects or being solely concerned with the information processing aspects of expectancies as opposed to examining the process holistically. Furthermore, the model is underpinned by broad range of expectancy research undertaken within social psychology. Nevertheless, a common drawback of integrated models such as Olson et al.'s (1996) is that they have to make choices about what to omit and can sometimes sacrifice depth for breadth. As such the model is not without its limitations. Largely these relate to its failure to include a discussion of the impact of dynamic and static cues (Cook, 1971) which continue to have relevance in contemporary research in the area; an omission of the concepts of cognitive rigidity and belief certainty when discussing the property of Certainty; negating to consider research on mood state which suggests that that specific moods heighten accessibility to specific thoughts when discussing the property of Accessibility; a truncated discussion of automaticity within the property of Explicitness, and a lack of clarity surrounding the terminology used in the model and the accompanying text. Each of the limitations will be raised in turn as the discussion of the model unfolds.

According to Olson et al. the concept of expectancies "…forms the basis for virtually all behaviour" (1996, p.211). Therefore for those interested in the process of interpersonal interaction their attraction is irrefutable. Initially, Olson et al.'s model (1996) will be discussed in relation to expectancy formation, before being returned to later to examine the properties and consequences of expectancies.

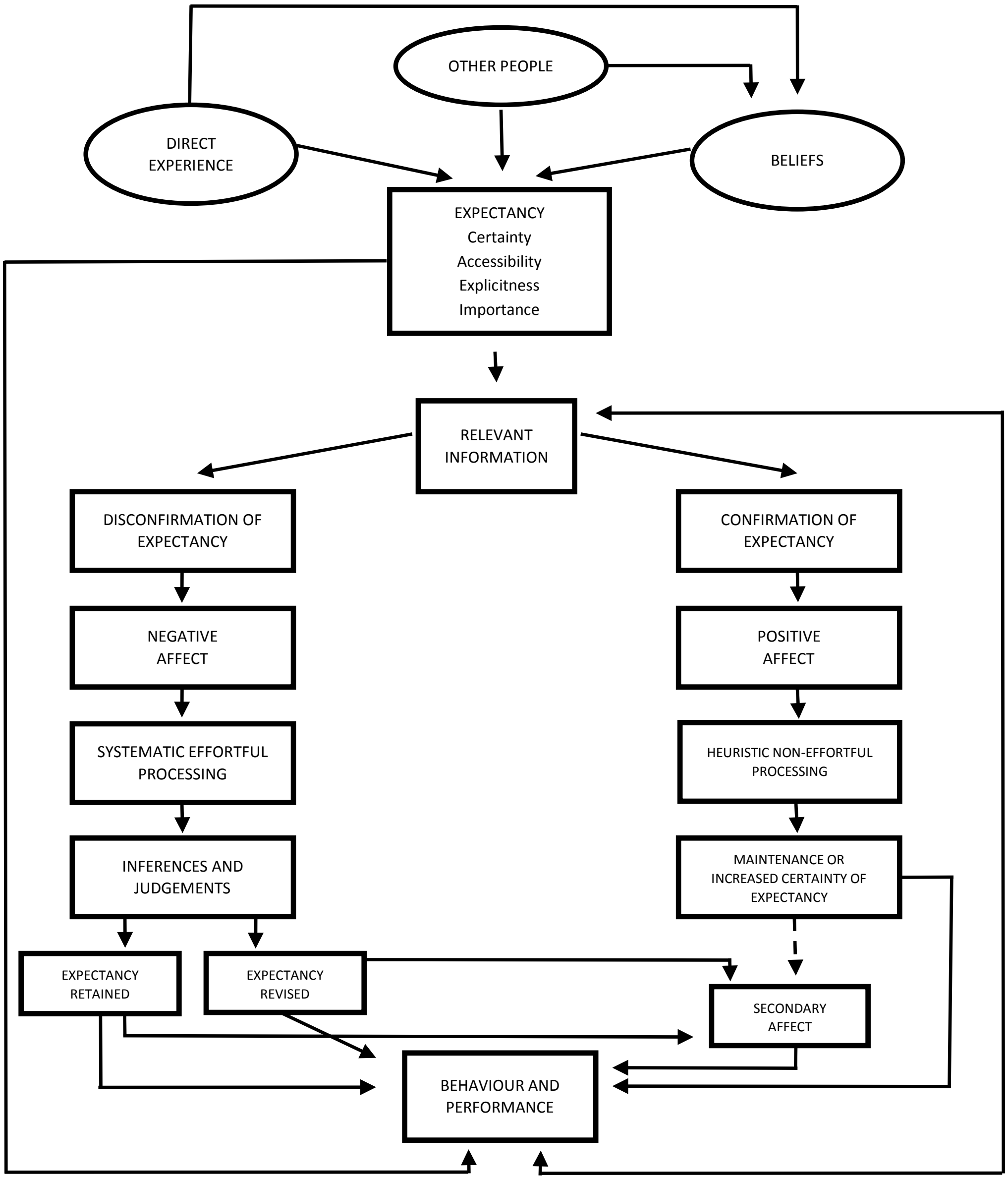

OTHER PEOPLE
DIRECT EXPERIENCE
BELIEFS
EXPECTANCY
Certainty
Accessibility
Explicitness
Importance
RELEVANT INFORMATION
DISCONFIRMATION OF EXPECTANCY
CONFIRMATION OF EXPECTANCY
NEGATIVE AFFECT
POSITIVE AFFECT
SYSTEMATIC EFFORTFUL PROCESSING
HEURISTIC NON-EFFORTFUL PROCESSING
INFERENCES AND JUDGEMENTS
MAINTENANCE OR INCREASED CERTAINTY OF EXPECTANCY
EXPECTANCY RETAINED
EXPECTANCY REVISED
SECONDARY AFFECT
BEHAVIOUR AND PERFORMANCE

Figure 1: Model of Expectancy Processes

2.3.1 Expectancy Formation

Despite widespread interest in expectancies, the majority of research has concerned itself with expectancy effects and not their antecedents. This lack of interest seems naive, since if the antecedents are not understood (i.e., expectancy formation), attempts to understand and influence the subsequent effects would seem limited. Furthermore, there is evidence that once a perceiver has formed an impression, it will be used to guide their future decisions and social interactions whether it is accurate or not (Greenlees, 2007). Recognising this inconsistency in the literature, Olson et al.'s (1996) model identifies several antecedents and begins by explaining that all expectancies are derived from beliefs (i.e., an individual's pre-existing knowledge or schema about the world). Schemas are mental constructs that house a body of knowledge, and often consist of beliefs, impressions, and information about social groups or individuals (Smith, 1998).

The beliefs identified by Olson et al. (1996) have three sources; direct personal experience, indirect personal experience and other beliefs.

2.3.2 **Direct Experience**

Direct personal experience is information about the target experienced directly by the perceiver. Its role as a source of expectancy formation is supported by other research (e.g., Warr & Knapper, 1968; Jussim, 1991). For example, a lecturer who has direct personal experience that a particular student never attends seminars may form the expectancy that the student will continue to be absent and therefore underperform in their assignment. Research has demonstrated that expectancies derived from direct personal experience are stronger, held with more confidence, more accessible, and more predictive of future behaviour than others (Fazio & Zanna, 1981). Olson et al. (1996) attribute this to the trustworthiness of such beliefs.

2.3.3 **Indirect Experience**

Defined by Olson et al. as “communication from other people” (1996, p. 214), indirect experience refers to expectancies held about a target despite the perceiver having had no direct interaction with the target. These types of expectancies are often referred to as reputational effects. For example, a lecturer with no direct experience of a student could form expectancies of them in a number of ways such as by speaking with other lecturers about the student’s reputation. Arguably, this process would work just as well to bias expectations about groups. For example, when a lecturer hears colleagues complaining that their third year undergraduates lack academic prowess, or that Asian students struggle to write coherently. Jussim’s (1991) assertion that expectancies are often based on acceptance of rumour, gossip, and hearsay buttress Olson et al.’s (1996) claim as to the importance of indirect experience. Furthermore, research interest in this area is varied and plentiful, specifically in relation to the domains of sport (Jones, Paull & Erskine, 2002; Plessner, 2005; Manley, Greenlees, Thelwell & Smith, 2010) and education (Batten, Batey, Shafe, Gubby, & Birch, 2011) thus indicating that the potential for indirect experience and reputation to influence beliefs and subsequent expectancies is considered important.

However, the impact of expectancy formation through indirect experience is likely to depend upon the ‘authority’ of the person providing the information (Cook, 1971). Specifically, “…credible ‘third party agents’” (White, Jones & Sherman 1998, p.15) have been identified as viable sources of expectancy formation for perceivers without direct experience of the target.

2.3.4 **Other Beliefs**

These refer to, 'beliefs that can be inferred logically from other beliefs' (Olson et al. 1996, p.214). Therefore, they do not necessarily depend upon direct or indirect experience, and instead can consist of a set of rules a perceiver may invent for themselves in order to guide their expectancy formation. Cook (1971) termed this process 'construction'. In the context of marking student assignments this could take the following form. This is a Chinese student's essay. I do not believe Chinese students' write well. This essay will be poorly written. The assumptive nature of this construction process illustrates the close relationship that can exist between stereotypes and subsequent expectancies.

While Olson et al. (1996) afforded space to antecedents to expectancies within their model, there is a disjuncture between the terminology and the text. Whilst the model labels the circled antecedents direct experience, other people, and beliefs, the text clarifies that 'other people' actually refers to indirect personal experience, and 'beliefs' refers to other beliefs that people might hold. This lack of clarity is unhelpful and therefore a re-labelling of these is overdue.

Additionally, the model omits to include the impact of static and dynamic cues on expectancy formation. Introduced by Cook (1971), and cited frequently in more contemporary literature (e.g., Manley, et al., 2010), static cues refer to constructs that remain relatively stable over time (e.g., gender, ethnicity). As such parallels can be drawn between these and category-based expectancies or stereotypes. Dynamic cues are changeable, and refer to things like body language or clothing. Cook (1971) suggested that both cues were used frequently in expectancy formation, but that dynamic cues were more likely to allow for the formation of accurate judgements. Later research has supported this argument demonstrating that dynamic cues have greater predictive validity than static cues [Jussim, 1993; Manley et al, 2008; Horn, Lox, & Labrador, 2010). For example, Jussim, Coleman, and Lerch (1987) demonstrated that dynamic cues (i.e., clothing and speech style) were more predictive than race for determining applicants' job suitability. This does not mean static cues are powerless however, since research has also demonstrated that race (Bodenhausen, 1998, 2005), gender (Manley et al., 2010) and physique (Lubker, Watson, Visek & Geer, 2005) are antecedents in expectancy formation, thus demonstrating that both cues have influence.

2.4 **Warr and Knapper's (1968) Schematic Model of Person Perception**

An alternative framework that illustrates the determinants and consequences of expectancies is the schematic model of person perception (see Figure 2).

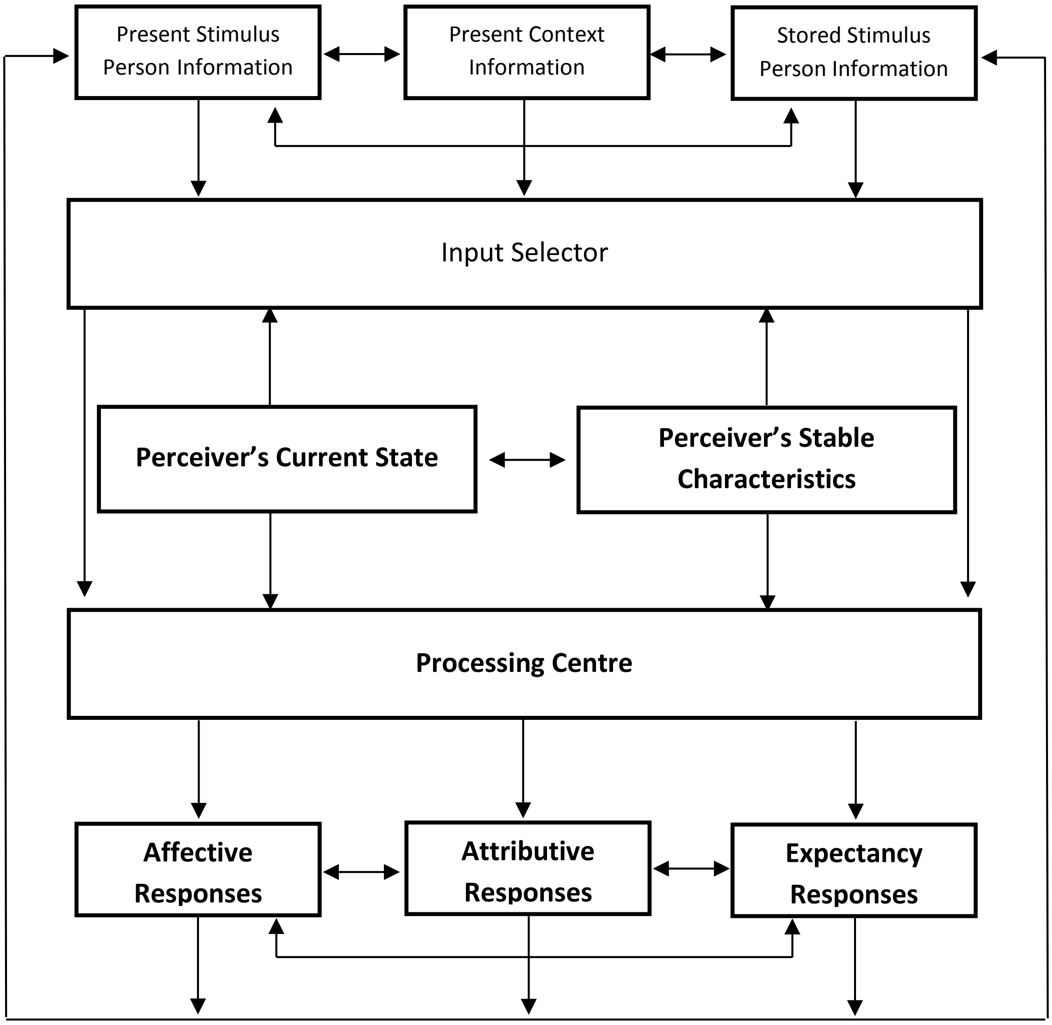

Present Stimulus Person Information
Present Context Information
Stored Stimulus Person Information
Input Selector
Perceiver's Current State
Perceiver's Stable Characteristics
Processing Centre
Affective Responses
Attributive Responses
Expectancy Responses

Figure 2: Warr and Knapper's (1968) Schematic Model of Person Perception

The model explains how information is processed during person perception. Similar to Olson et al. (1996), Warr and Knapper (1968) proposed that the role of direct experience is critical in the formation process; however, they break this down into three categories. First, present stimulus person information, refers to information received at the time of the interaction (e.g. clothing, gender, ethnicity). Second, present context information, which considers how information can be viewed differently according to the context it is presented in. For example, poor grammar in an assignment might be perceived less favourably than in a more informal communication with a student. Third, stored stimulus person information, which refers to prior knowledge held about the target. For instance, remembering that the last assignment by this student won an award.

Only a small amount of the information available from these three sources is processed (via the input selector).The content of this abridged information is determined by the perceiver's stable characteristics (e.g., their beliefs, attitudes, or personality traits) and the perceiver's current state (e.g., mood, arousal, motivation, etc.). This information is then processed, decisions made and inferences drawn. Three interdependent responses may then transpire; affective, attributive and expectancy.

Affective responses describe when the perception of a target evokes an emotional response such as pleasure or anger from the perceiver. Thus a lecturer, who perceives a student to present her work poorly, may feel emotions such as frustration and disappointment while marking her assignment. Attributive responses refer to when perceivers make judgements regarding the characteristics, states, and goals of a target. These judgements can happen in the absence of real evidence and are reminiscent of the construction process identified by Cook (1970). For example, a lecturer may make a judgement that a student has made little effort with the assignment because they are lazy. Lastly, expectancy responses are said to occur when people use the information gleaned from the person perception process to formulate stereotypic expectancies about how targets are likely to behave in the future. For example, a lecturer might mark a student's assignment and consider that marking their future submissions is going to be mentally demanding and uninspiring. Alternatively they might limit the amount of feedback they provide because they do not expect the student to read it. Thus attributive and expectancy responses are similar to schemas where an individual infers aspects of a target's character from the information that is presented (attributive response), and also develops a set of expectancies about how the target person is likely to behave (expectancy response).

Much of the research on person perception and expectancies has focused on the information detected between the perceiver and target on the basis of them being present in the interpersonal interaction. However, as Eysenck (2009) has noted, the principles of person perception and expectancy formation are not reliant on the perceiver being in the presence of the target for them to be influential. Therefore cues gleaned from information about the target outside of any direct interaction can also be influential in the expectancy formation process, e.g., the target's name. Names might be considered to be static cues or a form of present stimulus person information that act as a catalyst for the activation of category-based, stereotypic expectancies, and also elicit affective and attributive responses.

The information contained within a name allows perceivers the opportunity to create a multitude of expectancies including, but not limited to, a target's gender and ethnicity. In an intriguing demonstration of this, Krueger (2002) manipulated names on job applicants' curriculum vitaes in the USA by gender and ethnicity. The information about the candidate in the rest of the curriculum vitae remained the same. Males and applicants with white sounding names were called to interview fifty percent more than females and applicants with black sounding names.

Presumably the occupation in Krueger's (2002) research was considered in order that it did not obviously align itself to stereotypic expectancies related to gender or race (e.g., truck-driver, housekeeper) and thus become a confounding variable. Unfortunately however no information was provided about the job title the applicants applied for. Nonetheless, Krueger's findings have been replicated since in a variety of domains and are by no means isolated (Bodenhausen, 1988; Bertrand & Mullainathan, 2004; Carlsson & Roothe, 2007).

In the domain of marking it seems possible that the static cue (i.e., the name associated with the assignment) has the potential to impact many things. For example, the present-stimulus person information (e.g., gender, ethnicity) as identified by Warr and Knapper (1968), as well as indirect experience (e.g., everyone says that Asian students are diligent), and other beliefs (e.g., although this assignment is poorly written, English is likely to be their second language) as identified by Olson et al. (1996). The activation of such category-based expectancies subsequently has the potential to elicit certain affective, attributive and expectancy-based responses on the part of the perceiver.

2.5 Approaches to Information Processing

Person perception and expectancy formation processes require perceivers to draw upon information from a range of sources. These include, past knowledge, experience and beliefs, constructed knowledge invented by the perceiver, the perceptions of others, and present information about the target and context. This information is designed to help them make predictions about the future. This often requires a significant amount of cognitive effort, and therefore unsurprisingly the fundamental mechanism proposed to explain expectancies lies in the realm of information processing. Initial expectancies influence several aspects of the information processing including; a) the amount of attentional resources a perceiver will dedicate to the information, b) the information that is sought out and attended to, and c) the information that is remembered about a target (Bodenhausen, 1988; Fiske and Taylor, 1991; Macrae & Bodenhausen, 2000).

2.5.1 Schema-driven Processing

There are two broad approaches to information processing; schema-driven and data-driven (Greenlees, 2007). The most applicable type of schema to understand interpersonal interaction is the person schema. A person schema is a cognitive framework that encompasses a person's knowledge and beliefs about the characteristics of a specific type of person (e.g., Black, White, hard-working, good-natured etc.) and the relationships among these characteristics (i.e., Black and hard-working) (Fiske & Taylor, 1991). This schema will include judgements of the characteristics, mental states, and goals of that type of person (e.g., female students are conscientious; sport students are unfocused on their academic studies). Thus the schema for a “good student” could include knowledge about their gender or ethnicity, whether they have read widely, attended lecturers etc. Schemas can also include expectancies about how a person is likely to behave and respond to certain situations (e.g., female students will read the feedback given to them, or Asian students will attend tutorials).

Schema-driven approaches have dominated research, and subscribe to the view that the limitations of the human mind make it impossible to process all available information (Allport, 1954; Tversky & Kahneman, 1974; Kahneman & Tversky, 1996). Thus for reasons of cognitive efficiency people rely on categorical thinking or schemas (Chapman & Chapman, 1967, 1969; Zadney & Gerard, 1974; Darley & Gross, 1983; Snyder, 1984; Dijker 1987; Fiske & Taylor, 1991). When a tutor is marking an assignment, schema-driven theorists would argue that the tutor assigns the student to a category based on the cues in the early stages of an interaction (e.g., seeing their name on the cover sheet; reading their introduction) and then makes a judgement which forms expectancies for the remainder of the interaction (i.e., the marking process).

Early proponents of schematic thinking (also termed category activation), suggested that this process was unavoidable. Indeed Allport stated that, “the human mind must think with the aid of categories… We cannot possibly avoid this process. Orderly living depends on it” (1954, p.20). However thinking with categories makes individuals prone to cognitive biases. These are defined as, “systematic errors in judgement and decision-making …which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments” (Wilke & Mata, 2012, p.531). One type of cognitive bias are heuristics (Tversky and Kahneman, 1974; Kahneman and Tversky, 1996). These explain how perceivers use shortcuts in their reasoning to help guide judgements and draw on prior beliefs, experiences, and knowledge to aid their interpretation of events, as opposed to processing all available information. For example, a lecturer might use the heuristic that generally students who write a good introduction also write well for the remainder of their essay, and thus get good marks. This ‘rule of thumb’ way of operating has led to parallels being drawn between schematic thinking, heuristics, and stereotypes since the activation of person schemas or heuristics suggests that, “people often overestimate group members’ uniformity and overlook their diversity” (Smith & Mackie, 2007, p.145).

More recently the automaticity of schematic thinking first proposed by Allport (1954) has been subject to criticism (e.g., Bargh, 1994; Wegner & Bargh, 1998; Schwartz, 1998; Bargh, 1999; Macrae & Bodenhausen, 2000). Researchers proposed that perceivers might have some choice or control over the type of processing engaged in. For example, as well as being able to process information schematically, people could choose a more cognitively intensive information processing pathway, namely data-driven processing. Subsequently perceivers who continued to choose to process information schematically were labelled as cognitive misers (Fiske & Taylor, 1984) because their desire for cognitive economy meant that they were only willing to exert minimal effort in the person perception process. More recently Jussim has labelled these individuals as 'low wattage' (2012, p.4).

2.5.2 **Data-driven processing**

Data-driven processing represents a slower, more meticulous approach to information processing. Proponents of data-driven processing claim that perceiver's process information in a systematic and unbiased fashion as it becomes available (Fiske and Neuberg, 1987; Fiske & Taylor, 1991). As such people form expectancies of others by integrating every new piece of information about a target in order to form an individuated impression (Fiske & Neuberg, 1990; Pendry and Macrae, 1994). Received information is evaluated and assigned a weighting according to its relevance, and integrated into the evaluation of the target. Each new piece of information is then fitted into the overall evaluation such that a perceiver's impression is continually modified and amended (Fiske & Taylor, 1991). Data-driven theorists agree with schema-driven theorists that this process results in cognitive, affective, and behavioural responses, but maintain that these are responses to the perceiver's evaluation of the *information* as opposed to their evaluation of the *category* that the target belongs to. To use the previous example, when a tutor is marking an assignment, data-driven theorists would argue that a poorly written introduction would not trigger the expectancy that the rest of the assignment would be poor. Neither would the name or assumed gender and ethnicity of the student have much impact on the marking process, since this information would be assimilated with and reviewed alongside each new piece of information as it was encountered. However, claims that bias is avoided due to the systematic processing of each piece of information and a focus on information instead of category membership seem a little simplistic. One reason for this is that biases often exist and operate at an implicit level and therefore people hold biases which are unconscious. This concept will be explored in more detail later.

2.5.3 **Dual processing**

Fiske and Neuberg's (1990) continuum model of impression formation is one of many dual processing models which propose that perceivers can use both schema-driven and data-driven processing in tandem (Wood & Kallgren, 1998). These models consider perceivers as 'motivated tacticians' (Fiske & Taylor, 1991) who have a range of cognitive strategies at their disposal, and use the most appropriate according to their motives, needs, and goals. One of the model's propositions is that when perceivers have time and are motivated they make use of more individuating information and rely less on schematic thinking. Pendry and Macrae (1996) found that when participants were motivated, data-driven processing was used to form a more accurate perception of a target than when they lacked motivation.

Another proposition from the model is that when perceivers are cognitively busy and processing demands are high information is streamlined by using schema-driven processing (Fiske & Neuberg, 1990; Gilbert & Hixon, 1991). This is candidly explained by Allport (1954):

> We like to solve problems easily. We can do so best if we can fit them rapidly into a satisfactory category and use this category as a means of prejudging the solution.... So long as we can get away with coarse overgeneralizations we tend to do so. Why? Well, it takes less effort, and effort, except in the area of our most intense interests, is disagreeable (1954, p.20-21).

The effect of cognitive load on perceivers' ability to form individuated impressions of targets was clearly highlighted by Biesanz et al., (2001). Interviewers high in accuracy motivation were provided with a folder (by the researcher) which included a bogus personality profile for the applicant. This profile included scores indicating whether the applicant was well-suited to the position or not. Interviewers were then placed in either a highly distracting, mildly distracting, or no distraction group designed to replicate different cognitive loads. Interviewers free from distractions (low cognitive load) did not form expectancy-consistent impressions whereas those in the high distraction (high cognitive load) group did. This indicated that interviewers under high cognitive load had used schema-driven processing whereas those under low cognitive load had used data-driven processing to inform their expectancies about the candidates. This research provided support for other studies undertaken in this area (e.g., Neuberg, 1989; Fiske & Neuberg, 1990; Judice & Neuberg, 1998; Plessner, 1999), and led Biesanz et al. to comment that when cognitive load is high, 'even well-intentioned, accuracy-motivated perceivers can fall prey to their inaccurate expectancies' (2001, p.621).

The concepts of motivation and cognitive load have relevance to the marking process. It is assumed that tutors are motivated to form accurate impressions of student work which would necessitate using data-driven processing. However, marking is a cognitively demanding process which also occurs under time constraints and associated bureaucratic pressures. Research from other domains would therefore suggest that marking is likely to be dominated by a reliance on schema-based processing. Biesanz et al.'s (2001) research suggested that even high accuracy motivation was not enough to nullify the mediating effects of cognitive load on perceivers' ability to form an individuated impression of a target. This makes it hard to see how lecturers might avoid relying on expectancies when so many antecedents to their occurrence can be observed within the marking process.

Nevertheless, the impact cognitive load has on schematic thinking is a contentious issue. Some consider that whilst deficits in cognitive capacity generated by cognitive overload make it more likely that schematic thinking will occur, others believe that these same deficits will reduce a perceiver's ability to engage in stereotype activation and schematic thinking altogether. Gilbert and Hixon (1991) and Spear and Haslam's (1997) research demonstrated that stereotypes and schematic thinking were not activated under high cognitive load. However, both have fallen short in explaining what information processing style perceivers might have used. It seems unlikely that they would be engaged in data-driven processing since this resource-heavy style would further increase cognitive load when perceivers are already overloaded.

Presumably the desire for data-driven processing to take precedence over schema-driven processing resides in the association between data-driven processing and increased accuracy. However, engaging in data-driven processing does not necessarily increase accuracy of expectancies since people often hold biases which are unconscious and therefore cannot be overcome (Jussim, 1993, 2012). Additionally, a body of research spearheaded by Jussim (1990, 1993, 2012), has indicated that expectancies formed from schema-driven processing (and where relevant their subsequent stereotypes) are actually often accurate.

To further complicate this issue, some schemas or categories may have more potential to be activated and applied than others (Macrae & Bodenhausen, 2000). Categories of gender and race have been identified as, 'fundamental divides of the natural world' (Macrae & Bodenhausen, 2000, p.118), and therefore automatically activate categories in peoples' minds and incite schema-driven processing. Support for this is visible in Fiske, Lin, and Neuberg's (1999) Continuum Model, where they considered gender and ethnicity as 'privileged' categories which are available milliseconds into an interpersonal interaction. Lending support to these assertions Devine (1989) explained that people are exposed to stereotypic expectancies from early childhood and thus category activation is inextricably linked with stereotypic ideas. Research evidence supports the contention that ethnic groups hold specific traits (Packman, Brown, Englert, Sisarich & Bauer, 2005). Research conducted by the University of Chicago with 1,372 White Americans across three hundred communities demonstrated that Blacks were considered less intelligent, less industrious, more prone to violence, and less patriotic than Whites (Jonason, 2015).

Schemas related to gender appear even more resolute than those of race. Research on ‘typical’ traits considered women as as warm, sensitive, dependent, and relationship-oriented, whereas men were thought of as dominant, aggressive, independent, and task-oriented (Spence, Deaux & Helmreich, 1985). Williams and Best (1982) demonstrated the pervasiveness of this gender-related schematic thinking when they identified that results were similar in America, Asia, Africa, Europe and Australia. This research is dated however, and more contemporary work is required to identify whether these categorical expectancies remain. Nonetheless, it would appear that those categories that represent the fundamental divide outlined by Macrae & Bodenhausen (2000) appear to lessen perceivers’ ability to be the ‘motivated tacticians’ that Fiske and Taylor (1991) suggested. As such perceivers cannot help but use schema-driven processing and categorical thinking in expectancy formation.

2.5.4 **Moderators of schema and data-driven processing**

Conditions that may influence the activation of schema or data-driven processing has extended beyond those related to cognitive capacity (Locke et al. 1994; Blair & Banaji 1996; Lepore & Brown 1997; Macrae et al. 1997b; Wittenbrink et al. 1997). For example, the goal states of the perceiver (what they want or need out of the interaction with a target) have been found to be influential (Blair & Banaji 1996; Macrae, Bodenhausen, Milne, Thorn, & Castelli, 1997; Spencer, Fein, Wolfe, Fong, & Dunn, 1998). Specifically, the level of significance of the target to the perceiver has generated research interest (Neuberg & Fiske, 1987; Neuberg, 1989; Fiske and Depret, 1996). It has been noted that relationships considered important to the perceiver will generate the allocation of additional cognitive resources (Olson et al., 1996). As such, important targets are bestowed with more data-driven processing as perceivers attempt to form an individuated impression of them for their own self-interest. Thus targets considered important to perceivers, and with whom future interactions are expected, are less prone to inciting schema-driven processing and categorical thinking. The examination of relationship significance between target and perceiver has been extended to explore the role of power at play. Fiske and Depret (1996) contend that powerful perceivers will not invest the necessary cognitive effort to seek individuating information about less powerful targets. Furthermore, they will often attend to expectancy confirming information about such targets, thus reinforcing their own categorical thinking and paving the way for biased expectancy effects to transpire.

High-status perceivers such as teachers have been identified as a group most likely to elicit expectancy effects in the form of self-fulfilling prophecies (Smale, 1977). Consequently incorporating these issues into the arena of marking practices requires an acknowledgement of the power at play between lecturer and student. Lecturers determine student grades and the feedback they receive. They may also determine which students are deemed suitable to take specific modules, whether a student should progress to the next year, or whether to write them a reference. As such this is a relationship whereby the lecturer is powerful and the student is subordinate. Whether the student manages to transcend the schema-driven level of processing in the mind of the lecturer may therefore be largely out of their control. As Olson et al. (1996) suggest, it would also indicate that lecturers are unlikely to be sufficiently motivated to process student assignments in a more individuating manner unless they consider the relationship with their students as important.

However, it is possible that the target, whilst subordinate, may not remain entirely passive. Von Baeyer, Sherk, and Zanna (1981) claim that although subordinate targets may be susceptible to confirming perceivers' expectancies, when more tactically-minded targets recognise that a perceiver has something they want (e.g., a good grade on their assignment), they are motivated to behave in ways that shape and confirm the perceiver's expectancies to their advantage (e.g., to attend tutorials). This behaviour is strategic and designed to create a favourable expectancy in the minds of the perceiver. Bartram's (2016) research demonstrates how in the higher education environment the role of the target has become more dynamic. He explored how students' use emotional bargaining as a resource when requesting academic concessions. Whether this emotional bargaining resulted in students' gaining the concessions was not reported, but it does demonstrate evidence of the increasingly powerful role of the target.

Personality traits are also said to moderate information processing, expectancy formation and expectancy effects. One such trait is need for cognition, which is described as '… the tendency to enjoy and engage in effortful thought' (Sadowski & Cogburn, 1997, p.307). Individuals high in need for cognition tend to puzzle over problems, resolve inconsistencies, and search for the right answers (Cacioppo, Petty, Feinstein, & Jarvis, 1996). Perhaps unsurprisingly therefore, this type of person has been identified as more likely to invest their attentional energies into data-driven processing; thus reducing expectancy effects (Cacioppo et al., 1996). Alternatively, people low in need for cognition focus on information which is easier to process, adopting the 'cognitive miser' metaphor previously outlined (Fiske & Taylor, 1984). Given the career choices made by many lecturers it seems reasonable to assume that they might represent a group of people generally high in need for cognition. Therefore part of the reason expectancy effects might be attenuated in this population may be related to this trait overshadowing any effects brought about by the power relationship at play. There are also potential links here between accuracy motivation and need for cognition, with perceivers high in need for cognition also demonstrating high accuracy motivation, though this has yet to be explored.

Other moderators of schema-driven processing include general attitudes or level of prejudice toward the target (Lepore & Brown, 1997; Wittenbrink, Judd, & Park, 1997). Specifically, the stronger the attitude, prejudice or belief, the more likely schema-driven processing and categorical thinking are to occur. Devine (1989) suggested that the difference between prejudiced and non-prejudiced thinking might occur at the stage of category *application* as opposed to *activation*. She proposed that although all perceivers activate stereotypic categories and expectancies, less prejudiced individuals will use a process of controlled inhibition (the process of replacing stereotypic thoughts with their own non-prejudiced views) and therefore avoid the process of application. Subsequently however, Devine's work has been challenged. More contemporary research has found no evidence of categorical activation amongst the less prejudiced (Lepore & Brown, 1997; Wittenbrink, 1997), perhaps demonstrating that such people hold less prejudiced beliefs about these groups in memory and that category activation itself might be under some degree of control.

Mood state has also been considered to influence both information processing strategies and impression formation. Schwartz and Clore's (1996) review demonstrated that people evaluate things more favourably when in a good mood as opposed to a bad mood. In addition, irrespective of whether the perceiver is engaged in schema or data-driven processing their mood state can still influence what information is remembered (Forgas, 1995). However, generally speaking people who are feeling sad will engage in more data-driven, effortful, analytic processing (Clark & Isen, 1982; Isen, 1984, 1987; Schwarz, 1990). People in a happy mood rely more heavily on schematic processing, make use of existing attitudes and beliefs, and operate in a less meticulous way (Edwards & Weary, 1993; Bodenhausen et al., 2001). People in a sad mood are also less influenced by halo effects (Forgas, 2007), provide more accurate performance appraisals (Sinclair 1988; Sinclair and Mark, 1992; Forgas, 2011), make fewer judgemental errors, and use stereotypes less frequently than people in a happy mood (Forgas, 2007; Forgas & Laham, 2009). Applied to the realm of marking student assignments this suggests that perceivers are likely to process students' work more thoroughly, and be more accurate if they are in a bad mood.

Finally, one novel research area has explored the impact of the time of day on perceivers' decision-making. Bodenhausen (1990) collected information regarding his participants circadian rhythms (i.e., whether they were 'morning or evening types'), and then randomly assigned them to sessions at 09:00, 15:00 or 20:00. They were presented with the details of a man who had committed assault and was either called Robert Garner or Roberto Garcia. They were then asked to judge his guilt. People scheduled to undertake the task during their least favoured time of day were more prone to use schematic-processing and stereotypic expectancies to adjudge guilt to the Latino name. Thus, extrapolating from Bodenhausen's research, the time of day that lecturers mark student assignments may also influence their likelihood to process information schematically and rely on heuristics and stereotypic expectancies to inform their judgements.

2.6 The Automaticity Debate

The role of automaticity within category activation and expectancy formation is contentious. Some leading figures agree with Olson et al. (1996) that categories are activated and expectancies generated outside of conscious awareness (e.g., Devine, 1989; Greenwald & Banaji, 1995; Bargh, Chen and Burrows, 1996; Bargh, 1997; Chen & Bargh, 1997; Greenwald & Krieger, 2006) whereas others believe the perceiver plays an active, explicit role in this process (e.g., Snyder & Swann, 1978; Darley & Fazio, 1980; Neuberg, 1989, 1994; Snyder, 1992; Blair, 2002). Allport's (1954) influential early work forcibly identified category activation as automatic and unconscious, and his view went virtually unchallenged for over forty years (Macrae & Bodenhausen, 2000). However, some now question the role of automaticity entirely, while others suggest that perceivers can exert some control over the process (e.g., Horn, Lox, & Labrador, 2001; van Ryn & Fu, 2003). Another suggestion has been that the process might be conditionally automatic, occurring only under specific triggering conditions (Fiske, 1989; Bargh, 1994; Blair, 2002). An example of these triggering conditions was identified by Blair (2002). She found that automatic category activation and prejudice were influenced by; a) self and social motives, (e.g., preservation of self-image), b) specific strategies (e.g., stereotype suppression), c) the perceiver's focus of attention, (e.g., attentional load), and d) the configuration of stimulus cues (e.g., the context within which cues are received). Blair (2002) reported that group member's individual characteristics influenced the extent of category activation and stereotyping. She further claimed that the evidence presented in her research, 'stands in stark contrast to assertions that automatic processes are immutable and inescapable' (p.257) thus disputing Allport's original contention and those who acquiesced to his view.

Nonetheless, this new vanguard of research which claimed that category activation was conscious and preventable attracted criticism from those who adhered to the orthodox view. Perhaps the most vocal among these was Bargh who stated,

> …the field of social cognition has become overly optimistic about the 'cognitive monster' of automatic stereotype activation. . . . contrary to what research is actually showing, the conclusions drawn from the data have overestimated the degree to which automatically activated stereotypes can be controlled through good intentions and effortful thought (1999, p.362).

Additionally, supporters of automaticity are quick to point out that research which has explored conditional automaticity has been ambiguous, with some studies noting that category activation is contingent upon prejudice levels (e.g. Kawakami, Dion, & Dovidio, 1998; Blair, 2002) and others claiming the opposite. For example, Dunning and Sherman (1997) have demonstrated that implicit activation of the category of gender was unrelated to participants' levels of sexism, thus perhaps demonstrating that activation and application can indeed be separated. Furthermore, research exploring interventions such as category inhibition, thought suppression, and stereotype suppression as a means to regulate categorical thinking has been equivocal (e.g., Wegner, 1994; Wegener & Petty, 1997; Bodenhausen & Macrae, 1998). Some studies have even shown that attempts to suppress categories or stereotypes ironically increase the accessibility to them in memory (Macrae, Bodenhausen, Milne, & Jetten, 1994; Wyer, Sherman & Stroessner, 1998) and can therefore increase expectancy effects and bias. The notion of conditional automaticity and the potential influence of inhibitory mechanisms are relatively new challenges to the orthodox thinking in this area and thus this dispute is far from resolved.

Application of the orthodox view of automaticity to university marking practises is interesting. Perceivers are considered to hold schemas in long-term memory. These schemas simply need activating (through a priming stimulus such as gender or ethnicity). Once activated they can guide cognitive, affective, and behavioural responses in an entirely autonomous manner (Bargh, 1997) and consciousness need play no part. In line with this thinking, lecturers cannot help but think, feel, and behave in certain ways when they see a student's name and infer personal characteristics. Specific categories are activated in the lecturer's mind and the expectancies associated with those categories are ripe to be played out in a myriad of ways (i.e., if the name Aarav Singh activated the category Asian and hardworking, this might impact on the feedback the lecturer provides due to the expectancy that this student will engage with feedback and use it to improve future assignments). There has been scant acknowledgement of the role automaticity might play in the marking process. However, the growing belief that assessment is a social practice (Shay, 2008; Tuck, 2012) has led some to recognise that assessment processes can be implicit and unconscious. As such the exploration of markers judgements are incredibly difficult to explore since they are deeply internalised and therefore they are unable to fully articulate the processes through which they make such judgements.

Advocates of category inhibition (e.g., Wegner, 1994; Wegener & Petty, 1997; Bodenhausen & Macrae, 1998) would concede that schemas held in the lecturer's long-term memory would still be automatically activated upon processing a student's name. However, they would argue that the lecturer would then be able to use an inhibitory mechanism such as stereotype suppression to prevent the application of this schema. This would therefore avoid the expectancy-based and potentially stereotypic responses that might follow. In this instance, the lecturer's expectancies of the student would not impact upon the grade awarded or the feedback provided and the lecturer would be more likely to engage in data-driven processing as opposed to schema-driven processing whilst marking the assignment.

Finally, if category activation is to some extent controllable (Horn, Lox, & Labrador, 2001; van Ryn & Fu, 2003), or conditionally automatic (Fiske, 1989; Bargh, 1994; Wegner & Bargh, 1998; Blair, 2002), then a multitude of triggering conditions will determine whether category activation occurs, how the information is processed, and whether expectancy-based practices are evident in the marking process. Research in this area is in its infancy and though the work of Blair (2002) is laudable, much more research needs to be conducted before a holistic view of the determinants, moderators and mediators of conditional automaticity exists.

Nonetheless, the concept of automaticity has captured the imagination of numerous social psychologists because of the recognised effects that category activation and application have on interpersonal interaction. These categories generate expectancies or 'provisional hypotheses' (Darley & Gross, 1983) about individuals and groups. Importantly, these expectancies subsequently shape the processing of future information and are considered by many to '… exert a ubiquitous impact on social interaction" (Olson et al., 1996, p.216).

2.7 PROPERTIES OF EXPECTANCIES

According to Olson et al.'s (1996) model of expectancy processes, the activation of expectancies and subsequent information processing is only part of the picture. Importantly, it is the properties of these activated expectancies which determine their future consequences within interpersonal interaction (i.e., the expectancy response). These properties are certainty, accessibility, explicitness, and importance.

2.7.1 **Certainty**

Certainty has been defined as, "… the subjective level of probability associated with an anticipated outcome/event" (Olson et al., 1996, p.214). There are four determinants of certainty and there are parallels between these and the sources of expectancies that form the earlier stages of the model. Firstly, expectancies based on direct experience will have higher certainty than those based on indirect experience (Fazio & Zanna, 1981). Secondly, the determinant of consensus information means that observing others holding the same expectancy will increase certainty. For example, if the consensus amongst staff is that first year students will do the bare minimum to pass the year then this will increase certainty. Thirdly, the more accessible the expectancy the more certainly it will be held and the more likely it will be activated. Finally, whether expectancies have been previously confirmed will influence certainty. For example, in a university environment, if men have previously outperformed women on an assignment a tutor may be more certain that this will happen again.

Two additional determinants of certainty overlooked by Olson et al. (1996) but considered to determine expectancy formation and effects are cognitive rigidity and belief certainty. Cognitive rigidity has been considered a personality trait whereas belief certainty has been considered a state, or situational factor (Swann & Ely, 1984; Jussim, 1993). Schultz and Searleman (2002) claim that cognitive rigidity involves forming and persevering in using a mental pattern which includes expectancies and schemas. People high in cognitive rigidity are unlikely to alter their beliefs or expectancies even in the face of disconfirming evidence. They are likely to hold a deterministic view of human nature and therefore believe that the characteristics held by groups and individuals are unchangeable (Levy, Stroessner & Dweck, 1998). Conversely belief certainty is founded on the assumption that the context or environment in which beliefs are formed dictates the certainty with which a belief is held. It has been considered that people high in one or both of these constructs will rarely consider viewpoints that contradict their own, are overly confident in their expectancies, and more likely to elicit expectancy effects than people low in these constructs (Jussim, 1986, 1993). These assumptions are reinforced by Babad, Inbar, and Rosenthal's (1982) research with physical education teachers which demonstrated that those teachers with high cognitive rigidity were less friendly and more critical towards low-expectancy students. Teachers with lower cognitive rigidity behaved similarly to both high and low-expectancy pupils. Of course, the nature of the beliefs that people hold with such rigidity or certainty then becomes of interest since these expectancies have important consequences. More recent research collated results from a number of projects and demonstrated a relationship between high cognitive rigidity, high prejudicial beliefs, and low intelligence (Hodson & Busseri, 2012). While the link between intelligence, expectancies, and prejudice has yet to be fully explored understanding the cause of such biases will be imperative to addressing numerous social inequalities.

Resultantly a perceiver's level of certainty in their expectancies would seem to be determined by both personal and situational factors. Whatever the antecedent, people who possess such a high degree of certainty are most likely to maintain biased perceptions of targets and thus illustrate expectancy effects at work (Jussim, 1993).

2.7.2 **Accessibility**

Accessibility refers to the ease or speed with which an expectancy comes to mind (Olson et al., 1996). The more accessible the more likely the expectancy is to occur automatically and influence future interactions (Ford & Thompson, 2000). The categories of gender and race were previously described as "… fundamental divides of the natural world" (Macrae & Bodenhausen, 2000, p.118), and are therefore a good example of expectancy-laden categories whose accessibility is enhanced. Two key determinants of accessibility are frequency (e.g., Srull & Wyer, 1979) and recency (e.g., Higgins, Rholes & Jones, 1977). Specifically, expectancies which have been recently formed or frequently primed are highly accessible and more likely to be used to make sense of future interactions. For example, if a lecturer had recently helped a Chinese student with their assignment and commented on grammatical errors, they might expect similar errors when they later mark an assignment with a Chinese name on the submission sheet. Alternatively, if they had recently read an NUS report (2008) claiming that marking was biased according to gender and ethnicity they might be more aware of how bias might operate in their own marking practice.

The impact of priming on expectancy effects appears more significant when it is considered that a recently primed expectancy can remain accessible for twenty-four hours (Srull & Wyer, 1980). Perhaps more compelling, the impact of priming does not even depend upon the perceivers' awareness of the prime having taken place. This was illustrated in Bargh and Pietromonaco's (1982) research which involved flashing a priming word on a computer screen so quickly that participants could not identify the words. However, later when reading a description of a person's character, those participants primed with hostile words were more likely to interpret the person they had read about as hostile and aggressive. The researcher's concluded that even when participants were unable to consciously identify a word, encountering it was sufficient to make expectancies accessible and influence the interpretation of later information.

Another determinant of accessibility is mood state. The impact of mood related to the use of information processing strategies has been discussed, but its influence may be more wide-ranging than originally thought. Whilst not acknowledged by Olson et al. (1996), the mood that a perceiver is in seems likely to affect the types of thoughts that are accessible in a perceiver's mind when forming expectancies of a target (i.e., a good mood might make positive thoughts more accessible). The contention is that moods do this as a consequence of their impact on memory and judgement both through the process of recall and the use of feelings as a source of information (Schwartz, 1998). Specifically, perceivers are more likely to access and recall positive material from memory when they are in a happy mood (Schwartz & Clore, 1983). In terms of feelings, Schwartz and Clore (1983) claim that perceivers may simplify evaluative tasks by drawing upon a, 'How do I feel about it?' heuristic. Feelings then begin to influence judgements by serving as a source of information, or by influencing what comes to mind. Arguably, when the perceiver is in a good mood 'how they feel about it' is somewhat more positive that when they are in a bad mood and thus their judgement is also likely to be more positive.

Applied to marking this means that lecturers who are marking assignments when in a good mood may have greater accessibility to positive thoughts. They are more likely to access and recall positive aspects of the assignment from memory and use the feelings associated with being in a positive mood to make judgements on the work. Moods may therefore be considered as one antecedent to the halo effect (Thorndike, 1920) which when applied to marking can explain how perceivers may see qualities in an essay which complement their expectancy, but are not present or valid. Indeed research has already demonstrated that perceivers in happy moods were more likely than those in depressed moods to be influenced by peripheral cues rather than the quality of the argument presented and thus demonstrate halo effects (Sinclair & Mark, 1982).

Olson et al., (1996) identified expectancy disconfirmation as a key determinant of accessibility. Disconfirmation heightens accessibility since by its very nature it is surprising, and therefore noticeable. If a lecturer considers that an assignment was well-written *for an Asian student,* then the reason they have noticed this is because it contradicted their expectancy about Asian students' work. Some theorists claim that expectancy-consistent information is more likely to be attended to and encoded than expectancy-inconsistent information (i.e., individuals see what they want to see) (Chapman & Chapman, 1967; Miller & Turnbull, 1986; Harrison, Jr, 2001). However, Macrae and Bodenhausen (2000) contend that if both types of information are presented equally, then the information related to disconfirmation is most likely to dominate the perceiver's attention and encoding. Fundamentally Macrae and Bodenhausen (2000) attribute this to the increased attentional resources required to process information that does not fit with the expectancy. Expected (and therefore confirmatory) material is processed in a more schema-driven manner and therefore is relatively effortless, whereas unexpected (disconfirmatory) material requires a more data-driven approach and more cognitive effort.

2.7.3 **Explicitness**

Explicitness relates to whether the expectancy is generated consciously (explicitly) or unconsciously (implicitly). Given that argument surrounds whether category activation is conscious or unconscious it seems reasonable that the expectancies which follow have been subject to the same debate. Indeed, whilst Olson et al. (1996) do note that explicit expectancies are common in interpersonal settings, they also acknowledge that many expectancies exist without ever entering conscious awareness. In an attempt to definitively resolve the explicit versus implicit conundrum more sophisticated experimental designs, including attempts to obscure the relationship between the experimental stimuli, have been utilised (Devine, 1989; Gilbert & Hixon, 1991; Macrae, Bodenhausen, & Milne, 1995; Lepore & Brown, 1997; Wittenbrink et al., 1997; Kawakami et al., 1998). One experiment that demonstrated well how expectancies can operate at an implicit level was conducted by Chen and Bargh (1997). They presented subliminal cues (i.e., the faces of male African Americans or Caucasians) via a computer-based task to participants in the perceiver group and no subliminal cues to participants in the target group. A perceiver and a target participant were then placed in pairs and had to conduct a verbal game consisting of guessing a well-known catch phrase based on the clues given by their partner. Each took turns at being the clue-giver and the guesser. Next participants had to rate their impressions of each other on a trait-based questionnaire which included questions pertaining to the trait of hostility. Perceiver participants who had been exposed to the faces of African Americans recorded higher hostility ratings towards their game playing partners than did target participants who had not had such exposure. Chen and Bargh (1997) therefore concluded that subliminal cues were sufficient to activate implicit stereotypic expectancies.

The impact of implicit expectancies on behaviour is significant. When perceivers are conscious of their expectancies, the consequences can be more easily identified, monitored, prevented, challenged or encouraged (van Ryn & Fu, 2003). On the other hand, when they are implicit and occur "… behind the perceiver's back" (Schwartz, 1998, p.557), they become increasingly difficult to recognise and change (Wiers, van de Luitgaarden, van den Wildenberg, & Smulders, 2005), and thus their potential impact on interpersonal interactions is high.

Judgements regarding the level to which people hold implicit versus explicit expectancies vary and have been examined across numerous social categories. For example, Fiske (2002) has argued that only 10% of people in Western societies hold explicit racial biases. However, to counter this, she further observes that as many as 80% of people hold more subtle, implicit racial biases which can lead to, '… awkward social interactions, embarrassing slips of the tongue, unchecked assumptions, [and] stereotypic judgements" (2002, p.124). In a thought-provoking study, Payne (2001) briefly presented a photograph of either a Black male face or a White male face to perceivers (the ethnicity of perceivers was not disclosed). A photograph of an object was subsequently presented and perceivers had to decide within 0.5 seconds whether the object was a handgun or a handtool. Perceivers falsely claimed to see a handgun more often than a handtool when the preceding face was black, thus demonstrating the illusory correlation (whereby individuals only notice incidences that fit the stereotype) at work. In later research Payne, Lambert and Jacoby, claimed that biased expectancies operate implicitly and thus can, "… coexist with conscious intentions to be fair and unbiased" (2002, p.288).

This research would seem to support Olson et al.'s (1996) claim that many expectancies operate at an implicit level. Interestingly such racial bias is not unique in operating from Whites to Blacks. Some research has shown that responses made by Black American participants were indistinguishable from those of White Americans, with both groups demonstrating biased expectancies toward associating weapons with Black faces more than White faces (Correll, Park, Judd, & Wittenbrink, 2002). Bodenhausen (2005) explains that minority groups are often influenced by the same types of cultural stereotypes and biases as majority groups.

The above arguments have potential implications for the marking process. They would suggest that seeing a student's name at the top of their assignment could influence the marking process even if the information contained within the name is not consciously processed.

2.7.4 **Importance**

The final property of expectancies is importance, and its influence is determined by the perceiver's motivation towards the interpersonal interaction. Specifically it relates to how relevant the expectancy is to the needs of the perceiver. Similar to the role of motivation in determining the type of information processing activated, if forming an accurate expectancy is important to the perceiver they will engage in more data-driven, individuated processing, and expectancy effects are attenuated (Neuberg & Fiske, 1987). Similar results have been demonstrated when perceivers expect to have future interactions with a target (Neuberg, 1989). Indeed, expectancy effects between teachers and students have been shown to decrease once interaction time exceeded two-weeks (Raudenbush, 1984). This adds credence to Kelley and Thibaut's (1978) Interdependence Theory which outlines that people are motivated to spend time understanding individuals on whom their outcomes depend. Therefore there is an argument that category-based expectancies are likely to be less frequent and lower in strength when the perceiver has a vested interest in making accurate judgements about a target (Jussim, 1993) or has an extended period of contact time with them (Raudenbush, 1984).

2.8 EXPECTANCY EFFECTS

Now that the antecedents to and moderators of expectancies have been established and their properties determined, it is important to consider the consequences of formed expectancies on interpersonal interaction. These consequences (or expectancy effects) are considered by Olson et al., (1996) to manifest themselves in three ways, cognitively, affectively and behaviourally.

2.8.1 Cognitive consequences

Expectancies contribute to a number of cognitive consequences (e.g., Fiske & Taylor, 1991; Miller & Turnbull, 1986) and impact on cognitive functioning in a variety of ways. Indeed Olson et al.'s (1996) work identifies five potential cognitive elements which are influenced by expectancies; attention and encoding; memory; interpretation; attributions; and counterfactual thinking. Each of these will now be discussed in turn.

2.8.2 Impact on attention and encoding

While perceivers may process a wealth of information at any one time, it is the information that is attended to and encoded that is of importance since this will be saved as a schema, stored in memory, and later recalled. It is therefore important to acknowledge that, "People's expectancies…play a critical role in the selection of information from the environment to be encoded" (Higgins & Bargh, 1987, p.378).

In his examination of racial expectancies and stereotypes in sport, Harrison (2001) noted that expectancy-consistent information was more likely to be attended to and encoded than expectancy-disconfirming information. Applied to the marking process, this would mean an already formed expectancy that an assignment was written by a 'good' student would influence the lecturer to notice and process (i.e., attend to and encode) all the things the student does well. Interestingly, this cognitive bias would also lead the perceiver to discard information related to things the student did not do well (i.e., those things that disconfirm their expectancy). Therefore expectancies serve to bias information processing, are liable to cause the formation of inaccurate judgements on the part of the perceiver, and demonstrate halo effects at work. In this example, the lecturer may judge the assignment to be of a higher standard than it really is because their expectancy about the student's ability has lead them to attend to and encode only those aspects of their work which compliment this expectation. The grade and accompanying feedback is also likely to reflect this biased processing at work. This process is nicely wrapped up in Hamilton's comment about the influence of expectancies on the encoding process "…If I didn't believe it, I wouldn't have seen it" (1981, p.118).

One explanation forwarded to explain the salience of expectancy-consistent information is simply that it allows the perceiver to protect their original expectancy from disconfirmation (Olson et al., 1996). Presumably the need to do this relates to a self-serving bias whereby expectancy-consistent information justifies and reinforces beliefs and expectancies. Olson et al. (1996) also identify that expectancy-consistent information yields positive affective responses. Such information also provides little opportunity for inconsistencies to arise and therefore avoids the psychological state of cognitive dissonance which people are motivated to avoid (Festinger, 1957). A further explanation refers to the type of information processing perceivers are engaged in. When perceivers are processing information in a schema-driven fashion the lack of attention indicative of this approach results in inconsistent information being overlooked (Olson et al., 1996).

Chapman and Chapman's (1967) research demonstrated the extent to which perceivers attend to expectancy-confirming information. Participants were presented with drawings of faces which they were told had been completed by patients with specific mental illnesses. In reality there were no relationships between the drawings and the types of illness the patients had. Nonetheless, participants attended to and interpreted aspects of the drawings in line with the illnesses they believed the patients to have therefore reinforcing their original expectancies (e.g., enlarged eyes were thought to represent a patient's suspicious or paranoid nature). The influence of expectancies on encoded information was also demonstrated by Zadney and Gerard's (1974) simple experiment. Participants were shown footage of a student registering for a new class and were either told the student majored in maths or music. When participants thought he was a music major they recalled more music-relevant information from the footage and vice-versa. More recently a social experiment sponsored by Canon was conducted whereby six photographers were asked to take a single portrait photograph of a man which most represented how they saw him. Each photographer was given false personal information about the man prior to his arrival. The man arrived at the session dressed the same way and nothing about his demeanour altered. However, the photographers treated and photographed him differently according to the encoded information they had about him. The video ends with the tag line, “A photograph is shaped more by the person behind the camera, than by what is in front of it” (Canon, 2015 [Online]). Thus it would appear that when a target is assigned a category (e.g., female, Asian) this is a catalyst for the expectancy process to begin. Perceivers are then likely to attend to and encode additional and ambiguous information in relation to the activated category. In the context of marking it is feasible that perceivers who expect Asian students to write poorly may see an Asian name and subsequently recall numerous spelling mistakes in the work - thus reinforcing their expectancy that Asian students do not write well. In this way cognitive biases triggered by expectancies result in perceptual confirmation of perceivers’ original expectancies (Snyder, 1984) and the self-perpetuating nature of this process becomes apparent.

Nevertheless, as has been previously highlighted, although there is some support for the claim that people ‘see’ what they want to see (Harrison, 2001; Miller & Turnbull, 1986; Darley & Fazio, 1980; Chapman & Chapman, 1967) many researchers argue that expectancy-disconfirming information receives more processing since it requires an attempt to understand the inconsistency between what is presented and the original expectancy (Higgins & Bargh, 1987). More recently this claim has been tempered somewhat by the acknowledgement that this is only likely to occur if the perceiver has sufficient cognitive capacity to do so (Macrae & Bodenhausen, 2000; Sherman, Lee, Bessenoff, & Frost, 1998).

Irrespective of whether expectancy confirming or disconfirming information is attended to during interpersonal interactions, it is clear that expectancies have the ability to impact upon both the attention and encoding processes.

2.8.2.1 **Impact on memory**

The impact that expectancies have on memory has been discussed extensively, but only elements pertinent to this thesis are focused upon here. Research in the area of memory is contentious. Some research has suggested that information consistent with expectancies is better remembered than other information (Rothbart, Evans, & Fulero, 1979; Harrison, 2001), whereas alternate research claims the opposite (Hastie & Kumar, 1979). There is substantial crossover between the areas of attention, encoding, and memory, since it is assumed that information attended to and encoded is more likely to be remembered than information that is not subject to such scrutiny.

Expectancies have such a profound effect on memory that they can cause perceivers to falsely recall information from a situation when the information was not present. In his work examining the impact of perceptual biases, Jussim (1993) indicated that perceivers sometimes claim to remember others' behaviours in ways consistent with their expectancies even if that behaviour was not evident. In this way memorial biases and their subsequent inaccuracies contribute to the construction of a purely subjective social reality that exists in the mind of the perceiver and serves to maintain expectancies in the absence of supporting evidence (Darley & Fazio, 1980; Miller & Turnbull, 1986; Jussim, 1989, 1990, 1993).

Perceptual biases related to memory have been identified as determinants of inaccurate judgements when marking student work. Diederich (1974) and Rigsby (1987) found that the same essays were awarded higher grades when they were believed to have been produced by competent students. Presumably, knowledge about competence created expectancies in the minds of the teachers which contributed to them only attending to and remembering the competence-related information when marking. However these findings have been contradicted by research which found that manipulating student reputation had no impact on either grades or feedback (Batten et al., 2011) suggesting research in this area is equivocal.

More recent evidence of the role memory plays in assessment and feedback practices is explored in the work of Nash, Winstone, Gregory and Papps (2018). In an intriguing study spanning six experiments, they reported that students remembered past-oriented (evaluative) feedback better than future-oriented (directive) feedback. These results were unanticipated because much pedagogic research has evidenced a student preference for future-oriented (directive) feedback (e.g. Winstone et al., 2016) and cognitive psychology contends that memorial benefits are associated with information aligned with individual preferences. Perhaps the answer to why students did not remember their preferred type of feedback is linked to their broader expectancies about feedback. While it was not the focus of Nash et al.'s (2018) work to explore how such results might link to expectancies, it is possible that students simply remembered expectancy-consistent information better (Rothbart et al., 1979; Harrison, 2001). If students' exposure to past-oriented (evaluative) feedback has been typical of their experiences in HEIs then they will expect to receive this type of feedback again. In spite of evidence that future-oriented feedback is critical for deep learning (Higgins et al., 2002; Hattie & Timperley, 2007; Boud & Malloy, 2013) research classifying feedback has shown that it is rarely embedded into feedback practice (Walker, 2009; Orsmond and Merry 2011; McLean, Bond & Nicholson, 2015). Therefore much of the feedback students receive is evaluative (past-oriented) or descriptive in nature (i.e., you did not make adequate links to theory). This contention is strengthened further by an additional finding of Nash et al.'s (2018) which was that when adult participants were able to recall future-oriented (directive) feedback they often misremembered it in a critical, past-oriented (evaluative) style. However child participants did not make the same errors, suggesting that their more limited exposure to past-oriented (evaluative) feedback has not yet shaped their expectancies in the same way as their adult counterparts.

2.8.2.2 **Impact on interpretation**

Expectancies are also interpreted in line with expectancies (Olson et al., 1996). For example, expectancies based on male and female abilities to write an essay on eye make-up has been shown to influence perceivers' interpretation of academic performance (Biernat & Manis, 1994). Grades awarded to an essay supposedly written by Joan received a higher mark than the same essay supposedly written by John. However when perceivers were asked to provide feedback for various criteria on the essays no differences were apparent across genders. Biernat and Manis (1994) concluded that the student's interpretation was that John's essay, while not as good as Joan's, was good 'for a man'. Other research by Biernat and colleagues (e.g., Biernat, Crandall, Young, Kobrynowicz & Halpin, 1998; Kobrynowicz & Biernat, 1997) has addressed how subjective response scales can mediate perceivers' judgements and make them adjust their judgements alongside a category-specific standard. For example, when using a rating scale to judge how fast Rachel is as a 100-meter sprinter, the subjective meaning of 'fast' may be adjusted in terms of expectancies related to the category 'woman' (e.g. "pretty fast, for a woman"). However, when judging how fast Rachel is in seconds, by timing her with a digital stopwatch, Biernat and colleagues argue that no category-specific subjective calibration can occur, because a second is a second no matter which gender is being timed.

This discussion regarding the adjustment of subjective measures to align themselves with a category-specific standard is of interest when applied to the marking of student work. The marking of assignments is partly subjective and when non-anonymised the marker would be privy to certain categories to which the student belonged (e.g., gender and ethnicity). This might mean that lecturers use category-based knowledge when marking, perhaps considering the assignment "pretty good for an Asian student". Such thinking has parallels with the early work of Asch (1946) who wrote that the judgement of one piece of information is dependent on the information that accompanies it. He termed this process 'interactive effects'. For example, the overall judgement of a target might be different if the perceiver knew they were female and Chinese, versus them simply knowing they were female.

2.8.2.3 **Impact on attributions**

Attributions refer to the inferences a perceiver makes about the causes of events (Hamilton et al., 1990). Attributions about others can be internal and based on the target's dispositional characteristics (e.g., they are lazy) or external and based on situational factors outside of the target's control (e.g., they were in a bad group). A strong bias towards making dispositional as opposed to situational attributions about others is termed the fundamental attribution error (Christensen, Wagner, & Halliday, 2001**.** However, this bias reverses if outcomes disconfirm perceivers' expectancies. In this case they are likely to be attributed to situational factors rather than dispositional characteristics (Higgins & Bargh, 1987). For example, if a student is a high achiever and expected to do well, but then submits a poor essay the lecturer is likely to attribute this to a situational factor (e.g., the student must have pressures in their personal life), rather than as a reflection of the student's conscientiousness (e.g., that they did not apply themselves to the task). Olson et al., (1996) claim that the bias towards situational attributions is preferred by perceivers because their expectancy is confirmed and their structured view of the world is reinforced.

However, the attributional process is also complicated by the influence of an affective bias (i.e., how much the perceiver likes or dislikes the target). When a perceiver likes a target any detrimental mistakes are likely to be attributed as accidental or as indicative of the situation. Conversely the same mistake by a disliked target may be judged as deliberate or intentional, and attributed to stable, negative characteristics (White et al., 1998).

Presumably affective biases can operate at a group level as well as individual level. Therefore it is possible that lecturers who hold expectancies about specific groups of people may dislike the group that they perceive a student's name originates from. Subsequently a negative stereotypic expectancy may be activated which could stimulate a negative attributional process when marking the student's assignment. The consequences of these expectancy effects for the student might be visible in both their grades and feedback.

2.8.2.4 **Impact on counterfactual thinking**

Counterfactual thinking refers to thoughts about 'what might have been' (Roese, 1995), and involves reconstructing the past to create an alternative outcome. Expectancies are said to determine the occurrence and content of counterfactual thoughts. For example, if a student appeared knowledgeable and articulate in tutorials, the lecturer might expect a good assignment from the student. If this is not the case then the lecturer might engage in counterfactual thinking (e.g., 'if they had applied what we spoke about in tutorials then they would have done better').

The same cognitive process is likely to be engaged in by the student when receiving their feedback (e.g., ‘if I had written down what we discussed in tutorials I could have addressed some of the comments made in my feedback’).

There is distinct overlap between attributions and counterfactual thinking since they are both interested in providing explanations for why things happened. Just as is the case with attributions, the further the expected outcome is from the reality the more likely counterfactual thinking is to occur (Roese, 1995).

2.8.3 **Affective Consequences**

The scant coverage of affective responses within Olson et al.’s (1996) model is surprising given that Johnson et al. (2002) point to a wealth of research which suggests that affect and cognition cannot be easily separated. They cite the work of Isen and colleagues (Isen, Shalker, Clark, & Karp, 1978; Isen, Johnson, Mertz, & Robinson, 1985; Isen, Niedenthal, & Cantor, 1992) which, amongst other things, has shown that positive affect prompts positive material in memory and mediates the complexity and flexibility of recalled material.

Despite the limited scope of Olson et al.’s. (1996) analysis, they do contend that affective responses towards targets are influenced by perceivers’ expectancies following interpersonal interactions. This relationship was illustrated well in Dijker’s (1987) ground breaking study which aimed to identify the emotions that the indigenous Dutch population expected to feel when they had to interact with ethnic minorities living in the Netherlands. Participants were asked to rate how often they would experience negative or positive emotions, how they might hypothetically respond towards members of the group in certain situations, and how they generally felt about the group. Findings indicated that participants’ emotional responses to ethnic groups other than their own were more negative, thus illustrating the effect of expectancies on affective states and attitudes for the first time.

2.8.3.1 **Mood state**

One affective state overlooked by Olson et al. (1996) and now considered to have an impact on expectancies is mood state. While mood state has previously been discussed in terms of being a moderator of both information processing strategies and accessibility of expectancies, it also seems that information associated with a particular group may elicit an affective reaction from the perceiver. Known as integral affect (Bodenhausen, 1993) because of the role the group or target plays in eliciting a particular mood in the perceiver, this area of research has been central to many approaches attempting to understand stereotyping and prejudice in relation to negative mood and affective states.

Such research is of importance when these interpersonal processes are aligned to marking student assignments. It would appear as though specific moods or integral affect could be triggered purely by reading a student's name and making associations to gender and ethnicity. Given that perceivers in a good mood are likely to access more positive thoughts and therefore evaluate things more favourably (Schwartz & Clore, 1983; 1996), use schematic processing more readily (Edwards & Weary, 1993; Bodenhausen et al., 2001), be more prone to halo effects (Forgas, 2011), and be less accurate (Sinclair 1988; Sinclair & Mark, 1992), the mood state elicited by integral affect has wide-ranging implications for expectancy effects. In short, how the category the target embodies (e.g., Asian, Chinese, White British) makes the perceiver feel might be fundamental in influencing the grade and feedback they receive.

2.8.4 Behavioural consequences

Interest in expectancy effects is demonstrated most clearly by the abundance of research exploring its behavioural consequences (e.g., Darley & Fazio, 1980; Snyder et al., 1984; Harris & Rosenthal, 1985; Miller & Turnbull, 1986; Jussim, 1986, 1989, 1993; Chen & Bargh, 1997; Snyder & Stukas, 1999; Madon et al., 2004; Jussim & Harber, 2005). Whilst Olson et al. (1996) identified several behavioural consequences, including that people often behave in expectancy-consistent ways, and test hypotheses biased towards expectancy confirmation, the majority of research into behavioural consequences surrounds the self-fulfilling prophecy (SFP). This phenomenon arises when erroneous expectancies on the part of a perceiver impact the future behaviour of a target. Merton defined the SFP as, "… a false definition of the situation evoking a new behaviour which makes the originally false conception come true" (1948, p.195). Also referred to in the literature as behavioural confirmation, the SFP process is said to operate across 3 broad phases (Brophy & Good, 1974; Darley & Fazio, 1980). First, a perceiver must hold a false belief or expectancy about a target. Second, the perceiver treats the target in accordance with their false belief or expectancy. Third, the target interprets the perceiver's behaviour and responds by confirming the perceiver's original false belief/expectancy (Madon, Willard, Guyll, & Scher, 2011).

The self-fulfilling effects that expectancies can have on behaviour have been illustrated extensively (for a review, see Klein & Snyder, 2003), and bridge the interests of psychologists in multiple fields. However, it has most frequently been applied to educational settings. In such a setting the SFP operates in the following way; "… [a] teacher's expectations about a student's future achievement evoke from the student performance levels consistent with the teacher's expectations" (Jussim, 1986, p.429). For example, a lecturer may hold an expectancy that a student is lazy and did not work hard on their assignment. The lecturer provides feedback to that end, and the student behaves lazily and concludes that they are a lazy student. Jussim's (1986) early work further delineates the first part of the SFP process in an educational setting, stating that it is possible for the development of early expectations on the part of the teacher (perceiver) to be based on, a) pre-interaction information with the student (target); b) on superficial characteristics; or c) on a minimal amount of reputation information pertaining to achievement gleaned in initial interactions. It therefore appears that since little direct experience is accessible at this stage of the expectancy formation process, perceivers might be more reliant on indirect experience and other beliefs (Olson et al., 1996). Since a range of moderators have been identified as having the potential to stimulate early expectations including race, ethnicity, social class, gender, and developmental differences (Rist, 1970; Dusek & Joseph, 1983; Chen & Bargh, 1997; Van Matre, Valentine, & Cooper, 2000; Kuklinski & Weinstein, 2001), perceivers might be more prone to using stereotypic-based expectancies at this early stage of the SFP process.

2.8.4.1 **Early Research into the Self-Fulfilling Prophecy**

The most well-known study demonstrating teacher expectancies on student performance was conducted by Rosenthal and Jacobsen (1968). School children were given an IQ test and then approximately 20% of them were randomly labelled by the researchers as intellectual 'bloomers'. Teachers were informed that these children would demonstrate big gains in intelligence over the coming year. In reality, the children labelled as 'bloomers' should not have made any greater gains in intelligence than the other children. As Rosenthal explains, "The only difference … was in the minds of the teachers" (1994, p.177). Nonetheless, at the end of the school year those children labelled as 'bloomers' showed significantly greater IQ scores, leading Rosenthal and Jacobsen to conclude that teacher expectations had evoked the change and that 'Pygmalion' effects existed in the classroom (whereby positive expectations lead to positive performance).

This landmark study served as an impetus for many others which explored Pygmalion effects and the mechanisms through which they might occur. For example, Lanzetta and Hannah (1969) demonstrated that when students were labelled as 'bright' they received more positive feedback following correct answers, and more negative feedback following incorrect answers. They hypothesised that teachers were more motivated to provide feedback to the students labelled as bright as compared to those labelled dull since they believed feedback would have an impact on bright students. Nonetheless, research demonstrating evidence for the SFP was repudiated by research which showed few differences (e.g., Rubovitz & Maehr, 1971; Cooper & Baron, 1977;

Parsons, Kaczala, & Meece, 1982). Therefore it appeared as though the SFP effect was far from ubiquitous.

Despite this equivocal research base many continued to research the SFP. Rosenthal (1973) noted two central factors which mediated expectancy effects; climate, whereby teachers created a warmer climate for 'special' students; and input, where teachers taught additional and more complex material to 'special' students. He also found more informative feedback was provided for 'special' students, and they were given greater opportunities to respond in classroom discussions. Findings from later research also appeared to replicate this early work, demonstrating that teachers were more emotionally supportive of high-expectancy students (e.g., Rubovitz & Maehr, 1973; Chaiken, Sigler, & Derlega, 1974; Jussim, 1986); gave a higher quantity and more positive feedback to high expectancy students (e.g., Weinstein, 1976; Cooper, 1977, 1979; Brophy, 1983; Jussim, 1986); and gave high-expectancy students greater opportunities to demonstrate skills and master complex material (e.g., Rubovitz & Maehr, 1971; Brophy, 1983; Jussim, 1986).

As a result of this research and subsequent meta-analyses (e.g., Rosenthal & Rubin, 1978), many social psychologists concluded that there was a stronger relationship between teacher expectancies and student performance than the other way around (Crane & Mellon, 1978; Miller & Turnbull, 1986; Schultz & Oskamp, 2000). Others went as far as to question whether SFP effects were in fact inevitable (Jones, 1986, 1990).

2.8.4.2 **Moderators of SFP effects**

Although numerous moderators of SFPs exist only those most relevant to this thesis are discussed here.

While moderators of SFPs have generated a large body of research in other domains, its examination within the educational sphere had been rather sparse until recently (de Boer, Bosker, & van der Werf, 2010). Nonetheless early research indicated that student characteristics such as socioeconomic status, ethnicity, and gender might bias teacher expectations significantly enough to demonstrate SFP effects (Rist, 1970; Cooper, Baron & Lowe, 1975; Brophy, 1983). Later research by Jussim et al., (1996) explored the existence of SFP effects on students' gender, ethnicity, and social class, and found greater SFP effects for students from low socio-economic groups. Additionally, the SFP effects for students from African-American backgrounds were large, reaching effect sizes of .4 to .6. However, follow-up studies revealed that the perceived differences held by the teachers as related to ethnicity matched the actual differences in student achievement (Jussim et al., 1996; Madon et al., 1997). Therefore these results were used by Jussim et al., (1996) to argue that teacher expectations are largely accurate. While some might argue that these results could be evidence of an incredibly effective SFP, it is important to consider Merton's (1948) claim that only inaccurate beliefs can be self-fulfilling. Therefore since teachers' beliefs about students' achievements across ethnicities were accurate, these achievements did not reflect the SFP at work.

Jussim et al. (1996) believe that because teachers often have specific information available to them about students' abilities they do not have to base their expectancies on stereotypes. Furthermore, although Jussim (1986) suggested that additional information about students is desirable, he also said that when the *only* information available to a teacher is the social group a student inhabits *and* the teacher is aware of different levels of attainment across such groups it is better to use this information than to ignore it. Jussim's claims regarding accuracy would seem to be borne out by a recent meta-analysis conducted by Sudcamp, Kaiser, and Moller (2012). They found that teachers' judgement accuracy was high, showing an effect size of .63. However, as they identified, this figure is far from perfect and shows that teachers were inaccurate for a substantial amount of the time. Therefore significant opportunities continued to exist for the SFP to operate.

Auwarter and Aruguete (2008) explored the relationship between student gender, socioeconomic status, and teacher expectations, and found that teachers thought students low in socioeconomic status had poorer future prospects. Gender differences found that negative perceptions were stronger for boys with low socioeconomic status than for girls. This demonstrated that contemporary research continued to illustrate that stereotyped social groups might generate stronger SFP effects than others. Research has also shown that teacher expectations and behaviour varied according to student ethnicity (Tenenbaum & Ruck, 2007), with students from specific ethnic minorities more susceptible to SFP effects than others (McKown & Weinstein, 2002; Stiefel, Schwarz, & Ellen, 2007).

Situational factors are also considered to influence the SFP effects. When a target enters a new situation their sense of self-concept is less strong (Jussim, 1993), and in this unfamiliar environment they are more susceptible to confirming perceivers' expectancies (Jussim & Harber, 2005). This is exacerbated further if a target is motivated to get along with the perceiver, make a good impression (Zanna & Pack, 1975; Snyder & Haugen, 1995), and knows the perceiver controls resources that they want (von Baeyer, Sherk, & Zanna, 1981). Therefore first year university students in an unfamiliar environment may be a group who are especially vulnerable to expectancy effects in the form of SFPs. They may be motivated to make a good impression on their lecturers, and will likely recognise the power the lecturer holds. At the same time they may have unclear perceptions about their ability to be academically competent at degree level. Coupled with this, the means by which they will gauge their academic self-concept is through grades and feedback. Since the relationship between academic self-concept and student learning is well established (Marsh 1990; Moller, Pohlmann, Koller, & Marsh, 2009), and there is evidence to support the mediating role of teacher-assigned grades in this process (Trautwein, Ludtke, Köller, & Baumert, 2006) it would seem that students new to university might be particularly vulnerable to confirming perceivers' expectancies.

Interestingly, while results related to self-esteem and expectancy confirmation have been mixed (McNatt, 2000), there is some evidence that self-fulfilling prophecy effects are largest when they match the targets self-concept or self-esteem (Jussim, 1986). For example, if a lecturer provides positive feedback to students who hold a positive academic self-concept, and negative feedback to those with negative academic self-concept, the SFP effects will be large. Such self-verification (whereby people desire to confirm their self-concepts even if this self-concept is a negative one) has been explored within the educational context by Scherr, Madon, Guyll, Willard, and Spoth (2011). Their longitudinal study (across 6 years) examined the mediating effects of self-verification on adolescents' educational aspirations and subsequent academic attainment. Specifically, it examined mothers' false beliefs regarding how long their children would stay in education and what their academic profile would be. They stated that adolescents internalised their mothers' false beliefs about educational aspirations and self-verified those through their actual levels of academic attainment 40% of the time. These results are noteworthy for those who have the potential to influence students' perceptions of self in educational settings. They also suggest that these prophetic effects can endure across time.

2.8.4.3 **Summary**

The previous sections have detailed expectancy formation and expectancy effects and clearly indicated the role that error and bias can play in social judgement processes. Furthermore, the impact that expectancy-induced errors and biases can have on social inequalities has been explored and applied to educational contexts. Nonetheless, the view that stereotypic-based expectancies can create social reality does have its detractors, with numerous researchers keen to attest that such processes can sometimes also be accurate. However, the body of evidence advocating the existence of error and bias still outweighs the alternative.

2.9 CHAPTER TWO: MARKING PROCESSES

The following chapter will examine some of the challenges within marking processes, including marker reliability and halo effects. It will critically discuss research that has explored gender bias and ethnic bias within marking and consider whether anonymous marking is a credible strategy to reduce bias in assessment.

2.9.1 Perceptions of Assessment

Assessment is important. As Boud declared, “Students can, with difficulty, escape from the effects of poor teaching, they cannot (by definition, if they want to graduate) escape the effects of poor assessment” (1995, p.35). These words were underlined by Beadle who cautioned, “Make no mistake: this (*marking*) is the most important thing you do…. All the other stuff is of no use whatsoever if you don’t mark your books properly” (2012, [online]). It is therefore baffling that such an important process lacked research attention until the turn of the twenty-first century (Yorke, Bridges, & Woolfe 2000; Smith & Coombe 2006).

Despite the QAA Code of Practice specifying that “Institutions [should] have transparent and fair mechanisms for marking and moderation” (2006, p.16), confidence in university-level assessment appears to be low. Twenty years ago Newstead described the system as, “… flawed and in need of modification” (1996, p.543). More recently, Race has described the process as “… broken” (2003, p.5), and Knight claimed assessment to be, “… the Achilles’ heel of quality” (2003, p.107). These criticisms remain even with the introduction of a raft of procedures designed to increase confidence and reliability in assessment (e.g., moderation, grade descriptors, marking criteria, external examiners).

2.9.2 Reliability of Marking

One area that has received attention within marking processes is the area of marker reliability. Twenty years ago Newstead (1996) identified huge differences between markers when marking the same assignment. He cited the early work of Hartog and Rhodes (1935) who found a seventy-percent differential between markers of the same assignment. Newstead and Dennis (1994) later found significant variability between the grades awarded by external examiners, with the biggest differences identified as being between an excellent first and a lower second/third class. Despite these troubling research findings, which are supported by more contemporary research (e.g., Heywood, 2000; Meadows & Billington, 2005; Shay, 2006; Yorke, 2008; Crimmins et al., 2016), the issue of reliability has still to be adequately resolved.

One tool used to reduce marker variability and enhance reliability is marking criteria. Criteria usually consist of a sheet which identifies what is valued in the assessment task and provides a means to measure how well the student has done for each descriptor. These criteria were first designed by academics to demonstrate their expectations of the work and share with co-markers and students (Ecclestone, 2001; Sadler, 2005). However the processes behind the development of criteria have been opaque and subject to criticism (Reddy & Andrade, 2010; Panadero & Jonsson, 2013) and students often complain they find the end product vague and ambiguous (Carless, 2006). More recently a move towards the co-construction of marking criteria has gained momentum, and a more dialogic approach to assessment practice is emerging (Christie, Grainger, Dahlgren, Call, Heck & Simon, 2015; Crimmins, Nash, Oprescu, Liebergreen, Turley, Bond, & Dayton 2016).

Early research on gender bias in assessment demonstrated that using marking criteria reduced this bias (Pheterson, Kiesler, & Goldberg, 1971; Terborg & Illgen, 1975). Nonetheless, its ability to reduce variability, and enhance transparency and standardisation has been questioned. For example, Price (2005) found that training colleagues in the use of marking criteria did not reduce marker variability when marking was shared across a module. Similarly, Schaefer (2008) found that even when lecturers were trained to use criteria in similar ways, the cognitive process of marking was far more complex and error prone than the criteria could represent. Part of the problem might be that academics are reluctant to engage with marking criteria. Bloxham et al., (2011) used a 'think aloud' methodology with twelve lecturers and found that most did not use the available criteria when marking work. Additionally, when used at all, criteria were only referred to *after* the assignment had been read, and only to refine or justify thoughts regarding grade allocation. Other research supports Bloxham et al., (2011) in suggesting that markers ignore marking criteria or choose not to adopt them (e.g., Price & Rust 1999; Ecclestone 2001).

Woolf (2004) believes that academics actively and consciously resist the use of marking criteria, believing that it interferes with them being able to exercise academic judgment. Knight and Yorke (2003) claimed that this resistance comes from academics disliking the perception of marking as something simple enough to be marked reliably. They claim that markers draw on 'connoisseurship', the process of being able to use experience to make expert and reliable judgements of student work. Similarly, Ecclestone (2001) contended that markers develop 'mental models' of marking which negate the need for marking criteria. This notion of a mental model is supported by Woolf (1995) who said that academics develop fixed habits which may be unconscious but which impact on the marks awarded. Others have been less damning of academics motivation to engage with marking criteria, but do believe that marker variability continues due to tacit, unconscious beliefs held by the marker (Price, 2005; Shay, 2005, 2008; Tuck, 2012).

It is within the discussion of tacit knowledge, connoisseurship and mental models that the door is left ajar for the re-entry of judgements, expectancies and bias. It seems likely that experienced markers will use fixed habits, mental maps, and implicit judgements when marking which are all commensurate with the use of schema-driven processing. Alternatively inexperienced markers with less well established habits, mental models, and tacit knowledge might pay greater attention to the marking criteria and therefore process information in a more data-driven way. This contention is supported by the work of Dreyfus and Dreyfus whose model of professional decision making suggests that being an expert "… is characterised partly by a declining dependence on rules, routines and explicit deliberation" (cited in Ecclestone, 2001, p.305). Expert markers have also been considered sceptical of criteria despite evidence that their marking required as much moderation as novices (Ecclestone, 2001). Certainly there is some evidence that marker reliability scores are similar across novice and experienced markers regardless of their levels of engagement with criteria (Newstead, 2002). While researchers in assessment have not made reference to schema and data-driven processing in their work, they have acknowledged that experienced markers use tacit knowledge, whereas novice markers are more focused, pay greater attention when marking, and engage more with marking criteria (Price, 2005).

The vulnerability of subjective assessment to judgement and bias is something that that has been identified by those who consider marking to be a social practice (Connell et al. 1992; Mutch, 2003; Shay, 2005; 2008; Tuck, 2012). As Read et al. noted, the perceived quality of an assignment is ultimately constructed by the marker and their "… ways of understanding the world" (2005, p242). Certainly reliability of marking is lower in non-science based subjects where it is harder to distinguish what is entirely correct or incorrect (Yorke, et al., 2000). The degree to which lecturers use implicit biases such as mental maps, fixed habits, tacit knowledge, and connoisseurship to formulate judgements is difficult to research since by their very nature these processes often operate subconsciously. However failing to remedy this issue and therefore allowing markers to exercise their academic judgements unchecked means accepting that the marking process remains susceptible to expectancy effects and bias.

Concerns regarding implicit bias in marking are a real issue since there is evidence that the expectancies formed as a result can influence grades. Higher marks have been awarded to essays thought to have been written by more intelligent students (Diederich, 1974) and students of a higher social class (Darley & Gross, 1983). Stock and Robinson (1987) claimed that perceivers' expectancies play as much of a role as the student's work in determining the grade awarded. Smith summed this up nicely when he explained that,

> … in a sense, information already available in the brain is more important in reading than information available to the eyes from the print on the page, even when the text is quite new and unfamiliar (1982, p. 9).

More recently, Yorke et al. (2000) asked markers a series of questions pertaining to the factors which influenced their judgements when marking student work. These included time pressures, level of interest in the work, the academic level of the work, pressures from senior management and energy levels. Parallels can be drawn here between the factors identified by Yorke et al. (2000) and some of the moderators of schema-driven and data-driven processing. For example, under conditions of high cognitive load (i.e., time pressures; pressures from senior management) perceivers are more likely to rely on their expectancies and process data schematically (Fiske & Neuberg, 1990; Snyder & Stukas, 1999). Additionally, the goal states of the perceiver (i.e., level of interest in the work; the academic level of the work, wanting to be liked) are important in determining whether they will allocate additional effort to process information in a more individuating manner (Blair & Banaji 1996; Macrae et al., 1997; Spencer et al., 1998). Furthermore, biased judgements and reliance upon expectancies is said to occur more frequently when perceivers have to work at non-optimal times of the day (Bodenhausen, 1990). This suggests that marker expectations and biases might be attenuated if HEIs allowed staff sufficient time to mark student work.

Acknowledgement of the specific role that tacit expectancies and cognitive overload play as an antecedent to biased judgements and expectancies has been explored by Hunter and Docherty (2011). Their research suggested that tacit expectations about assessment are so "… deeply instinctive that they cannot be articulated at all given the limited time and resources available to academics" (2011, p113). Results demonstrated that it was only through a process of moderation that these expectations and subsequent grade variability were reduced.

In an effort to reduce variability in marking calls for the implementation of grading scales have been made (Yorke et al., 2000; Rust, 2007). Yorke et al., (2000) advocate this on the basis of doubts that the human mind can effectively discriminate when using large scales (such as percentage scales). This is supported by theories of judgement analysis which claim that integrating numerous cues is challenging (Elander & Hardman, 2002), and by Allport's seminal work which claimed that '… the human mind must think with the aid of categories' (1954, p.20). However, while grading scales may help with issues of variability and reliability how far they could eradicate expectancies and bias is questionable because of their unconscious nature.

2.9.3 **Halo Effects**

The term halo effect was first used by Thorndike in the 1920's and is now a well- researched expectancy-based cognitive bias (e.g., Dion, Berscheid & Walster, 1972; Dennis et al., 1996; Forgas & Latham, 2009; Forgas, 2011). Halo effects are said to occur when, "… the perception of one trait (i.e. a characteristic of a person or object) is influenced by information about another, often irrelevant trait" (Forgas & Laham, 2009, p.54). For example, knowing a student is Chinese might also mean a lecturer perceives them as conscientious. How some information about a target is, a) more influential than other pieces of information, and, b) effects a perceiver's perception of later information was termed trait centrality (Asch, 1946). Given that gender and race are traits so embedded in our psyche that they cannot fail to be activated (Macrae & Bodenhausen, 2000), it is clear that they may be difficult to ignore once they have been primed. Therefore these traits may be more likely to evoke halo effects than others.

Dennis stated that halo effects are important, "… in any situation where ratings are used to provide feedback on performance" (2007, p.1169). Relevance of this bias to the marking of student assignments is therefore worthy of attention. A review of research on halo effects would identify their pervasiveness as equivocal (i.e., Dion, Bercheid, & Walster, 1972, Landy & Signall, 1974; Murphy, Jako, & Anhalt, 1993; Forgas, 2011), however results in education have often suggested their existence. For example, Dennis et al., (1996) claimed that 25-30% of grade differences awarded to student projects could be explained by halo effects. They clarified this by suggesting that the variance could be attributed to factors which influenced the first marker (who had supervised the student and worked closely with them) but not the second marker (who often had little knowledge of the student). Whilst these results appear noteworthy, they are difficult to accept with confidence. Due to the non-experimental nature of the study (which was based on the analysis of archival data), there was no option to include a control group. Without being able to compare between control and experimental groups it is difficult to ascertain whether differences in marks awarded were due to the relationship the project supervisor had with the student or other confounding variables. Given that the concept of hawks and doves (where hawks mark to the left of the normal distribution curve and doves mark to the right) existed when Dennis et al., (1996) conducted this work, it seems short-sighted to omit this limitation.

Furthermore, other potential confounding variables were not controlled in this study. For example, the students name was clearly visible to the markers. Given that research has demonstrated halo effects can operate on the basis that popular student names gain higher essay scores than non-popular names (Harari & McDavid, 1973; Erwin & Calev, 1984) the authors cannot be sure that the second marker (with either no, or very little personal knowledge of the student) did not use other forms of cognitive bias in their marking of the work. There was certainly potential for them to do this since expectancy effects can operate on “thin slices”, or minimal information which can impact upon behavioural consequences (Ambady & Rosenthal, 1992).

An additional antecedent of halo effects which have demonstrated the awarding of higher marks to student work is reputation. Reputation effects would constitute perceivers making target-based expectancies (i.e., expectancies generated from knowledge about the target’s prior behaviour), as opposed to category-based expectancies (expectancies generated from knowledge of the category or group the target belongs to) (Jones & McGillis, 1976). Although reputation effects have been examined as far back as the 1970’s (e.g., Diederich, 1974; Babad, 1980; Rigsby, 1987), Malouff, Emmerton, and Schutte‘s (2013) more recent work claims to provide the first experimental evidence of bias in the assessment of university-level work. Staff and teaching assistants were randomly assigned to watch and then grade a student presentation which was either of poor or high academic quality. They were subsequently asked to mark a written assignment by the same student. Perceivers who saw the student give a high quality oral presentation assigned higher marks to the later written assignment than did perceivers who watched the poor presentation. The authors claimed this demonstrated strong evidence of the halo effect at work. While these findings regarding reputation effects are not unanimously supported elsewhere (i.e., Batten et al., 2011), Malouff et al., (2013) found the difference in grades awarded to be 4%. They used this moderate effect size (Cohen, 1988), to argue for anonymous marking to be implemented in HEIs.

However, Malouff et al.’s., (2013) assertion that the differences in grades were due to halo effects appear less resolute than they have claimed. The criticism once more revolves around the lack of a control group within the methodology, and therefore a lack of ability to control for marker stringency/variability. Without participants first marking a control presentation and essay later interpretations of differences in grades lack reliability. The differences found could simply have been as a result of the subjective, unreliable nature of marking itself. It is feasible for example that participants who saw the better academic presentation gave higher grades to the later written assignment because they are ‘doves’, and had nothing to do with them activating the reputation information to initiate a halo effect. Nonetheless, the potential impact of reputation within the marking process appears not to have been lost on the student body. Carless’ research identified several perceived biases evident in the marking and feedback process, with one student claiming that, “If the lecturer thinks that the student is hardworking or lazy then this will influence the mark” (2006, p.227).

Similar criticisms to those of Malouf et al. (2013) can be levelled at Dennis (2007) who examined the grades awarded over a 4-year period to student projects. Similar to the procedure adopted in Dennis et al.’s earlier work (1996), one marker was the project supervisor and knew the student well, whereas the second marker had no contact with the student. Dennis (2007) claimed that, a) halo effects were at work, b) that 29% of grade variance was idiosyncratic to the grader, and c) project supervisors did not show more halo effects that second markers. However, once again this research was non-experimental being (based on archival data), and therefore could not control for marker stringency/variability. Curiously Dennis (2007) does admit that he took no precautions against unequal standards in the grading process, and notes that variation in marker stringency might be interpreted as a halo effect. Latterly however he dismisses stringency as “… ultimately an empirical issue” (Dennis, 2007, p.1171), which seems rather myopic. These surprising results demonstrated that student reputation and interpersonal affect did not cause the halo effects identified in the study since halo effects were evident across all markers. Dennis (2007) suggested that the knowledge another person would also be marking the same work may have reduced the impact of reputation and interpersonal affect on halo effects, but failed to theorise further about why a continued halo bias was found.

It is therefore evident that a lack of experimental studies exist regarding halo effects related to gender bias in marking. Without the ability to control for the confounding variable of marker variability it makes it very difficult to infer anything of substance from the results. In addition, the small number of papers that have used experimental designs have fallen short methodologically since they have also failed to include a control group and undertake the necessary procedures to attempt to control this important variable. Furthermore, both non-experimental and experimental studies have often failed to explicitly acknowledge whether marking criteria were used. Whilst the inclusion of control groups and marking criteria would not eradicate all problems it would represent an attempt to increase reliability of the work and communicate a stronger commitment to methodological rigour. Nonetheless, Kember (2003) has been critical of the appropriateness of using experimental designs for evaluation purposes in higher education stating that seeking control is both reductive and impossible within more naturalistic research.

As well as a number of methodological shortcomings in the research examining halo effects there also appears to be a lack of clarity regarding the differences between halo effects and stereotypes. Forgas (2011) explains halo effects are different to stereotypes because the initial piece of information is based on the individual's trait or characteristic and not on a generalised characteristic of a group. However, this differentiation is problematic since it casts doubt on whether a number of studies (including his own) have actually been testing for stereotypes as opposed to halo effects. For example, Forgas (2011) explored the role of affective states on halo effects. Participants were asked to read a philosophy essay identified as either being written by an older male professor or a younger female professor. A photo of the professor was attached to the essay. Affective results aside, the male professor was judged as significantly more competent and likeable than the female professor. Of interest here though is that these results were presented by Forgas (2011) as evidence of halo effects, yet how far the participants used information based on the professors' individual traits or characteristics (presumably primed from the photo), versus how far they employed a generalised characteristic related to a group the professors belonged to (e.g., male, female) was not explored.

Additionally, Dennis et al. (1996) and Dennis' (2007) research compared the marks awarded to student projects between the project supervisor and a second marker. Differences in marks awarded were interpreted as evidence of halo effects at work. However, in the 1996 study it was claimed that second markers of student projects varied in their knowledge of the student whose work they were marking, and in the 2007 study it was claimed the second marker would only have had prior contact with the student in about 5% of cases. Therefore, how far the second marker could call on the students' individual traits and characteristics (and therefore exercise the halo bias) is uncertain. Without knowledge of these individual traits and characteristics, it is therefore possible that the second markers mobilised stereotypic information based on other available information (perhaps from the student's name which was visible in both studies). This makes Dennis' claim that "There was no evidence that the students' supervisors showed more halo effects than other graders" (2007, p.1174) invalid since it is arguable whether some markers would have been able to move beyond stereotypic generalised characteristics of a group that the student belonged to.

2.9.4 **Gender bias**

Gender bias has been explored in a variety of different ways and as both a category-based expectancy and a halo effect. It is the most widely researched form of bias in assessment. The original impetus for the research stemmed from the perception that males were evaluated more favourably than females even when ability was equal. Within the most prestigious British Universities; Oxford and Cambridge, there have been notable variations in male versus female attainment for some time. This has led feminist researchers to claim that the entire culture of HEIs is gendered (Reay, 2000; Francis, Robson, & Read 2001; Turner & Gibbs, 2010). Historically, males have gained more First and Third Class degrees whereas females have gained more Second Class (e.g. McCrum, 1994, 1995; Smith & Naylor, 2001; Simonite, 2005; University of Oxford Gazette, 2008). These gender-based differences in overall degree classification at Oxford and Cambridge have also been replicated at assignment level at numerous HEIs across the United Kingdom (Francis et al., 2001; Robson, Francis, & Read, 2002). One of the traditional explanations put forward for such disparity has been that marking is biased according to gender (Bradley, 1984; Goddard-Spear, 1984).

With the exception of Martin (1972) who found female students gained higher marks for composition than male students, the earliest papers in this area seemed to demonstrate a pro-male gender bias in operation. For example, working on the principles of expectancy confirmation and disconfirmation, Goddard-Spear (1984) hypothesised that teachers would give lower marks to female assignments within traditional male science-based subjects. Her hypothesis was upheld. She then replicated this research solely in relation to chemistry, a subject where females had historically underperformed. Participants marked six assignments from three different pupils. Each student's ability was described as either good, average or poor, and names were changed on the assignments to create the gender manipulation. She found that female chemistry students received lower grades than their male counterparts regardless of reported ability level. Nonetheless, more recent findings on gender bias in relation to subject area have been contradictory, with Breda and Ly (2014) finding a pro-female bias in oral examination within mathematics (a male dominated subject), and a negative bias towards them in literature (a female dominated subject). In contrast Enzi (2015) found evidence of bias towards female students in mathematics and for male students in German language. Therefore the research base for gender bias according to subject area is contradictory.

Bradley's (1984) early work in this area explored gender bias in relation to whether the marker knew the student. She examined non-anonymised student work which was marked by a first marker (who knew the student), and a second marker (who had no relationship with the student). She found that second markers awarded marks closer to the mean for females and marked at more extreme ends of the percentage scale for males. Bradley (1984) used the concept of trait centrality (Asch, 1946) to explain these results, claiming that the central trait (i.e., gender) had invoked expectancy effects and bias in the second marker which disadvantaged women. However, Newstead and Dennis' (1990) research was sceptical of gender bias. They criticised Bradley's (1984) work on the basis that she had assumed differences present between the marks awarded by the first and second marker indicated gender bias. Instead Newstead and Dennis (1990) proposed these differences could be attributed to the relationship the first marker had built with the student through supervision. They claimed that this relationship would not be present for the second marker. Subsequently they replicated Bradley's (1984) methodology and found no evidence for the existence of gender bias. Specifically, they cited that, a) there was no evidence that females were marked less extremely (i.e., markers used the same breadth of percentage scales for women as they did for men), b) where markers disagreed, marks were not resolved upwards more for males than females, c) there were no differences between a university using blind marking versus one that did not. Nonetheless, it was true (if not statistically significant), that where markers disagreed about grades awarded, these disagreements were resolved upwards 67% of the time for male students and only 44% of the time for female students.

Bradley (1993) wrote a critical rejoinder to Newstead and Dennis (1990) claiming that their interpretation of having found no bias was short-sighted since they had overlooked the possibility that both first and second markers expectancies might have culminated in gender bias. Instead they mistakenly assumed the first marker would be immune to gender bias, presumably because they knew the student and would therefore see beyond their gender. Bradley (1993) continued that, even if knowledge of the student were to prevent against gender-based category biases being activated, this does not mean the first marker should be held up as bias free. There are numerous other biases that may have been invoked, (i.e., liking of the student, attractiveness of the student, ethnicity of the student), and therefore first markers may simply have exchanged one bias for another. This is something that Newstead and Dennis (1990) should not have been blind to, because their research clearly stated that multiple biases can operate at any one time and may work in opposite directions. Bradley (1993) further claimed that differences in standard deviations across departments (where female work was anonymised in one department and non-anonymised in another), were not reported, and these were sufficiently different to identify gender bias in operation. Additionally, she pointed to Newstead and Dennis' (1990) strategic choice of statistical testing and claimed that a different test would have moved the p value from being insignificant at p=0.084, to being significant at the 5% level.

Dennis and Newstead (1994) claimed that Bradley's (1993) contentions were not well supported and replied with further research which re-evaluated and extended their original data sets. The more recent data found only a small difference in the amount of projects whose grades were marked upwards across genders (44% for males and 42% for females). Furthermore, they examined whether second markers extended marks for male students above female students as compared to the first marker and found no significant difference. Instead for them the most pervasive bias centred on personal knowledge of the student and not their gender. A later paper by Dennis et al., (1996) analysed the marks awarded by two markers on student dissertations. Once again, one of the markers was the main supervisor and had an existing relationship with the student. Although Dennis et al. (1996) found that 30 percent of the variance in the first markers grade was likely to be as a result of their knowledge of the student, they did note that the biases in marking were more influential for males since this bias elevated male student's grades above those of females. Nonetheless, their recommendation for future practice was to avoid "… as far as possible, the assessment of work by those who know the student' (p. 517), therefore somewhat downplaying the gender differences found. More recent research conducted in Swedish schools heeded Dennis' (1996) call when exam scripts were re-graded anonymously and examined for gender bias. No evidence of bias was found (Hinnerich, Hoglin & Johannesson, 2011). However, even when student work has been non-anonymised more contemporary research has shown little evidence of gender bias (Van Ewijk, 2011; Sprietsma, 2013; Birch et al., 2015).

Research by Krawcyzk (2017) has echoed the design of Newstead and Dennis (1990) and Dennis and Newstead (1996). In a study involving 15,000 students he explored grade differences on theses between advisors (who knew the student), and referees (who did not). He concluded that although some gender differences were evident (specifically that male students gained higher grades from referees and females from advisors), these did not necessarily affect the final grade of the thesis. However, they did suggest that assignments not usually marked by two people could be biased against males. Nonetheless, overall they claimed no significant effects across gender, and therefore support the body of research by Newstead and Dennis in this area. Nevertheless, a criticism can be levelled at work examining gender differences as measured by grades awarded by different markers on the same project. Research in social psychology has suggested that the impact of stereotypes is attenuated when participants expect future interactions with a target (Berscheid et al 1976, Neuberg &Fiske 1987, Neuberg, 1989) because stereotypes are revised when more direct information is available (Brophy, 1983). Therefore markers who have gained knowledge of a student through a working relationship may be less likely to be influenced by gender stereotypes than second markers who are not party to additional knowledge. This was a criticism implicit in the work of Bradley (1993) but not fully developed.

Notwithstanding the fact that research on gender bias has been contradictory and ambiguous, there have also been limitations with methodological design, generally though not exclusively associated with early research in the area. For example, several studies employed non-experimental methodological designs which prevented them from including a control group. Therefore conclusions drawn from the data could not account for marker stringency (i.e., Newstead & Dennis, 1990; Dennis et al., 1996; Dennis, 2007; Van Ewijk, 2011; Sprietsma, 2013; Hinnerich, et al., 2015), and subsequently any conclusions regarding gender bias lack reliability. Additionally, many studies on gender bias have either, a) omitted to mention marking criteria at all (e.g., Newstead & Dennis, 1990; Dennis & Newstead, 1994; Sprietsma, 2013); b) stated that marks were awarded for subsections of projects (e.g., introduction, methods) but did not note whether specific criteria were provided to guide their decision-making on each section (Dennis et al., 1996); or c) created marking criteria specifically for the project itself rather than using an established and therefore more ecologically valid set of criteria (Goddard-Spear, 1984; Malouff et al., 2013). In one early study it was even noted that, "… when grading teachers applied their own separate criteria" (Martin, 1972, p.38). Given that the inclusion of standardised task-based criteria had been shown to decrease gender bias long before the majority of these studies were published (e.g. Pheterson, Kiesler, & Goldberg, 1971; Terborg & Illgen, 1975) this omission is regrettable.

While the internal validity of experimental studies in this area have been described as strong, some research has been critical of their external validity. Malouff et al. (2013) claimed that many experimental studies used participants who were not real markers (i.e., teachers or lecturers) to mark the work. Furthermore, they claimed that the work itself was generated for the experiment and therefore was not authentic student work. However, it appears as though the criticisms offered by Malouff et al. (2013) may be somewhat exaggerated. For example, both Martin (1972) and Goddard-Spear's (1984a, 1984b) early work did use experienced teachers for marking purposes although it is true that in both cases the student work was produced especially for the investigation. Furthermore, Malouff et al.'s (2013) criticisms only seem to extend to the older research in the area, since newer research has often used real essays and real teachers to mark work (e.g., Read et al., 2005; Van Ewijk, 2010; Sprietsma, 2013; Hinnerich et al., 2015; Krawcyzk, 2017).

Some research has explored the area of assessor gender on marks awarded to students. In a longitudinal study in American elementary schools Ouazad (2009) found that male teachers awarded higher grades to male pupils, but female teachers did not show any gender-based bias. In HEI settings research on assessor gender has been minimal, but Read, Robson and Francis, (2004) found that female and male markers look for and prioritise different issues. Specifically, female markers seek out information related to presentation and effort more than males who prioritise the student's ability to construct an academic argument. Therefore the issue of gender bias is perhaps not as simple as purely considering the gender of the student in isolation.

Much of the more contemporary research has recognised this and has started to explore how a range of other characteristics combined with gender might impact on teachers' perceptions. Findings have illustrated that, despite the integration of diversity training in the workplace and a growth in awareness of the impact of stereotypes, perceptual biases still remain. For example, Parks and Kennedy (2007) found that teachers rated boys who were Black and unattractive as being the least competent student group. More recently Auwarter and Aruguete (2010) discovered that teachers expected students from low socioeconomic backgrounds to do worse in class, and this finding was exacerbated for boys. How these combined characteristics of a target interact in the mind of the perceiver seems worthy of future investigation. As far back as 1946, Asch wrote of the concepts of trait centrality and interactive effects. Trait centrality describes how some personality traits hold more weight in determining how a person is perceived than others. Interactive effects describe how one piece of information interpreted by a perceiver can be altered by the accompanying piece(s) of information. Applied Parks and Kennedy's (2007) research, it appears that the teachers altered their expectancies of their students depending on whether the descriptor of 'competent' was linked with either, a) male, b) female, c) Black, or d) White, and with either e) attractive, or f) unattractive.

The concept of trait centrality and interactive effects has yet to be fully researched in social psychology literature and has been presented as a complex issue. Macrae and Bodenhausen

(2000), provided some possibilities related to how perceivers contend with targets who fit multiple category memberships. The first is that all relevant stereotypes and biases are activated and acted upon. The second relates to the perceiver engaging in a process of low-level inhibition whereby a competition for dominance between the categories arises (Stroessner, 1996; Bodenhausen & Macrae, 1998). It is hypothesised that those more salient, accessible, and relevant to the perceiver will then be activated and the losing categories are inhibited and removed (Bodenhausen & Macrae, 1998).

In summary, gender bias in marking is the most widely researched bias in assessment, but results surrounding whether a bias exists are equivocal. Importantly however, previous research disputing the existence of gender bias has often suffered from methodological shortcomings which make it hard to accurately gauge how much gravity to afford such claims. Furthermore, much of the research has been school rather than University based. Perhaps as a result of such ambiguous findings research into gender bias has diversified in recent years. It has explored teacher expectations about students' gender and their writing capabilities (Francis, Read, & Robson, 2004), gendered writing styles (Read et al., 2005), assessor gender (Ouazad, 2009), and the interactive effects of gender with other characteristics (e.g., Auwarter & Aruguete, 2010; Krawcyzk, 2017). Ostensibly this diversification has only served to further muddy the waters, though there is some evidence of gender-based expectancies operating within the realm of marking and assessment.

2.9.5 **Ethnic Bias**

Ethnic bias within assessment has been examined less frequently than gender bias, and historically research has also been school as opposed to university focused. In light of claims that marking is biased on the basis of ethnicity when student names are visible (NUS, 1999, 2008, 2012), it is surprising that calls for anonymous marking have not stimulated more research activity at university level.

2.9.5.1 **Ethnic Bias in Schools**

Although Rubovits and Maehr (1973) claimed to have completed the first piece of research on race in the classroom, this title actually belongs to Coates (1972). He manipulated information on children's test performance and subsequently observed the feedback the perceiver chose to award to Black or White students from a collection of pre-defined feedback phrases. Furthermore, perceivers had to complete a questionnaire which judged the personality traits of the children they saw. Findings demonstrated that White adult males used more negative feedback statements and judged the personality traits of Black students less favourably than White students. White females did not differ in the feedback they gave to children, but did judge Black children more negatively on personality traits. Research conducted soon after by Rubovits and Maehr (1973) had a similar focus and found that school teachers not only treated children labelled as ‘gifted’ better, but this preferential treatment was also extended to White students over Black students. Specifically Black students were paid less attention, often ignored, seldom praised, and often criticised. Perhaps most interestingly was the finding that Black ‘gifted’ students were the worst treated of all, even above their ‘non-gifted’ Black counterparts. Rubovits and Maehr concluded that their results indicated evidence of “… white racism” (1973, p.210) in the American classroom.

Compelling though the findings of these pioneering studies were, they were not unequivocally supported at the time (e.g., Kehle, Bramble, & Mason, 1974). Nonetheless, early meta-analyses in the area from Dusek and Joseph (1983) and Baron et al. (1985) which reviewed a wide range of bases for teachers’ expectancies, found that with regard to race, White Americans created more positive expectancies on the part of their teachers than did Black or Mexican Americans. It should be noted however that these early studies were all conducted in America, and therefore how generalizable the findings are to different cultures is difficult to gauge. Some studies also only included female participants (Rubovitz & Maehr, 1973), and most failed to disclose the race/ethnicity of their participants (Coates, 1972; Rubovitz & Maehr, 1973). Furthermore, many of the studies used undergraduate students as participants as opposed to qualified and experienced teachers and therefore the expectancies, perceptual biases, and prejudices of these groups may differ (Burgess & Greaves, 2009; Tenenbaum & Ruck, 2007). For example, some school teachers working in America during this time period will have been exposed to more diverse ethnic groups in their classrooms than undergraduate students would have experienced in the white dominated university environment of the 1970s and 1980’s. Application of Social Identity Theory (Tajfel, 1979) provides an understanding of how in-group bias could operate in this scenario. Generally speaking, individuals draw favourable comparisons with people who occupy the same social group as themselves (in-group) and draw unfavourable comparisons with those who do not (out-group). However, this bias can be reduced through contact with people in the out-group as part of a process labelled Intergroup Contact Hypothesis (Allport, 1954). Accordingly, teachers who have contact with a wide variety of races and ethnicities are less likely to demonstrate differential expectancy effects and treatment of students whose race is different from their own.

Tenenbaum and Ruck (2007) conducted a further meta-analysis of the American-based research. Key findings illustrated European American children's race generated more positive expectations, than African American or Latino children, although the highest expectations were held for Asian American students. An additional analysis also demonstrated that teachers preferred European American children to African American children.

In the small body of research which has examined the existence of grade bias according to ethnicity, van Ewijk (2010) explored whether ethnic majority teachers marked students who matched their own ethnic status higher than those that did not. Conducted in Dutch primary schools and with real teachers, this study manipulated student names and proposed that grades awarded may be influenced by the expectancies teachers held about the ethnicity of a student with that name. How expectancies might affect the grades awarded might be as a result of several moderators of expectancy effects which have been discussed extensively in an earlier part of this chapter. Van Ewijk (2010) found no direct evidence for grade bias, but did note that teachers generally reported having lower expectations and less favourable attitudes towards ethnic minority students.

These findings are interesting in as much as teachers lower expectancies and attitudes did not extend as far as eliciting differences in grades. Perhaps these findings reflect that teachers were involved in the process of category application but not category activation. This would support the earlier work of Devine (1989) who claimed that while all human beings will activate stereotypes, prejudices, and expectancies, less prejudiced people can initiate the process of controlled inhibition and therefore replace the stereotype with non-prejudicial views, thus curtailing biased behaviours. Nonetheless, although van Ewijk (2010) failed to find evidence for bias in grades awarded, other research has found such bias. For example, Ouazad's (2009) longitudinal study included 20,000 pupils in United States schools and found that non-White students marked by a White teacher gained significantly lower marks than if the same work was assessed by a non-White teacher. The bias amongst White teachers was said to explain 22% of the gap between White pupils and other ethnic minorities. While details of Ouazad's (2009) methodology were not available in the paper, it is clear that real teachers and students were used. Furthermore, any concerns that White students grades might have been higher simply because they were more gifted, rather than because teachers were biased were controlled for. Students repeatedly took objectively scored tests throughout the research. Whilst it is true that a students' abilities across assessment type are not always consistent, this does provide some attempt to control for ability as a confounding variable.

Research conducted in German primary schools has also demonstrated evidence of grading bias according to ethnicity. Sprietsma (2013) and Kiss (2013) randomly assigned native (German) or immigrant names to student essays in order to explore whether teacher expectations would culminate in grading differences. They both found that essays identified with an immigrant sounding name gained significantly lower grades. However, the extent to which the bias in Sprietsma’s (2013) research can be exclusively identified as ethnic bias is debatable since the names of students were only manipulated according to ethnicity and not according to gender. An essay written by a girl always bore a girl’s name. Therefore it is difficult to distinguish whether the grading bias was as a result of the gender of the student, the ethnicity of the student, or a combination of the two. There are researchers who have commented on the possibility of individuals being doubly disadvantaged as a result of belonging to more than one persecuted group and eliciting interactive effects on the part of the perceiver. Thomas and Miles (1995) have called this the ‘double jeopardy’ effect.

Burgess and Greaves (2009) provided the first UK-based examination of ethnic bias in school assessment. Having first tested for systemic differences in assessment they found significant differences across ethnic groups. Specifically, Black pupils of African or Caribbean descent were awarded lower marks when their names were visible in comparison to White pupils. Conversely, students of Indian, Chinese, and mixed White-Asian descent were awarded higher marks for non-anonymised assessments. Burgess and Greaves (2009) contended that expectancies about students’ progress are likely to be based on the past performance of other members of that ethnic group. This seems a reasonable assumption when it is considered that all expectancies are derived from beliefs (Olson et al., 1996). These beliefs incorporate an individual’s pre-existing knowledge, and often include interpersonal expectancy-based information about specific social groups (Smith, 1998). The results of this initial research in the UK exploring bias within schools based on ethnic status and other characteristics have also been replicated more recently (e.g., Campbell, 2015).

A meta-analysis of non UK-based research on grading bias was recently conducted by Malouff and Thornsteinsson (2016). The majority of research included was school-based and covered various criteria for bias (including ethnicity). Results suggested that students who belong to stigmatised groups can receive lower grades. However, it also noted that results were homogenous, with some studies demonstrating a reverse bias. Although it is difficult to draw direct comparisons with ethnicity due to a range of biases being explored in this paper, there was evidence of differential treatment of students according to biasing characteristics. Furthermore, this paper did make a link between its findings and implicit biases, demonstrating a rare acknowledgement of the overlap between psychology literature and assessment bias.

2.9.5.2 Ethnic Bias in Universities

Ethnic bias within HEIs has largely ignored grade bias and instead focused on admission processes (Shiner & Madood, 2002; Gittoes & Thompson, 2005; Madood, 2006; Shiner & Madood, 2010; Noden, Shiner, & Madood, 2014), or how the gender and ethnicity of the lecturer might impact student evaluations (Glascock & Ruggiero, 2006; Bavishi et al., 2010). Research conducted using meta-analyses (Malouff & Thornsteinsson, 2016) or archival data (Hinton & Higson, 2017) have explored ethnic bias in relation to grades as part of a wider exploration of causes of bias within marking (e.g., gender, past performance of the student, socio-environmental background), but have not considered it in isolation. Furthermore, Malouff and Thornsteinsson's (2016) meta-analyses included research that was largely school-based therefore making any comparisons to HEI settings difficult.

2.9.6 Anonymous Marking

As far back as 1996, Newstead claimed that a possible solution to bias within marking was to anonymise student work. The most popular method has been to assign students a number which they present at the top of their work instead of their name. Newstead's (1996) idea gained momentum in the late 1990's when the NUS campaign for anonymous marking began. Initially the campaign only lobbied for examinations to be marked anonymously, but latterly they called for this to be extended to both essays and coursework. This has met with some resistance; Professor Frank Furedi at Kent University being amongst the most vociferous. He stated that:

> I can live with anonymous marking of exams. However, when it comes to coursework it is a different matter. Coursework takes place in the context of a relationship that informs expectations and the assessment of outcome. In a sense the marking of an essay represents the continuation of that relationship (cited in Baty, 2007, [Online]).

Recent research has demonstrated that relationships between lecturers and students are stronger when work is marked non-anonymously, therefore lending support to Furedi's claim (Pitt & Norton, 2018). Nonetheless, despite its detractors a number of HEIs have moved towards a policy of anonymous marking.

However, it has proved difficult to gauge how widespread the implementation of anonymous marking has been. Malouff et al., (2013) also appear confused since they noted that student anonymity is both uncommon and that it has been formally adopted throughout the United Kingdom. Wes Streeting, the Vice President for education at the NUS (2006-2008) described its implementation as ‘patchy’. Pockets of research provide clues; such as Fowell, Maudsley, Maguire, Leinster, and Bligh (2000), who reported that 17 out of 21 medical schools in the United Kingdom anonymised undergraduate student assessment. Baty (2007, [Online]) cited Swansea University as having introduced anonymous marking for both coursework and examinations in 2000. However, information available on the University website (Swansea University [Online]) indicated that only examinations needed to meet this requirement, and cited feedback and learning as reasons why anonymity might not be uniformly applied. Current QAA guidelines (2013 [Online]), allow Universities to exercise discretion, outlining that HEIs may wish to consider which forms of assessment it applies to. The evidence of its adoption outside the UK is also unclear, suggesting that research addressing how comprehensively anonymous marking has been adopted is long overdue. The picture is further complicated by the fact that many HEIs choose a pick-and-mix strategy whereby some assignments are marked anonymously (usually examinations) and others are not.

The catalyst for anonymous marking was influenced by the NUS (2008) claim that 44% of students’ unions believed that marking was biased according to gender and ethnicity. To support this they cited Belsey’s (1988) research (at University College Cardiff), which used archival data (from between 1977-1981), and revealed that when marking was non-anonymised 42% percent of men and 34% of women gained either a First or a High Second. After the introduction of anonymous marking, these figures were 42% and 47% respectively. Faculty members were not convinced that these results reflected evidence of gender bias at work and stated that women were simply less committed, easily distracted, and less interested in research than men. Belsey (1988) noted that unsurprisingly these type of comments heightened the conviction among some staff that seeing a female name on an assignment did not improve her chances of doing well. Nonetheless, the NUS report failed to identify other research available at the time which demonstrated opposing findings. For example, longitudinal research by Perry-Langdon (1990) found little evidence of female students grades changing pre and post implementation of anonymous marking.

The NUS report also identified that marking was biased according to the ethnicity of the student. It cited two Universities; the University of East London (UEL) and the University of Glasgow (UOG) Dental School, where non-anonymised student work saw Black students do worse than Whites. Specifically, Black students’ grades at UEL were on average 4.2% lower than White students. At the UOG, Asian students represented 20% of the degree cohort, but also constituted 80% of those who failed. While there was no explicit evidence that these figures represented ethnic bias at work both universities subsequently began to mark anonymously. There was evidence from school-based experimental studies that Black students receive either lower (Piche, Michellin, Rubin, & Sullivan, 1977) or higher scores (Fajardo, 1985) than their White counterparts, but at the time that the NUS wrote their report there was no evidence of minority students grades changing pre and post implementation of anonymous marking.

Therefore, despite performance differences in assessment being researched in relation to a host of biasing characteristics, only a handful of studies have explored the efficacy of anonymous marking in reducing such differences in a higher education context. Belsey's (1988) early research was followed by Shay whose work detailed how the University of Cape Town implemented anonymous marking in 2003. The aim was to minimise;

> …the possibility that irrelevant inferences be subconsciously used to discriminate for or against students, in particular inferences based on gender, race and any other kind of information which can be made on the basis of a student's name (2008, p.161).

After the implementation of anonymous marking, mean examination scores decreased from 62% to 55% and score distributions also fell by up to ten percent. Exam failure rates also increased significantly from 7% to 33%. One of the suggestions for the lower grades was that student anonymity had liberated markers from feeling that they needed to give disadvantaged students the benefit of the doubt. Therefore anonymous marking was said to eradicate a positive feedback bias which had previously resulted in the sympathetic marking of disadvantaged students work. Subsequent interview data supported this view revealing that positive discrimination and bias was widespread.

> You note the name and think the language isn't going to be good. And with that you'd have an element of, you know, how would I do in a second language? Here's somebody carrying two bags of cement on their shoulders, not one… And so you go a bit easy (Shay, 2008, p.163).

Australian based research by Owen et al. (2010) explored gender bias in marking when examination scripts were anonymised and non-anonymised. In contrast to Shay (2008), they found no evidence of systematic bias and concluded that anonymous marking was not required. However, the most extensive examination of the impact of anonymous marking on grades at University level was conducted at Aston University in the UK. Hinton and Higson (2017) used archival data from 31,710 students across a twelve-year period (2000-2001 and 2012-2013,) and analysed the demographic data for each student alongside all of their summative assessment grades. Aston University implemented anonymous marking for all assessment types in the 2005/2006 academic year, thus allowing grade comparisons to be drawn pre and post anonymous marking. While some changes were reported for gender, ethnicity, and socio-economic status these were negligible in practical terms. They concluded that,

> Despite the supporters of anonymous marking claiming that its implementation has led to fairer assessment in Higher Education, the present study suggests that anonymous marking initiatives – at least in the present case – have done little to eliminate between-group mean performance differences (Hinton & Higson, 2017, p.12).

Notwithstanding these results, they claimed it would be 'reactionary' of them to recommend that HEIs using anonymous marking should alter their practice on the basis of one study. Instead they called for more research exploring the impact of anonymous marking on grades. Pitt and Winstone (2018) answered that call the following year. Although the flavour of their research was more feedback than grade-focused, they reported that anonymous marking did not differentially effect specific groups of students over others in terms of grades. Therefore the most recent research seems to point to there being little evidence of implicit bias in the marking process related to personal characteristics. While these findings are welcome, they do only represent results from three universities and therefore how commonplace they are has yet to be established. Moreover, Hinton and Higson's (2017) research was conducted using archival data. The nature of such non-experimental research has been criticised previously in this thesis, specifically pertaining to its inability to account for confounding variables.

If evidence of actual bias proves to be less widespread than might have first claimed, the most important function of anonymous marking might be to reduce perceptions of bias for students. This was acknowledged by Owen et al. who commented that the perception, "…can be as undermining of…confidence as actual bias" (2010, p.18). The concept of fairness has been explored more robustly by Pitt and Winstone (2018). Their student participants experienced both anonymous and non-anonymous marking over a semester. Statistical analyses revealed that overall students did not perceive anonymous marking to be fairer, although female students did believe that they were treated more fairly when work was anonymised as opposed to non-anonymised. This notion of being treated fairly is important since it may actually increase performance in similar ways as has been evidenced in job performance (Hinton & Higson, 2017), and can help in creating better relationships between students and lecturers.

Nonetheless, even if HEIs introduce anonymous marking to bolster confidence in the assessment process through reducing perceptions of bias there are still many forms of assessment where anonymity is impossible. For example, in the case of oral presentations, practical assessments, and dissertations, students are identifiable, and consequently the potential for expectancy effects and perceptual biases to exist would remain high. Therefore perhaps at best students can only perceive assessment to be free from bias for some of the time.

More important than the logistical arguments are those of a pedagogical nature. The principal area of contention within the anonymous marking debate surrounds the concept of feedback and generally suggests that marking anonymously would reduce quality. Whitelegg (2002) was the first to voice concerns, claiming that anonymous marking would disrupt the feedback loop, depersonalise feedback, and therefore damage the dialogue between student and lecturer. This is concerning in light of the growing body of evidence which suggests that student's value personalised feedback (e.g. Ferguson, 2011; Bols 2013; Laryea 2013; McArthur & Huxham, 2013) and that depersonalised feedback can create feelings of detachment (Pitt, 2017). Feedback on non-anonymised work has been considered more relational (Price et al., 2010) and has been perceived by students to have a greater impact on learning (Pitt & Winstone, 2018). Moreover, as Whitelegg (2002) first speculated, a lack of personalised feedback has been shown to weaken the relationship between lecturer and students (Pitt & Winstone, 2018).

> With essay work I think [anonymous marking] reduces you to a distance from the lecturer and you don't form any kind of relationship where you are able to have feedback. With essays if you have had your mark and your feedback you feel as though there is a connection (Whitelegg, 2002, p.2).

However, Brennan (2008) is critical of these arguments and stated that although it might be comforting for tutors to provide personalised feedback, and this might meet student's needs, the question is one of priority. This priority revolves around whether HEIs privilege tutor and student preferences over attempts to eradicate systemic bias in assessment. He further stated that strategies to address the feedback issue could be implemented. For example, students could only be allowed to submit their next assignment once they have met with their tutor to receive feedback on their previous assignment. Nevertheless, the feasibility of this for modules with high numbers of students and time-poor lecturers is debatable. Furthermore, weaker or shy students may not feel able to meet with a tutor (Whitelegg, 2002), and following Brennan's suggestion to its conclusion their inability to do so would prevent them from submitting subsequent assignments and invoke academic penalties.

Other options from Whitelegg (2002) included marking anonymously but providing feedback non-anonymously (although no indication about how this might be implemented was forthcoming), and to de-anonymise student work once marking is completed, although the additional administrative burden this entails would likely make it unpopular. Furthermore, the ability for tutors to de-anonymise work has led to claims that anonymous marking is a farce. In a revealing report, Maclellan (2001) claimed that marking is often not truly anonymous. She reported that although 39% of tutors stated that student work was submitted anonymously, 43% claimed that anonymous marking never happened. Unfortunately her research did not explain how this was possible. However, later work by Bloxham et al. supported the problematic nature of de-anonymisation when they quoted a tutor saying, “At this point I can de-anonymise it *(the assignment)* to see who’s the lucky recipient” (2011, p.666). The ability for tutors to easily de-anonymise student work compromises procedures designed to uphold anonymity in the first place. Furthermore, it leaves the door ajar for the potential for inequitable and expectancy-based biases to be perpetuated at the latter stages of the marking process. Tutors could de-anonymise the assignment and then upon seeing the student’s name re-evaluate the grade and feedback they have just given.

For a range of pedagogical and practical reasons the controversy over anonymous marking has polarised academics for many years, and it is clear to see that huge variations in practice are in operation across the sector. NUS (1999, 2008, 2012) campaigns were instrumental in the revival of this debate and their earlier publications included interpretations of research which definitively supported their argument. However, it is clear that much of the research they based their evidence on was non-experimental and lacked both methodological rigour and external validity. A more balanced view of the existing research at the time of their report was that it was equivocal (e.g., Newstead & Dennis, 1990; Bradley, 1984; Perry-Langdon, 1990). Latterly the NUS (2016 [online]) acknowledged the dearth of contemporary research addressing anonymous assessment and the lack of conclusive evidence regarding its implementation. They consequently called for further research. However this further research has shown little evidence that anonymous marking reduces performance differences according to personal characteristics (Hinton & Higson, 2017; Pitt & Winstone, 2018), and therefore indicates that either bias does not exist, or is not generated by these factors. These findings may force the NUS to re-evaluate their position in the anonymous marking debate.

2.10 CHAPTER THREE: FEEDBACK

Feedback has been defined as "… information provided after instruction that seeks to provide knowledge and skills or to develop particular attitudes" (Hattie & Timperley, 2007, p.102). It has been identified as the most comprehensively documented contributor to student achievement (Hattie, 2009). The purpose of feedback is to aid students' understanding about what they were meant to learn, identify how well they did this, and to provide information about how they might bridge the gap between actual and desired performance (Sadler, 1989; Li & DeLuca, 2014). As such feedback can be classified as having both evaluative and educative functions (Dochy & McDowell, 1997; Hattie & Timperley, 2007). Numerous meta-analyses and reviews of feedback having underlined its centrality to student learning and achievement (Hattie, Biggs, & Purdie 1996; Hattie & Jaeger, 1998; Hattie & Timperley, 2007; Li & De Luca, 2014), and it is an important indicator of the quality of the student experience at university (Higgens, Hartley, & Skelton, 2001). Until quite recently it was claimed that, the examination of feedback specifically within HEIs remained relatively small (Weaver, 2006; Walker, 2009; Sadler, 2010). However Li and De Luca's (2014) work on assessment feedback reviewed 37 empirical studies published between 2000 and 2011 which were all conducted with undergraduate students, therefore suggesting that the body of research is less scarce than has been previously indicated.

Nonetheless, it is true that research on feedback has been diverse and has broadly included, a) the examination and classification of feedback types in order to ascertain their usability (e.g., Mutch, 2003; Hyatt, 2005; Brown & Glover, 2006; Walker, 2009, Nicol, 2010), b) principles of good feedback (e.g., Gibbs & Simpson, 2004; Nicol & Macfarlane-Dick, 2006, c) positive and negative feedback practices (e.g., Nicol & Macfarlane-Dick, 2006; Price, 2010; Nicol, 2010, Boud & Malloy, 2013; Small & Attree, 2015), d) students' perceptions of feedback (e.g., Weaver, 2006, Lizzio & Wilson, 2008; Poulos & Mahoney, 2009; Sadler, 2010; Ferguson, 2011; Pitt & Norton, 2016), and e) strategies to improve feedback (e.g., Rust, 2003; Nicol, 2010; Sadler, 2010; Orsmond & Merry, 2011; Yang & Carless, 2013). It is important to note at this point that the feedback being explored here specifically relates to the written feedback on a students work and not feedback they may gain at other times, such as in tutorials, teaching sessions, or other more informal situations.

Reasons for the recent growth of research into feedback within the university sector are multifaceted. However, the massification of higher education and a culture of consumerism has placed a reliance on written feedback that was not previously required, and may be catalysts for this shift. This is a point emphasized by Higgens et al., (2001) who noted that as a result of increasing student numbers and staff workloads, face-to-face contact time between students and tutors has decreased. This in turn has led to the increased reliance on written communication as the primary form of feedback. Support for this contention is evident in Hyland's (2000) research which found that 40% of students had never had a tutorial about their assessed work. However, not only have opportunities for students to learn from different modes of feedback diminished, but the quality and quantity of the written feedback they receive may also have suffered. This is a point championed by Boud and Malloy (2013) who contended that modularised teaching structures reduce the time frame within which feedback can be provided. Therefore students get less practice and feedback about their performance. Furthermore, detailed knowledge of students and their work over time, combined with multiple opportunities to view student work, has been eroded by summative assessment practices. These reductions in type, quality, and quantity of feedback have worrying implications for students' ability to learn through the feedback process.

Another reason for the increased impetus of feedback focused research in HEIs might be the introduction of the National Student Survey (NSS) in 2005 and the concomitant publication of those results. Designed to assess the quality of degree programmes across seven areas, the domain of feedback has consistently attracted the lowest scores (Nicol, 2010; Boud & Malloy, 2013; Bols, 2013). This is not an issue confined solely to the United Kingdom since research in Asia and Australia have shown similar trends (Rowe & Wood, 2008; Krause, Hartley, James & McInnis, 2009).

This section will now explore attempts to analyse and classify feedback in order to ascertain its usability for students; outline positive and negative feedback practices; discuss student perceptions of feedback; and examine biased feedback in relation to both gender and ethnicity.

2.10.1 **Classifying Feedback**

A number of attempts to classify feedback have been undertaken in the hope of developing an evidence-based platform upon which to develop feedback guidance to students (e.g., Ivanic et al., 2000; Hyland, 2001; Mutch, 2003; Brown et al., 2003; Hyatt, 2005; Brown & Glover, 2006; Kumar & Stracke, 2007). However, these have been criticised by Mutch who lamented:

> The evidence presented (on feedback) is often sparse, the reasons for presenting it are rarely spelt out and … seem to relate more to the memorability of the example than to any structured approach (2003, p.26).

He cited several studies which made bold claims based on very small (or undeclared) sample sizes and a lack of structured analysis (e.g., Falchikov, 1995; Ivanic et al., 2000; Thorpe, 2000), and further identified that comments such as 'most tutors' and 'frequently' were unhelpful. He then called for a more systematic approach to examining tutor feedback and subsequently examined comments on undergraduate written assignment feedback sheets. Mutch (2003) hoped that his research would direct attention to significant parts of practice beyond merely describing salient expressions.

He examined the text in three ways; by modality, area of concern, and developmental content. Findings revealed that tutors provided feedback on a full range of issues (although there was a bias towards knowledge and understanding). Feedback was effortful and considered, but often presented in a categorical manner (i.e., commanding students to do things) which was considered to influence how it might be interpreted. Additionally, Mutch (2003) identified a conversational style of feedback which conveyed messages of 'implied development'. Such messages were considered prone to ambiguous interpretation by students who may not recognise what is being implied through such academic discourse.

Whilst Mutch's (2003) criticisms of earlier work are not unfounded, and his research was laudable in its attempt to develop the field, it also suffers from several limitations. For example, Mutch (2003) acknowledged just over half of the assignment sample had been returned to the students. Therefore only analysis of summary feedback sheets was possible for those assignments and the opportunity to analyse in-text feedback was lost. Resultantly, it is debatable whether a holistic and representative account of feedback practice was possible; especially since in-text and summary comments have been revealed to include different feedback functions (Kumar & Stracke, 2007). In a related point, summary comments made on feedback sheets may also have lacked context without having read the preceding comments on the scripts. It is also pertinent to question the extent to which this research adds to the discussion on feedback since Mutch negated to make any claims beyond the fact that, "…most academics in this sample were trying their best, often in difficult circumstances, to give helpful feedback to their students" (2003, p.24-25). He also failed to develop a taxonomy for other researchers to analyse feedback practice, and noted himself that his findings were "disappointing" (Mutch, 2003, p.24).

One of the earliest classification systems for written feedback was devised by Brown et al., (2003) (see appendix i). Designed for use in science-based disciplines, it identified eight feedback categories: identifying errors; correcting errors; explaining misunderstandings; demonstrating correct practice; engaging students in thinking; suggesting further study; justifying marks; and suggesting how to approach future assignments. 'Giving praise' was also added later (Glover & Brown, 2006). Although this system has been used in research by Walker (2009) and Orsmond and Merry (2011), the feedback categories identified by Brown et al. (2003) appear not to have emanated from an analysis of tutor comments, but rather from a set of guidelines outlining what constitutes good feedback practice. Thus using Brown et al.'s (2003) more deductive categories to analyse feedback data could be argued to be more about exploring whether tutors are adhering to good feedback practice as opposed to reflecting and examining their *actual* practice in a more inductive way. It is therefore likely that Brown et al.'s (2003) framework is insufficient to categorise all existing feedback on assignments and therefore using it would only reflect a partial view of the data.

Nonetheless, research which has adopted either Brown et al.'s (2003) original framework or the updated version (Brown & Glover, 2006) has revealed some interesting results. For example, in her work on usable feedback with students Walker (2009) found tutor comments on content were most common (41%), followed by motivational (32%) and skills development based comments (21%). Under the theme of content and skills development the largest proportion of comments were for corrective feedback and did not provide an explanation for how to correct the error. Worryingly, a third of students in her research claimed that comments on their assignments did not help much if at all. Conversely, Orsmond and Merry (2011) who used Brown et al.'s (2003) original framework found that 'giving praise' was the most used category. This contradicts both Walker's finding above (2009) and Orrell's (2006) work which demonstrated that feedback was largely error-focused, but is consistent with the work of Glover and Brown (2006). Orsmond and Merry's (2011) findings also contradicted Walker (2009) since their results demonstrated that tutor feedback explained errors and misunderstandings to students instead of simply highlighting errors. However, they agreed with the finding that feedback was largely assignment specific and focused little on future assignments or feeding forward.

Kumar and Stracke (2007) adopted a more inductive approach in their analysis of feedback on a PhD thesis. Although their paper identified the work as an interim analysis, it coded in-text and summary feedback and developed a taxonomy of practice. They based their coding on the functionality of feedback (i.e., what feedback comments do) (see appendix ii), but despite lengthy discussions within the research team some data still required double or triple-coding when a consensus could not be reached. They identified three categories of feedback; referential, directive, and expressive, and found evidence that in-text and summary feedback comments were often presented in different styles. For example, in-text comments included more referential (i.e., presentational, organisational and content), and directive (i.e., suggestions, questions and instructional) feedback whereas summary comments were dominated by expressive feedback (i.e., praise, criticism, and opinion). However, despite them analysing all feedback comments and providing a potential framework for future use, their analysis does only extend to one PhD thesis. Research using larger numbers of assignments to create a framework are more likely to generate a model that is representative and transferable to assignments outside of the original research data.

Orrell (2006) adopted more qualitative methods in her examination of feedback practice. In study one she used a think aloud approach with 16 academics while they marked student work. In study two she interviewed those academics about their views on assessment and in study three she compared their academic behaviour (gleaned from study one) and contrasted this with their views on assessment (study 2). However, while the think aloud approach uncovered some interesting points, it would have been an inappropriate method to explore biases and expectancy effects within this PhD due to the implicit or unconscious nature of these cognitive processes (Allport, 1954; Devine, 1989; Greenwald & Banaji, 1995; Bargh, Chen and Burrows, 1996; Bargh, 1997; Chen & Bargh, 1997; Greenwald & Krieger, 2006). Orrell's (2006) research findings provided a rather gloomy portrait of assignment feedback, noting that it was largely error-focused and therefore damaging to students' egos; a finding that was later supported elsewhere (Walker, 2009). Nonetheless, such negative findings do sit in opposition to the findings of other research in the area (e.g., Mutch, 2003; Hyatt, 2005; Glover & Brown, 2006; Orsmond & Merry, 2011) thus demonstrating the difficulties with trying to ascertain what feedback looks like. These inconsistencies in feedback provision have been bemoaned elsewhere (e.g. Duncan, 2007; Lizzio & Wilson, 2008; Poulos & Mahoney, 2008).

Hyatt's (2005) research used data generated from a corpus of sixty 6,000 word Masters level assignments to provide arguably the most comprehensive series of functional feedback categories to date (see appendix iii). Fundamentally concerned with power relations within HEIs, and therefore cognisant that writing is a social practice, Hyatt (2005) outlined that students are often confused by feedback terminology. Conventions of academic discourse which are synonymous with feedback practice have been cited as unintentionally exclusive (e.g., Lea & Street, 1998; Mutch, 2003; Carless, 2006; Jonsson, 2013), and the alienation of students through the use of linguistic feedback practices which represent and reinforce institutional identity and power is a significant barrier to student learning. Hyatt (2005) therefore hoped that his representation of the ways in which tutors provided feedback might help them to reflect on how useable it is for students and raise awareness that writing does not take place in a vacuum. While he was careful to note that his work on educationally-based scripts may not presume a wider commonality, the categories generated were considered sufficiently uniform for this framework to be adopted as the taxonomy of analysis for this thesis. His findings demonstrated that comments on content and stylistic-related comments dominated over developmental comments, thus reinforcing the findings of other research in the area (Glover & Brown, 2006; Walker, 2009; Orsmond & Merry, 2011). Moreover, Hyatt (2005) identified a relationship between markers who used imperatives in their feedback (i.e., should, must, have to) and a large amount of stylistic (44%), content (34.5%) and structural (17.8%) comments. He explained these findings as being illustrative of the territories within which tutors applied the power structures of academic discourse.

2.10.2 **Positive and Negative Feedback Practices**

Boud and Malloy (2013) cited Nicol and McFarlane-Dick's (2006) work as the most influential account of feedback practice in higher education. Generated from a synthesis of the literature, their seven principles of good practice are grounded in the model of self-regulation, and suggest that feedback should facilitate students in the self-regulation of their own performance. Specifically Nicol and McFarlane-Dick (2006) stated that good feedback practice:

1. helps clarify what good performance is (goals, criteria, expected standards);
2. facilitates the development of self-assessment (reflection) in learning;
3. delivers high quality information to students about their learning;
4. encourages teacher and peer dialogue around learning;
5. encourages positive motivational beliefs and self-esteem;
6. provides opportunities to close the gap between current and desired performance;
7. provides information to teachers that can be used to help shape teaching.

However, whilst knowledge about good feedback practice is desirable and might be a useful reflective tool for tutors, it is arguable how far Nicol and McFarlane-Dick's (2006) work constitutes an outline of feedback purpose or *practice* versus feedback *outcomes*. As such it becomes important to consider how tutors' might facilitate those outcomes through their feedback. For example, what type of comments might facilitate self-assessment or encourage positive motivational beliefs and self-esteem? This is something that has been considered by Thorpe (2000). He stated that good feedback included; interest or empathy towards the student; acknowledgement of self-revelation by the student; questions to prompt further thinking; approval and praise; suggestions for addressing problems; reinforcement of the student's approach; recognition of strengths; offers of further help or contact, and questioning ideas which the tutor thinks misguided. Furthermore, good feedback should be delivered in a timely fashion while it maintains relevance and can be integrated into future learning opportunities (Gibbs, 2010). The issue of integration is an important one since Carless (2006) found that students could not benefit from tutors' comments if they were too specific to a particular assignment.

Guidelines on feedback have arisen as a result of research findings suggesting feedback practice has fallen short of delivering on these aims and outcomes. As has been identified through an examination of feedback classification schemes, much feedback is focused on content and stylistic elements as opposed to developmental comments which can feed-forward into future learning (Glover & Brown, 2006; Walker, 2009; Orsmond & Merry, 2011). Additionally, tutors have often relied on feedback which uses imperatives, thus aligning their message with one which tells the student what they should do. Lea and Street (2000) have argued that this only serves to reinforce the power at play in academic institutions and can interrupt the opportunity for learning to occur. Researchers have also suggested that students do not always understand the feedback they are provided with since tutors rely on academic discourse and conventions which students have not been inducted into (Lea & Street, 1998; Mutch, 2003; Carless, 2006). In an attempt to address this, marking criteria or rubrics have been suggested both as a strategy to induct students into this academic discourse and to standardise feedback practice and expectancies between tutor and student. However, these in turn have been criticised because they lack personalisation (Agius & Wilkinson, 2014) and are perceived by students to demonstrate a lack of respect (Lizzio & Wilson, 2008) and be vague and ambiguous (Carless, 2006). To further muddy the waters, research has also demonstrated that the personalisation of feedback can impact negatively on students' motivation and affective responses such that their attention is diverted from the feedback and they fail to adequately process it (Skipper & Douglas, 2012). These equivocal research findings make it difficult for HEIs and tutors to know which feedback processes to adopt since, "… the general picture is that the relationship between its form, timing and effectiveness is complex and variable, with no magic formulas" (Sadler, 2010, p. 536).

2.10.3 Student Perceptions of Feedback

In an attempt to better understand feedback effectiveness research attention has extended to include the student voice. Sadler (2010) considers this to be a reflection of the movement towards a more student-centred approach to higher education, although Gibbs (2013) maintains that it might also reflect a concern over NSS results on assessment and feedback. Certainly the NUS was sufficiently concerned with NSS results to produce a charter on feedback and assessment in 2010. This charter included calls for formative feedback; having access to face-to-face feedback; receiving feedback an all types of assignment; timely feedback (i.e. personalised feedback within 3 weeks and group feedback within 1 week of submission); anonymous marking; support in critiquing work; guidance on understanding marking criteria; and a choice over the format of feedback.

However, despite the NUS charter, which might arguably standardise HEIs approach to feedback, research has discovered that student perception and interpretation of feedback; its usefulness, worth, usability and other indices of impact have been variable and contradictory (see Li & De Luca, 2014 for a comprehensive review). For example, whilst Higgins, Hartley and Skelton (2002) consider students to be conscientious consumers of education, with 97% of students in their study reading their feedback and 82% paying close attention to it, their interview data revealed that students generally spoke of negative experiences with feedback. Similar contradictions can be seen across the literature. Hyland (2000) and Ding (1998) claimed students valued and desired feedback while Orsmond et al. (2005) revealed that students considered it to be meaningful. However, other research highlighted that students do not read feedback (Hounsell, 1987; Crisp, 2007), might not understand or use it (Lea & Street, 1998; Gibbs & Simpson, 2004; Winstone et al., 2016), view it as a luxury rather than a necessity (Senko, Belmonte, & Yakhkind, 2012) and believe it has little impact (Sadler, 2010). This led Poulos and Mahoney (2008) to claim that students do not hold a homogenous view of feedback. Given the time invested in the provision of feedback many of these results are both demoralising for tutors and damaging for students.

However, given that there is a body of research which suggests that students are engaging with feedback, perhaps the judgement that it has a limited effect on learning might be due to a lack of understanding as opposed to a lack of engagement. This is something that has been addressed in the literature. Price et al. (2012) have been keen to note that feedback can only be effective when the student understands it and is willing and able to address it. Yet many researchers have noted that students have difficulty in making sense of comments (Norton & Norton, 2001) and interpreting academic discourse (e.g., Lea & Street, 1998; Orsmond, et al., 2000; Jonsson, 2013; Winstone et al., 2016). For example Weaver (2006), in her examination of whether feedback was usable for students found that a large percentage (the exact number was not specified), did not understand common phrases used in feedback. Specifically, 37% could not interpret what was meant by a request for more critical reflection, and 33% did not understand the judgement that their analysis was superficial. Furthermore, 35% did not consider the comments clear or easy to read. Whilst these figures do reflect a large percentage of students who are able to interpret, understand, and use the feedback they are given, it still leaves roughly a third of students who attend to but do not understand the comments provided and whose future learning is potentially compromised as a result.

More recently Winstone et al. (2017) involved 31 undergraduate psychology students in focus group activities which explored how students used feedback. Their results clustered around the four psychological processes of awareness, cognisance, agency and volition to explain that students struggled to understand and decode feedback, did not understand how to implement it, felt disempowered, and were not always receptive to feedback. Given that feedback can only be effective if it used, these results paint a bleak picture.

When students have specifically been asked what types of feedback they find helpful or unhelpful, research findings seem to have generated a number of common themes. Weaver (2006) and Carless (2006) found that students seemed to dislike comments which were too general or vague, lacked guidance, were negatively oriented, unrelated to assessment criteria, and provided no guidance on how to improve. Price, Handley, Millar and O'Donovan (2010) echoed these findings with relation to negatively-oriented comments and also added that students disliked ambiguous feedback, tick box feedback, and illegible feedback. Given that Walker's (2009) research identified corrective feedback with no guidance for improvement as the most frequent type of feedback, and Walker (2009) and Orrell (2006) identified that feedback was predominantly error-focused, there seems to be a mismatch between what students want and what they receive.

Ferguson's (2011) research asked students to rate their preferences for different types of feedback. He found that students did not view feedback on the details e.g., spelling, grammar, referencing etc. as important or useful. Instead students preferred feedback on the approach to and structure of the work, and on the specific ideas examined. To emphasise this point, the most frequent comment from students was; ‘Focus on the fine detail is not useful, what is needed is an explanation of how to improve’ (Ferguson, 2011, p.56). Students also conveyed that one word, or short responses (e.g., very good, structure, expression) were unhelpful, and in line with Price et al. (2010) identified that ticks and crosses, though common, were ineffectual. Furthermore, students wanted to read feedback that was encouraging and motivational. When student perceptions are compared with research which has highlighted that stylistic and content-related comments dominate over developmental comments, the discrepancy between student needs and actual feedback is evident once more (Hyatt, 2005; Glover & Brown, 2006; Walker, 2009; Orsmond & Merry, 2011).

Ferguson’s work also identified the fragile nature of students’ self-confidence and the role feedback can play in shaping this construct. One student commented that, “If all comments were negative, I would never write a paper again” (2011, p.57). Another highlighted the importance of phraseology, and noted the different emotive responses possible from reading, “You could have tried this”, instead of “You did not do this” (Ferguson, 2011, p.57). Ninety-percent of students said a balance was required between negative and positive comments. The impact of feedback on self-confidence and self-efficacy has been explored elsewhere (Hattie & Timperley, 2007; Van Dinther, Dochy, & Segers, 2011; Nash, Crimmins & Oprescu, 2015) and is connected with a wider body of research that explores the relationship between emotions and learning (e.g., Weiss, 2000; Hattie & Timperley, 2007; Shields, 2015). Specifically, the internalisation of feedback is likely to be moderated by the student’s level of emotional maturity and whether they are in a receptive emotional state for feedback to be absorbed (Nash et al., 2015; Pitt & Norton, 2016). A lack of either of these things is likely to culminate in what Pitt and Norton (2016) have termed ‘emotional backwash’ whereby feedback messages are ‘… eclipsed by the learner’s reactions’ (Race, 1995, p.67) and guilty of causing academic paralysis (Nash et al., 2015). Interestingly Ferguson’s (2011) research was conducted with postgraduate students who will have already demonstrated a relatively high level of academic competence. Therefore it is perhaps surprising that their self-confidence remained fragile. This lends even more weight to Mandhane, Ansari, Shaikh and Deolekar’s (2015) contention that effective feedback needs to be performance rather than individual focused.

Essentially students have claimed that they want to know what to change and how to change it (Nicol & Macfarlane-Dick 2006). However, it has proved difficult to obtain a congruent picture of the types of feedback that tutors provide, and account for the inter-individual needs of students at the same time. For example, some students have claimed to thrive on negative feedback:

"Saying I didn't do so well makes me feel bad and spurs me onto wanting to get a better mark next time" (Pitt & Norton, 2015, p.6). Furthermore, although there is a movement towards integrating more opportunities for dialogic feedback, some students prefer only written comments (Yang & Carless, 2013) while other prefer meeting with tutors as well as written text (Blair & McGinty, 2012). Essentially students' preferences vary (Hepplestone & Chikawa, 2014)

Cognisant of the diversity of student preferences on feedback, Winstone et al. (2017) have applied a budgeting methodology to a recent study. Rather than allowing students the freedom to state that everything is important, this methodology allows researchers to understand the relative importance of different characteristics amongst a target group. Participants are provided with a budget and asked to 'buy' the qualities that are the most important. The budget then increases and more qualities are able to be bought. In this way researchers can see a process of prioritisation in the results and Winstone et al. (2017) hoped to be able to show which aspects of written feedback were most important for students. Study one showed that out of nine lecturer qualities available the ability of the lecturer to provide good feedback was ranked first by the students. Study two showed how students ranked the 10 characteristics of written feedback (from most to least important).

> 1) Highlights the skills I need to improve for future assignments
> 2) Suggests where I could get advice or help
> 3) Explains why the mark was appropriate with reference to grade descriptors
> 4) Includes comments that invited me to come and talk about the essay
> 5) Provides encouragement for things that were done well
> 6) Comments on the professionalism of my writing styles and/or how to improve it
> 7) Corrects grammatical errors/advises me how to improve my grammar
> 8) Shows how my mark compares to others in the cohort
> 9) Comments on my understanding of the topic
> 10) Highlights how well I have met the learning objectives
> (Winstone et al., 2017, p.1246)

Once again in relation to these findings it appears that what students want has not always been what they have received. Research has identified patterns of feedback that are predominantly corrective and error-focused (e.g., Orrell, 2006; Walker, 2009) as opposed to being improvement and future-oriented. However, this picture is further confused by recent research which demonstrated that even though students wanted future-oriented feedback they are less likely to remember this style of feedback over a past-oriented, evaluative style (Nash et al., 2018). Furthermore, there is a legitimate argument that students might not be best placed to make judgements about what they need (Huxham, 2007) and may only have a partial conceptualisation of feedback. This is something that has been picked up by Price et al.

> … we have a situation where students evaluate feedback with the benefit of first-hand experience of using feedback but without the pedagogic literacy to fully understand its role in learning processes (2010, p.286).

2.10.4 **Biased Feedback**

The potential for feedback on student work to be biased has largely been ignored in educational literature. Lip service has been paid to related concepts by a handful of researchers. For example, Chory-Assad (2002) and Lizzio and Wilson, (2008) have explored perceptions of fairness, specifically related to grades awarded and unambiguous marking criteria. Fairness as judged by the tone of comments has also been considered important (Lizzio & Wilson, 2008). Furthermore, Carless (2006; 2009) has raised the issue of trust, claiming that students will only act on information they deem to be trustworthy and warned that trust cannot be assumed. Moreover, Carless's (2006) research discovered that students had little faith in grading processes and considered these to be inconsistent.

Explicit links to bias have only been made fleetingly. For example, Poulos and Mahoney's (2009) qualitative data uncovered a theme related to the credibility of feedback. In part this referred to student perceptions that lecturer biases influenced the feedback provided. However, this bias was not considered to be directed towards the students themselves. Rather it highlighted how a lecturer's views on a particular subject matter might bias their marking of a student's work who did not agree with their ideology. Other work has failed to address bias but has written about the potentially related issue of mistrust. For example, in Price et al.'s (2010) examination of the importance of the relational dimension of feedback they maintained that the relationship between the tutor and the student is fundamental to the feedback process. They stated that mistrust of the feedback provider may culminate in a lack of motivation to engage with feedback, but failed to identify why or how this mistrust might manifest itself. Therefore it is possible that if the student mistrusts the tutor on the basis that they consider the tutor to be biased they will not develop a good relationship with them and are unlikely to learn from the feedback provided.

Moreover, as has previously been noted, the emotional state that a student is in dictates whether they absorb the information within feedback (Race, 1995; Nash et al., 2015; Shields, 2015; Pitt & Norton, 2016). Thus if student perceives their tutors' feedback to be biased this could evoke an emotional reaction which is sufficiently strong to be a barrier to their learning. This argument is supported by research findings from Orsmond et al., (2005) who found that how students responded to feedback was influenced by their perception of the tutor providing it.

Mutch's (2003) work also recognised that students from diverse backgrounds have different capacities to recognise and respond to feedback but did not consider that students from diverse backgrounds might also *receive* different feedback. Therefore discussion of concepts related to bias have been both scarce and tentative, and potential links to the concept have largely been missed within the educational literature. For many researchers this discussion might be outside the remit of their research question or objectives, and it is important to recognise that although the research on feedback has been diverse it does not have a long history. Nonetheless, recent research has decried how the cultural practices related to feedback in the higher education context have remaining isolated from ideas and research from outside the education sector (Boud & Malloy, 2013) and bias within feedback would seem to be one such issue.

2.10.5 **Biased Feedback and Gender**

There are however some notable exceptions to the examination of biased feedback in HEIs. In a body of research which examined essay writing and gender-related issues more broadly, it was Read et al.'s (2005) paper that most comprehensively explored the possibility of gender bias within written feedback in higher education. Their examination of in-text and summary feedback on undergraduate history essays explored differences in the perceived qualities of the essay as denoted from the feedback, and what shape or form this feedback took. Their findings revealed that feedback often differed markedly on the same essay and that tutors explicitly contradicted each other on some segments of the work. For example, while some tutors commented that an introduction was well-written, another considered it to be "appalling". This resonates with the views that feedback provision is far from consistent (Duncan, 2007; Lizzio & Wilson, 2008; Poulos & Mahoney, 2008; Li & De Luca, 2014) and implies that despite large numbers of HEIs using marking criteria to standardise the marking process many tutors continue to rely on more tacit forms of knowledge (Ecclestone, 2001; Price, 2005; Sadler, 2005) or connoisseurship (Knight & York, 2003) when marking student work.

Despite Read et al.'s (2005) work being the most relevant example of the examination of gender bias and feedback, much of their focus pertains to the gender of the tutor and not the student. Specifically, they have explored whether differences in the espoused feedback practices of male and female tutors correspond with the feedback on the assignment. They subsequently analysed these findings on the basis of differential feedback according to tutor gender. There was some examination of whether the reported differences correlated with different practice on the texts of female versus male students, but this was limited. However, that is not to belittle their findings altogether. Those of most interest here were that double the amount of female tutors compared to male tutors noted that lack of self-confidence might impact on female students' academic performance. However, this heightened awareness on the part of female tutors did not transfer to them providing increased amounts of positive feedback to female students. In fact, both male and female tutors provided four times more negative comments than positive comments in their feedback generally. This demonstrates that not all expectancies culminate in biased practice once category activation has occurred. Furthermore, Read et al. (2005) reported no sizeable differences in the feedback given according to tutor gender.

Despite explicit demonstrations of sexism declining in recent years it remains surprising that no examination of whether male and female students gain different feedback on the basis of their gender has been published in educational journals. Read et al.'s (2005) work might have been considered a potential catalyst for such research, but this has not come to fruition. Perhaps a more potent stimulant might have been the NUS report (2008). However, the development of such a body of research has been conspicuous by its absence both within the educational domain and beyond. There is one aberration however, and it falls within the realm of workplace feedback. Jampol (2014) has examined how gender bias is maintained through performance feedback. Her series of studies established that the feedback provided to female authors on their work was more compassionate but less accurate than that provided to males. This pattern was also more likely to occur amongst participants who held stereotypical views about women (e.g., women are emotionally weak). Jampol (2014) ascertained that the telling of what she termed 'white lies' to women contributed to the maintenance of the glass ceiling, since inaccurate feedback often prevents women from receiving the information necessary for performance improvement. Interestingly her work also underlined the implicit nature of gender bias since no participants reported any awareness of the existence of such a bias.

2.10.6 **Biased Feedback and Ethnicity**

Research on ethnic bias and written feedback within education is also embryonic. This is surprising given the diverse range of approaches taken by researchers to explain the under-attainment of ethnic minorities at all levels of education (Richardson, 2015). A body of dated research has explored the differential treatment of high expectancy versus low expectancy students in school-based settings (e.g., Lanzetta & Hannah, 1969; Rosenthal, 1973, 1974; Weinstein, 1976; Cooper, 1977, 1979; Brophy, 1983; Jussim, 1986) and determined that there is a feedback bias in favour of high-expectancy students. However, it would seem that no immediate connection was made between the ethnic groups that low versus high-expectancy students might inhabit. Consequently research did not develop in this domain within educational journals.

However, research exploring ethnic bias and feedback within educational contexts has been examined in the social psychology literature. As far back as 1998, Harber's research explored the contention that White students would provide more positive feedback to an essay perceived to have been written by a Black student than to a White student. This hypothesis was proved correct. Moreover, the feedback bias proved itself specific to certain types of feedback. Harber (1998) reported that feedback on essay content (which was more aligned to subjective evaluation, such as originality, argument, and choice of theory) was more dominant for White students than feedback on essay mechanics (which was aligned to more objective evaluation, such as grammar, spelling and referencing) which was more dominant for Black students. He explained this by noting that the provision of objective evaluation protects the marker from accusations of impartiality or bias since they have an external reference point from which they can draw support (i.e., dictionaries, presentation and referencing guidelines etc.).

In an extension of the examination of feedback types and the motives for their provision, Harber et al.'s (2010) later work found that Black students were given more favourable feedback because Whites were concerned about their self-image. In this research trainee teachers conducted the marking and had their egalitarian self-images manipulated through the completion of a version of the Social Issues Survey. The first version of the survey was designed to reinforce pro-minority views, the second to reinforce anti-minority views, and the third to reflect a neutral view. Given that research has demonstrated that the more uncomfortable Whites are with minority groups, the more positively they evaluate them (Littleford, Wright, & Sayoc-Parial, 2005) it was unsurprising that the markers whose anti-minority views were reinforced provided the most positive feedback to Black students.

Although Harber's (1998) work is certainly noteworthy, it is also important to acknowledge that similar to many earlier research papers related to gender bias it did not use authentic student essays or qualified marking tutors. Part of the protocol was that student markers were told their feedback was going to be returned to the author of the work. Therefore it is possible that this priming of self-presentational concerns may have had a greater impact on student markers (who were ultimately marking the work of their peers), than it would have had on more experienced tutor markers. This was a limitation that Harber (1998) acknowledged, noting that additional research exploring whether this bias extends when the feedback supplier and feedback recipient are of unequal social status is desirable. Another limitation relates to how clear the ethnicity of the author was to the student marker. Author ethnicity was only made apparent when markers were directed to read a demographic sheet supposedly completed by the author of the work. It is possible that the marker failed to attend to this information and thus the intended priming of the ethnic group may not have been activated. Furthermore, this sheet was handwritten, and given the research which demonstrates the impact of legibility bias (e.g., Greifeneder, Zelt, Seele, Bottenberg, & Alt, 2012) it is therefore difficult to disaggregate whether any bias emanated from the ethnicity of the writer, other information contained within the demographic sheet (e.g., indicators of social class, age etc.), or their handwriting. Moreover, the study did not include a control group, and consequently it is difficult to ascertain whether differences in the feedback were as a result of the knowledge of the ethnicity of the writer or simply reflected different feedback styles of the marker. Similar criticisms have been identified earlier with regard to research exploring marker severity (e.g., Malouff et al. 2013).

Harber's (1998) findings, though novel in their application to education, reflected a growing trend in social psychology research exploring intergroup evaluation. This body of research has demonstrated that Whites consistently provide positively biased assessments of minority groups (e.g. for a review see Devine, 1989). A number of explanations have been forwarded for this trend. These include, self-presentational motives; whereby Whites do not want to appear prejudiced to others or themselves (Vorauer & Kumhyr, 2001; Littleford, et al., 2005), and sympathy motives; whereby dominant groups feel uncomfortable criticising less dominant groups (Jones, Farina, Hastorf, Markus, Miller, & Scott, 1985). There are also a host of more implicit expectancy-based processes which might explain this research trend. These include Whites having lower expectations of work produced by minority groups and therefore feedback registers their surprise when these are exceeded (Biernat & Manis, 1994); and Whites judging other Whites more severely for poor work (Jussim et al., 1987) subsequently providing more critical feedback to them in comparison to other ethnic groups.

Ostensibly it might be considered beneficial to be provided with more positive and favourable feedback, and in some cases also higher grades (e.g., Croft & Schmader, 2012). However, Harber (1998) takes a similar view to that expressed earlier by Jampol (2014) which is that it is damaging for academic development and progression. For example, it may prevent Blacks from developing higher achievement motivation, and/or fail to maximally challenge intellectual capacity, therefore thwarting development. This is a concern echoed by Croft and Schmader who queried, "How can we learn from our mistakes if we're unaware they exist?" (2012, p.1139). In their Canadian-based research they coined the term feedback withholding bias to explain how White markers gave equal amounts of positive feedback to other Whites and minority students, but withheld provision of negative feedback to minorities. While outwardly their results would seem to support other research on feedback bias there are a number of shortcomings with their methodology which should be examined. For example, once again student markers and non-authentic essays were used. Additionally, although three so-called 'filler essays" were used alongside two target essays it was unclear whether these were analysed in an attempt to account for marker variability. The paper simply stated that the inclusion of the filler essays allowed participants to view assignments which varied in academic quality and thus presented a 'meaningful range' (p.1140). Moreover, there were limitations with the way in which feedback was measured and subsequently interpreted. Negative and positive feedback were simply measured according to the total number of centimetres of text that had been highlighted in blue (negative) or yellow (positive). This reflects a somewhat reductionist approach to measuring what is essentially an emotive type of information. In this way, the length of the comment is considered to be more meaningful than the message contained within it. Consequently a comment that said 'impeccable work' on a Black student's essay would be considered as providing less positive feedback than a comment which stated 'your work is of a very high quality' on a White essay. Arguably the meaning is lost through the quantitative measuring process and is also misrepresentative. Finally, Croft and Schmader (2012) administered their motivation related questionnaires after participants had marked the assignments and therefore their responses might have been influenced by their perceptions of the assignments.

The ramifications of the type of feedback provided to Whites and ethnic minorities is incredibly complex however. While it is clear that the feedback withholding bias (Croft & Schmader, 2012) may prevent ethnic minority students from progressing in the same way as White students, there is also concern that threatening feedback may be especially damaging to minority students who are already mistrustful of educational establishments and prone to disengagement (Cohen, Steele, & Ross, 1999). Cohen et al. (1999) identified a raft of self-protection strategies that ethnic minority students might engage in to buffer the impact of negative feedback. One such example related to how minority students attribute the feedback, since if they perceive the negative feedback to stem from racist bias as opposed to a real need to improve their work they may fail to act on it and thus miss an opportunity to improve. Whilst Cohen et al. (1999) do identify some strategies previously used with minority youths to address this dilemma, it seems that maintaining a balance between keeping minority students motivated and engaged while also providing non-biased feedback continues to be a challenge.

Richardson et al.’s (2014) work was the first article published in an educational journal and conducted within HEIs in the UK which examined different feedback practices amongst ethnic minority students. They highlighted that ethnic differences in academic attainment have been acknowledged for many years and that research has consistently highlighted their ubiquitous nature (e.g., Owen, Green, Pitcher, & Maguire 2000; Broecke & Nicholls, 2007; Richardson, 2008; Richardson 2015). Richardson (2015) explains that White students are almost twice as likely to gain a good degree (first class or 2:1) than a non-White student. Asian students are more likely to get a better degree than Black students, and Chinese students are more likely to outperform Asians. He continues that only about half of the attainment gap can be explained by differences in academic ability. Various explanations have been forwarded to explain the other fifty-percent of this gap. These have included claims that ethnic minority students have unsatisfactory experiences within HEIs (Singh, 2011). However, other large-scale research has rebutted these claims (e.g., Connor et al., 2004) and NSS data has identified few differences in this regard.

Richardson et al. (2014) therefore contended that perhaps one area which might uncover some differences across ethnicities and help to unravel the attainment gap might be the feedback awarded. They subsequently analysed both in-text and summary feedback using a computer programme called OpenMentor. Findings revealed that there were no differences in the number of comments provided across ethnic groups. There were differences in the type of comments provided however, with Black students tending to receive more negative comments. Thus Richardson et al.’s (2014) findings contradicted much of the research conducted in the social psychology domain regarding inter-group evaluation and that specifically conducted within education settings on feedback (i.e., Harber, 1998; Harber et al., 2010; Croft & Schmader, 2012). Although Richardson et al. (2014) claimed the receipt of more negative comments for Black students was a matter of concern, they also noted that since these differences would only equate to a small effect size they were not considered important. Moreover, when marks were taken into account alongside the feedback the small differences between ethnic groups disappeared, therefore indicating that students were given feedback appropriate to the mark awarded. Richardson et al. (2014) therefore concluded that explanations for the attainment gap could not be attributed to differences in feedback.

The work of Richardson and colleagues is both worthwhile and valuable in examining whether feedback has a role to play in the under attainment of ethnic minority students. It is clear that Richardson et al.’s (2014; 2015) work has emerged from a desire to explore and explain the attainment gap that exists between White and ethnic minority students as opposed to an interest in exploring whether or not academics are biased in the feedback they provide. However, this means that although his work examined whether differences existed it fell short in its consideration of why such differences might exist. As such the work can be criticised for its failure to adopt a theoretical lens through which to examine and explain its findings. The adoption of an atheoretical approach to research within the higher education domain is not uncommon, and calls have been made for the field to increase its theoretical engagement in order to gain credibility (see Tight, 2004; 2014).

The examination of the potential for a feedback bias to exist across genders and ethnicities has been woefully slow to develop within the educational domain. This is despite the domain of social psychology having conducted numerous studies on inter-group evaluation and a few specific studies on feedback to ethnic minority groups. The reticence of educationalists to leave the comfort of their own sphere and engage with ideas and research from elsewhere has created an isolation that is dangerous. Uncritical acceptance of claims regarding biased marking practices (such as those promoted by the NUS) means educationalists forgo the opportunity to underpin their practice with policies which are both theory and evidence driven. Further interrogation of the NUS report (2008) would have identified a number of shortcomings. For example, it missed a raft of research within the social psychology domain that contradicted their claim regarding non-white and female students gaining lower grades. Admittedly finding such research is difficult when “…a significant part of the more interesting research into higher education is not published in specialist higher education journals… but in those devoted to particular social science disciplines” (Tight, 2004, p.409). However, it is incumbent on those seeking answers to such difficult questions to be meticulous in their search for such literature. Additionally, since the social psychological literature has often demonstrated the existence of a positive bias this raises new questions which paint a different picture of the marking process and are therefore likely to require different interventions to rectify it.

3 RESEARCH METHODS

3.1 Introduction

The NUS (1999, 2008, 2012), and prominent voices across the sector have repeatedly made claims that marking is biased on the basis of student gender and ethnicity. They have used this argument to underpin a drive towards anonymous assessment. The aim of this thesis was to explore this claim of bias through examining whether expectancy effects, as primed by knowledge of the student name, would lead to biased feedback.

This study used mixed methods research to address the following research questions. Do expectancy effects as primed through knowledge of the following characteristics impact upon feedback in a way that suggest biased practice:

i) Student gender
ii) Student ethnicity
iii) Student gender and ethnicity

A description of mixed methods research and the underpinning philosophical approaches for this thesis will now be presented.

3.2 Introduction to Mixed Methods

Although it is possible to find discrete examples of mixed methods research (MMR) as far back as the late 1970s (e.g. Denzin, 1978) it was only in the late 1980's that this type of research gained momentum. As the approach began to develop sufficient popularity and credibility it was referred to as the "…third methodological movement" (Tashakkori & Teddlie, 2003, p.5), "…a new star in the social science sky" (Mayring, 2007, p.1), and, "…a research paradigm whose time has come" (Johnson & Onwuegbuzie, 2004, p.14).

Several definitions of MMR exist and reflect how it has evolved in its short history. However, Johnson, Onwuegbuzie and Turner's attempt to define it is most comprehensive since it involved input from many mixed methods researchers. Their definition demonstrated that MMR transcends merely being about methods and encompasses an entire methodology.

> Mixed methods is the type of research in which a researcher or team or researchers combine elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the purposes of breadth and depth of understanding and corroboration (2007, p.123).

Promoted as a paradigm which allows researchers to approach problems intuitively (Creswell & Plano Clark, 2011) and use "… multiple ways of making sense of the social world" (Greene, 2007, p.20), mixed methods has promoted itself as an accessible approach to inquiry. Often research questions, research outcomes, and consequences are promoted as more important than the methods themselves. At times therefore MMR appears to have intentionally distanced itself from the more established research paradigms and their philosophical underpinnings. Nevertheless, some have emphasised that its goal is not to replace either qualitative or quantitative approaches to research. Instead it "… bridges the schism", positioning itself at the mid-point of the research methods continuum, with qualitative and quantitative approaches residing on opposing sides (Johnson & Onwuegbuzie, 2004, p.15).

Perhaps cognisant that MMR might be criticised as an 'anything goes' approach, Creswell and Plano Clark (2011, p.5) have judiciously identified some core characteristics of MMR. These are as follows:

1) MMR collects and analyses persuasively and rigorously both qualitative and quantitative data (based on research questions).
2) MMR mixes (or integrates or links) the two forms of data concurrently by combining them (or merging them), sequentially by having one build on the other, or embedding one within the other.
3) MMR gives priority to one or both forms of data (in terms of what the research emphasizes).
4) MMR uses these procedures in a single study or in multiple phases of a program of study.
5) MMR frame these procedures within philosophical worldviews (paradigms) and theoretical lenses.
6) MMR combine the procedures into specific research designs that direct the plan for conducting the study.

However, despite having offered these core characteristics, they also acknowledged the complexity and variety inherent in the conduct of MMR and noted that a "… limitless number" (2011, p.68) of designs exist. This diversity may be reflective of mixed methods short history and the subsequent lack of guidance researchers have had to operate within. Indeed, Johnson and Onwuegbuzie (2004) shrewdly noted that it is time that methodologists caught up with the types of research being conducted by practising researchers.

MMR has been labelled as the paradigm of choice in educational research (Johnson & Onwuegbuzie, 2004) where questions have been raised about the appropriateness of traditional experimental designs (Kember, 2003). Its strengths are considered to include, a) the use of both qualitative and quantitative research procedures which nullify the weaknesses of using a single approach, b) not being restricted to using specific data collection tools, c) being able to answer questions that cannot be answered by using a singular approach, d) encouraging the development of new research paradigms (e.g. pragmatism), e) being applicable to the real world, and, f) being able to provide stronger evidence due to the merging and verification of findings (Johnson & Onwuegbuzie, 2004; Creswell & Plano Clark, 2011). The limitations of MMR are considered to be, a) its short history, which may mean people need convincing of its value (much as was once the case with qualitative research), b) the blending of philosophical positions and paradigms which purists may argue against, c) the need for researchers using MMR to be competent in both qualitative and quantitative research, d) the time consuming nature of such research, and, e) the need for researchers to learn about a new paradigm and understand how to mix methods appropriately (Creswell & Plano Clark, 2011).

Further to these identified limitations, MMR has been subject to additional criticism. Perhaps the most vocal critic is Giddings who claimed that MMR is simply "… positivism dressed in drag" (2006, p.195). She contended that the drive for and expansion of MMR reflects economic concerns (i.e. to gain research funding), rather than an altruistic desire to develop the field of research methods. She further argued MMR has paradoxically regressed the field of research methods since it has retained the terms qualitative and quantitative at a time when qualitative researchers no longer have to define themselves in opposition to their quantitative counterpart. She maintained that the fusion of qualitative and quantitative approaches under the umbrella of MMR hid methodological diversity and subsequently increased the potential for positivist ways of thinking about research to thrive undetected. She further noted that many studies which claim to use MMR are rarely constructionist or subjective in their worldview, but instead are heavily cloaked in postpositivist philosophy, design, and analysis, with the qualitative aspects being tokenistic. Giddings (2006) is not alone in her criticism of MMR or of the judgement surrounding the secondary status qualitative research is afforded within it (Denzin & Lincoln, 2005).

Such vehement criticism has attracted responses, most notably from Creswell, Shope, Plano Clark, and Green (2006). Their paper cited a number of qualitative researchers who advocated the use of MMR and also identified some key MMR research projects that privileged qualitative approaches within their research. They outlined that while some purist qualitative researchers may feel threatened by the rise in MMR, both its acceptance among this community and the use of interpretive, critical theoretical frameworks within mixed methods studies are on the increase.

3.3 Methodological Philosophy

Research related to higher education has increased significantly since the turn of the 21st century. Ostensibly this has been identified as being a combined result of the massification of higher education, the changing perception of HEIs as businesses, and the subsequent interest from key stakeholders with the quality of the processes (Brendan & Shah, 2000). This flurry of research interest has used a variety of philosophical and methodological approaches (Tight, 2013; 2014). According to Tight (2004) this reflects both a tendency for researchers to work within familiar approaches from their own disciplines and adopt an institutionally-centric approach. While on the one hand, this diversity has created a wide-ranging and innovative field of research (Altbach, 1997), it has also prevented a more cohesive, synergistic, research community from developing which has led to the field being described as fragmented.

Furthermore, research into higher education has been criticised for being largely atheoretical. Tight's (2004) exploration of over 406 articles published in the year 2000 across 17 higher education based journals outside of North America (where the field is better established), demonstrated that at best theory was only implicit in the majority of the research. He concluded that, "Higher education researchers, for the most part, do not appear to feel the need to make their theoretical perspectives explicit, or to engage in a broader sense in theoretical debate" (Tight, 2004, p.409). However, although his follow-up study (which included analysis of 567 articles published in 15 leading higher education journals) still referred to theoretical application and development as being "… fairly low level" (Tight, 2014, p.107), this was rationalised by highlighting that the focus of educational research has been more pragmatically oriented. However such pragmatism need not always be implemented at the expense of theory. Indeed MMR does advocate the use of an explanatory framework or theory which should underpin the project (Creswell & Plano Clark, 2011).

3.3.1 Philosophical Background

Creswell and Plano Clark (2011) use the term worldview instead of paradigm and identified that there are four worldviews and accompanying philosophical characteristics that underpin research.

Postpositivist Worldview	**Constructivist Worldview**	**Participatory Worldview**	**Pragmatist Worldview**
Determinism	Understanding	Political	Consequences of actions
Reductionism	Multiple participant meanings	Empowerment and issue oriented	Problem centred
Empirical observation and measurement	Social and historical construction	Collaborative	Pluralistic
Theory verification	Theory generation	Change oriented	Real-world practice oriented

Table 1: Basic characteristics of four worldviews used in research (Creswell & Plano Clark, 2011, p.40)

The postpositivist and constructivist worldviews are familiar to most researchers since they align themselves to the quantitative and qualitative research paradigms respectively. While postpositivism adopts a slightly softened version of positivistic ontological and epistemological assumptions it is still largely characterised by belief in a singular truth, the notion of objectivity, experimental designs, deductive reasoning, and quantitative data collection (Willig, 2008). Alternatively constructivists believe in multiple truths which are socially constructed and inductively revealed through communication from their participants (Creswell & Plano Clark, 2011). The participatory worldview is perhaps less well-known and is more often aligned with qualitative research (although this is not always the case). It is underpinned by political concerns such as societal improvement, specifically though not solely through addressing issues related to the marginalisation and unfair treatment of specific groups. Often people conducting participatory research work closely with groups experiencing discrimination. The pragmatist worldview is the belief system most closely aligned to MMR. Rather than being overly caught up in the methods of research this worldview is more concerned with the research questions asked and the consequences of the research itself. Armitage and Campus (2007) see it as the most appropriate paradigm for real life research since it encourages multiple methods of data collection and is concerned with what works and practice. Such a view would be endorsed by Garratt and Hodkinson who claimed that,

> … the selection of appropriate preordained sets of different paradigmatic rules, then, is not a solution. A more constructive way forward begins with the acknowledgment that the selection of criteria should be related to the nature of the particular piece of research that is being evaluated’ (1999, p.527).

Many mixed methods researchers now consider the pragmatist approach to provide the best foundations for MMR (e.g. Tashakkori & Teddlie, 2003).

This thesis most closely parallels the pragmatist worldview for the following reasons: it is interested in exploring the consequences of knowing a student’s name on feedback provided (consequences of actions). It is centred on the problem of accusations of bias in the marking practices of HEIs (problem centred), and was undertaken by analysing different types of feedback (pluralistic). It addressed an on-going and contentious issue within HEIs through an examination of marking practices (real-world practice oriented). If the methodology for this thesis is considered in relation to the key elements that differentiate paradigms at each stage of the research process (Lincoln & Guba, 2000; Creswell & Plano Clark, 2011), then ontologically the research explores both singular and multiple realities (i.e., between groups differences and participants individual comments). Epistemologically data collection has been guided by adhering as far as possible to usual HEI marking practices and thus it is pluralistic, practical, and adheres to what works in practice. With regard to axiology (e.g., value-free or value-laden, biased or non-biased) it is recognised that the researcher is not value-free and therefore brings both subjective, and thus potentially biased (e.g., personal history, gender, HEI experiences as a student, HEI experiences as a lecturer) and objective, and thus potentially unbiased (e.g., aspects of the methodological design) perspectives to the project. These potential biases were addressed by adopting an element of reflexivity into the early stages of analysis through the use of a critical friend. Methodologically the project includes aspects of deductive and inductive reasoning and analysis.

Nonetheless, according to mixed methods researchers, worldviews only provide a *general* philosophical orientation and can therefore be combined. Indeed, Crotty emphasized that they are not "… watertight compartments" (1998, p.9). This flexibility bodes well for this project since it includes philosophical aspects of each of the four worldviews. For example, it involves some empirical observation and measurement (post-positivist worldview). It aims to enhance understanding through the representation of multiple participants' meanings (constructivist worldview). It is partly political, and is empowerment and change oriented (participatory worldview), since it hopes to influence HEI policy in order to reduce the discrimination against specific social groups.

Postpositivist Worldview	**Constructivist Worldview**	**Participatory Worldview**	**Pragmatist Worldview**
Determinism	Understanding	Political	Consequences of actions
Reductionism	Multiple participant meanings	Empowerment and issue oriented	Problem-centered
Empirical observation and measurement	Social and historical construction	Collaborative	Pluralistic
Theory verification	Theory generation	Change oriented	Real-world practice oriented

Table 2: Identification of the four worldviews used in this research (highlighted in yellow)

3.4 Mixed Methods Research Designs

Creswell and Plano Clark (2011) have identified a typology of six MMR designs which help researchers to select and adapt a particular design to their study's purpose and questions. These typologies include: the convergent design, the explanatory design, the exploratory design, the embedded (or nested) design, the transformative design and the multiphase design. Despite what might be interpreted as a rather rigid approach to identifying MMR typologies, these designs have been generated from a comprehensive exploration of types of MMR conducted since the 1990's. Furthermore, Creswell and Plano Clark do acknowledge that although these approaches are the most well used there are likely to be a, "… limitless number of unique combinations" (2011, p.68).

This thesis most closely resembles the convergent design. The convergent design is underpinned by a pragmatist philosophy and includes two key purposes which fit with this thesis, 1) to use different types of data to obtain multiple ways of understanding the research problem (i.e. in-text feedback content analysed deductively and numerically. Summary feedback content analysed inductively), 2) to synthesize qualitative and quantitative results in order to develop a holistic understanding of the issue (i.e., to integrate the descriptive statistics obtained from the in-text feedback and the themes that emerged from the summary feedback). Broadly the convergent design entails the researcher collecting both qualitative and quantitative data at the same time in the research process. Data are only merged during the '… point of interface' (Morse & Neihaus, 2009, p.56) which occurs at the final stage of the writing process. This thesis did not collect quantitative data per se, since the in-text feedback was textual rather than numerical. However, the feedback was analysed deductively using an existing feedback classification system (Hyatt, 2005), and then presented in the form of descriptive statistics, trends and frequency counts. Moreover, an unequal emphasis across the qualitative and quantitative strands is not unusual, although some purist definitions of MMR do highlight the importance of having a fully developed qualitative and quantitative strand at each phase of the design. Nonetheless, others have highlighted that mixed methods "… combine *elements* of quantitative and qualitative research" (Johnson et al., 2007, p123). In terms of this thesis it was considered more important to stay congruent with the pragmatist philosophy of real world practice than it was to impose a set of methods which threatened this in order to balance the qualitative and quantitative weighting of the research.

3.5 Research Design

This section of the thesis will explain participant recruitment, participant demographics, the development of the research pack, procedures, and processes of data analysis.

3.5.1 Participant Recruitment

After ethical approval had been granted from the University of Winchester, Heads of Sport Departments at Universities across England, Scotland, and Wales were emailed. These included Russell Group Universities, pre-1992, and post-1992 institutions. Monthly follow-up emails and assessment packs were sent to those Heads of Department whose staff agreed to participate. In order to encourage participation and compensate participants for their time each individual was paid thirty pounds. However, even with this incentive it took over two years to recruit a sufficient number of participants.

3.5.2 Participants

Sixty sport and exercise academics (n males = 30, n females = 30; mean age 34.82 years, S.D. = 11.09 years) were recruited from eight HEIs across England, Scotland, and Wales. Twenty-three percent of participants (n = 14) were drawn from pre-1992 institutions and seventy-seven percent (n = 46) were from post-1992 institutions. Ninety-five percent of participants (n = 57) defined their ethnicity as White British with the remaining five percent identifying themselves as Asian/Asian British – Indian (n = 1), Other White Background (n = 1), and Mixed – White and Black African (n = 1).

The participants had a range of teaching experience and marking loads across a standard academic year as illustrated below.

Figure 3: Participants Teaching Experience

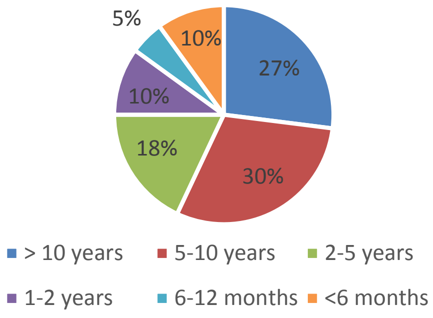
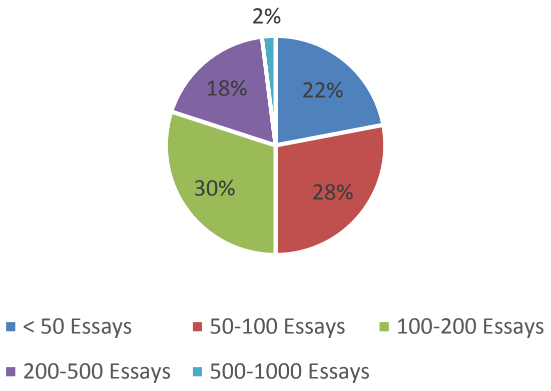


Figure 4: Annual Marking Loads



Participants represented a range of five academic positions. Professor, (*n*=1), Principal Lecturer, (*n*=5); Senior Lecturer, (*n* = 19); Lecturer, (*n* = 16); Associate Lecturer, (*n* = 19). The experimental protocol was explained to the participants and ethical approval and written informed consent obtained. Participants were also debriefed as to the purpose of the research via email once they had completed the marking process.

3.5.3 **Development of the Assessment pack**

Each assessment pack (see appendix iv) included a demographic questionnaire, and two essays (control and experimental). The essays each included an assessment criteria sheet. A question was printed at the end of each assignment which asked which factors influenced the participant's perceptions of the work. The essays had previously been submitted by students at the University of Winchester for a first year undergraduate Research Methods module and these students gave their consent for their work to be used for this research.

On the cover sheet for each essay the student's name was presented in size thirty-six font and in bold capital letters. The name was obviously important for participants to notice since the intention was that this static cue would act as a stimulus for category-based stereotypical thinking and expectancy effects to occur. It is usual practice in expectancy-based research to prime participants through both written and oral instructions (e.g., Jones, Paull & Erskine, 2002), and it has been demonstrated that even 'thin slices' of information (such as a name) can influence social judgement and interactions (Ambady & Rosenthal, 1992). Below the student name were instructions for the marking process. These instructions asked participants to provide feedback 'in line with current teaching practice', and 'as if it was to be returned to the student'.

Given that any cognitive stimulus remains accessible in the mind for as long as twenty-four hours after the event (Srull & Wyer, 1980) this increased accessibility to the student's name was considered to remain active throughout the marking process and thus have the potential to be expectancy forming. Interestingly however the impact of priming information does not have to be consciously noted by the perceiver in order for it to be impactful (Bargh & Pietromonaco, 1982). Consequently, although at the end of the marking process participants were explicitly asked to reflect upon what had influenced their perceptions of the essay it is possible that, a) the name and ethnicity of the student could have impacted upon participants' expectancies without their conscious awareness, b) the priming process failed to activate a stereotypic expectancy effect because participants inferred the true nature of the experiment and consciously impression-managed their attitudes. This supports Fazio's (1987) belief that if individuals are aware of what the research is exploring their expectancies are largely reactive to measurement. The decision not to inform participants about the true nature of the research until after data collection was completed at the debriefing stage is therefore validated. For those who believe expectancies operate unconsciously the inclusion of such a question may seem redundant, since participants would not be aware of what had influenced their perceptions, but it was considered important to include because some researchers do adhere to the view that expectancies are conscious. Furthermore, if the relationship (between the participant and the student) was not perceived as important to the participant they may not allocate sufficient time and resources to effectively prevent the occurrence of expectancy effects or engage in self-presentational strategies to mask them (Neuberg & Fiske, 1987; Jussim, 1993).

3.5.4 **Procedures**

In order to examine between group differences in feedback provision, participants were randomly and equally divided (n=10) into 6 groups of markers. Each group were assigned a condition which referred to the gender and ethnicity of the student whose work they would mark in the experimental condition; White British male (WBM), White British female (WBF), Asian Male (AM), Asian Female (AF), Chinese Male (CM) and Chinese Female (CF). Each participant marked a control essay authored by Samuel Jones, a male, White British name, and an experimental essay authored by either a WBM (James Smith), WBF (Natasha Brown), AM (Jagjit Sidhu), AF (Avinash Puri), CM (Zhi Rong Liu), CF (Mei Lin Pang). The content of the experimental essays were the same, it was only the student name that was changed.

The experimental design attempted to improve internal validity by controlling for the confounding variable of marking stringency through the existence of the control essay. As such if no feedback differences were seen between conditions when participants marked the control essay, but differences were noted when participants marked the experimental essay this provide confidence that these differences emerged as a result of expectancy effects related to the gender or ethnicity of the student and not simply differences in feedback practice between groups.

Although participants were instructed to provide feedback ‘in line with current teaching practice’, and ‘as if it was to be returned to the student’ this did not produce identical practice. Fifty-seven out of sixty participants (95%), provided in-text feedback for both the control and experimental essays. However the provision of summary feedback was more variable. While twenty-eight participants (47%) provided summary feedback on both essays, twenty-four participants (40%) provided no summary feedback on either essay. Eight participants provided summary feedback on one essay but not the other. Summary feedback was missed on 3 control essays (5%) which had a male White British name assigned to them and on 5 experimental essays (8%) all of which were written by female students. Specifically one was a WBF name, one was AF and three were CF. Similar inconsistencies were noted in Read et al.'s (2005) research and perhaps highlights both the disparate marking practices which exist across HEIs in the UK alongside intra individual variations.

However, upon closer inspection the above figures were slightly misleading. At the end of each essay participants were explicitly asked the following question; 'what factors influenced your perception of the essay?' It became evident that two participants responses to this question were not intended for the researcher but instead had been written in a style designed to be returned to the students. These responses have therefore been coded as summary feedback. The coding criteria for this process involved applying syntactic rules regarding sentence structure. For example, when the lecturer had addressed the student by name, i.e., 'Sam, while you included reference to arguments on two sides of the issue', or had used the personal pronoun 'you', i.e., 'You draw on a good range of literature appropriately and demonstrate understanding of their agreements', these comments were deemed to represent summary comments. Once this criteria had been applied 30 participants (50%) had provided what constituted summary feedback on both assignments as opposed to the original 47%.

In terms of the final question asking 'what factors influenced your perception of the essay', fifty-eight participants (97%) in the control group responded and fifty-four participants (90%) in the experimental group. Two participants failed to answer the question for either of the essays they marked. Four participants only answered the question for the control essay presented as being authored by a WBM and did not answer it when they marked an AF, a CF, an AM and a CM name. Whether this was simply because the participants perceived the question to be excessive when asked a second time is impossible to ascertain, although it is interesting that the question was overlooked when marking essays authored by non-White British students.

3.5.5 **Process of Data Analysis**

Tight has identified documentary analysis as one of a handful of methods that constitutes the "… 'bedrock' of educational and social science research" (2014, p.100). Although the analysis of feedback on assignments may be considered an unorthodox form of textual data; since it analyses responses to a text rather than the key body of the text itself, it does act as a record of the thoughts of the marker and as such is important to interpret and understand.

Irrespective of the type of text being examined, content and thematic analysis are the typical approaches used within the umbrella of textual analysis. Identified as being "… useful for examining trends and patterns in documents" (Stemmler, 2001, p.2), their purpose is to describe the content, structure, and function of text-based messages and, at times, interpret meaning from the data (Frey, Botan & Kreps, 1999). Such analytic methods have been considered to lack the kudos of their more sophisticated cousins (e.g., discourse analysis) and have also been considered synonymous with each other. This has led to both techniques being poorly understood and demarcated (Braun & Clarke, 2006). Some have claimed that content analysis is more closely aligned to the quantitative domain due to its deductive analytic approach, categorisation of textual data, and reporting of descriptive statistics and frequency counts (Joffe & Yardley, 2004). Nonetheless, while it is true that content analysis is more concerned with establishing categories of information (description) as opposed to patterns of experience (meaning), many still consider it to have a place within qualitative methods (e.g. Sparkes & Smith, 2014). Moreover, there are specific types of content analysis (e.g., hierarchical content analysis) that explore categories at a latent level, focus on the process of indwelling (Maykut & Morehouse, 1994), and adopt inductive analytic procedures. This demonstrates that some forms of content analysis can align themselves more closely to the qualitative aspects of practice usually associated with thematic analysis.

Given this lack of clarity, it is therefore important for researchers using either of these forms of analysis to define their terms and clearly explain their procedures so that trustworthiness and credibility can ensue (Biddle, Markland, Gilbourne, Chatzisarantis, & Sparkes, 2001; Cote, Salmela, Baria, & Russell, 1993). Throughout this thesis the term content analysis will be used to refer to the procedures used to analyse the in-text feedback data. This analysis will be deductive in nature since it uses a predetermined set of categories to organise the data (i.e. Hyatt, 2005) and seeks only to describe the data at a manifest level. Although using thematic analysis was considered for the summary feedback, one defining feature of this method is that it seeks patterns of experience as opposed to categories of meaning. Given that participants' summary feedback and reflections did not represent them recalling a personal experience but rather detailed the information they wanted relayed to students, it was considered that hierarchical content analysis would be more appropriate in this instance. As such the term hierarchical content analysis will be used to refer to the procedures used to analyse the summary feedback. The procedures followed for each type of content analysis are explained in the following sections.

This thesis therefore explored two ways participants constructed their feedback comments in order to gain a holistic understanding of the feedback landscape. It aimed to identify, analyse, and interpret the feedback provided as well as searching for expectancy-induced similarities and differences according to gender, ethnicity, and the interactive effects of both gender and ethnicity.

3.5.6 **Analysis of In-Text Feedback**

Criticism has been levelled at research which has claimed to adopt content analysis procedures but failed to provide detailed explanation of how analysis was conducted (Biddle et al., 2001). Conversely step-by-step explanations of data analysis procedures have gained researchers' kudos through having their work considered trustworthy, credible (Cote et al., 1993), and transferable (Hardy, Jones, & Gould 1996). Therefore what follows is a detailed explanation of procedures followed for in-text feedback.

Assignment in-text feedback (n=120) was text-to-text transcribed. This process not only included textual data, but any mark made on the text including punctuation, symbols, circled text, inserted arrows etc. The process involved several stages. The first thirty essays underwent preliminary emergent analysis by two people, thus introducing a level of investigator triangulation into the procedure. Defined as the "… use of multiple observers/investigators in a single study" (Denzin, 1970, p.223), this process has been identified as adding credibility to research findings and as a suitable, though as yet largely unexplored, technique within MMR (Archibald, 2016). The practice has also been credited with enhancing procedural reliability since it initiates discussions between researchers which can reduce ambiguity (Weber, 1990).

Furthermore, the use of a critical friend added an element of reflexivity to the data analysis. Reflexivity asks researchers to "… explore the ways in which a researcher's involvement with a particular study influences, acts upon and informs such research (Nightingale & Cromby, 1999, p.228). There are considered to be two types of reflexivity. Personal reflexivity considers how the researcher's background, beliefs, morals, values, gender, ethnicity etc. might shape the research, whereas epistemological reflexivity considers how the research design and analysis might reflect assumptions about the area to be researched. While the notion of reflexivity has received scant explicit attention within MMR, in qualitative research it is considered central to understanding how the researcher influences and shapes the research process. However, the notion of personal reflexivity moves beyond simply reflecting on our individual biases to a consideration of how these impact on our understanding of the data and the interpretations we draw from it (Willig, 2008). Awareness of these issues prompted the inclusion of a critical friend in the early stages of data analysis for the in-text feedback. The role of a critical friend is to "… provide a theoretical sounding board to encourage reflection upon, and exploration of, alternative explanations and interpretations of … the analysis of the data as it is generated (Sparkes & Smith, 2014, p.183). Additionally, borrowing from the domain of phenomenological methods the researcher aimed to suspend or bracket quick, taken-for-granted judgements and assumptions in an attempt to engage in the contemplative phase of epoche. Amongst other things, this process involved reflecting upon my own assumptions regarding the time and effort that first year students had invested in writing their essays and the biases I hold about poor quality written English.

Preliminary coding of the data used a table that consisted of three columns (see Table 3). The first column documented the paragraph and line/sentence of the assignment that had been edited or commented upon by the marker; the second column detailed the initial coding category for the feedback (e.g. Request for rephrasing) and the final column provided details on the action the marker had taken, for example the exact comment written, or the way in which they had edited the assignment text for example, "[underlined sentence and wrote "is this sentence necessary?"]". The final column also replicated the section of assignment text written by the student in order to provide some context for the feedback comment made by the marker. Markers' feedback has been presented in red font in this third column in order that it can be distinguished from the original student text. If a marker had underlined the original text, this was reproduced with a red line. If they had identified a section of text and commented on it, this identification was represented with yellow highlighting. If a marker had circled the original assignment text, this was indicated within the transcript by highlighting the text in grey. If they had crossed through any text this was represented by a double strike through.

Para: line	**Type of comment**	**[action] and 'extracted data'**
1:1	Request for rephrasing	**[underlined sentence and wrote** "is this sentence necessary?"**]** "There are numerous arguments both for and against the legalisation of drugs in sport."
2:6	Punctuation correction	**[entered full stop, replaced word, entered comma and crossed through comma]** "….damaged by drugs (Dimeo, 2007). ~~what's more~~ Moreover, Gifford, (2004)…."

Table 3: Example of initial transcription process

Variations in transcription and coding practices were initially reviewed after 10 essays had been marked. Any disagreements regarding coding categories were discussed with a critical friend and agreement on labels were reached. Another 20 assignments were then coded and we met for the final time to ensure consistency of application of codes across all in-text feedback. Where there was disagreement a researcher would query this on the electronic version of the document and suggest an alternative (see highlighted section in Table 4). These alternative suggestions were then discussed and agreed upon.

Para: line	Type of comment	[action] and 'extracted data'
1:5-7	Informal language correction Previously you have said 'informal language error' The marker hasn't actually corrected the work so, for consistency, I think we should keep it as 'informal language error'. What do you think?	**[highlighted sentences and wrote** "A little colloquial – try to keep phrasing scientific & formal"**]** "Even though many drugs within sport are illegal, athletes still take them, but for what, the fame, the medals, the sponsorship? Still the underlying remains, should drugs in sport be legalised?"

Table 4: Example of researcher triangulation regarding coding

Once the feedback from thirty essays had been transcribed a table was composed in order to document the thematic structure that was emerging (see appendix v). The table includes themes that the types of comments could be categorised into. Three higher order themes were identified; critical feedback, ambiguous feedback, and constructive feedback alongside a series of lower order themes (which were colour-coded to establish which higher order theme they referred to). Therefore, it is likely that the 30 later transcriptions were highly influenced by the understanding of the first 30 essays.

Some prior attempts to analyse feedback on student assignments have only counted each criteria once irrespective of the number of times a marker might comment on something (e.g., Read et al., 2005). Given that this thesis aimed to provide a holistic and accurate view of feedback across student assignments each comment has been considered here in its own right. It might be the case that punctuation was the bugbear of this particular marker, but if this was the case then, a) we would expect this to remain the case across both the control and experimental essays, and, b) the student is going to experience receiving feedback which consistently identifies poor punctuation. To represent the data otherwise is unlikely to stay true to the students experience when they read the feedback and would ignore the important impact that repetitive criticism might have on student confidence.

3.5.7 **Hyatt's (2005) Feedback Classification System**

Following this initial categorisation in-text feedback was further analysed according to Hyatt's (2005) feedback classification system since it was considered judicious to use an already published feedback framework. This corpus-based analysis of Master's level assignments generated 7 categories and 20 subcategories and provided the most comprehensive framework within which to analyse the data (see appendix iii). Nonetheless, the preliminary inductive coding of the data was worthwhile since it demonstrated some emergent themes within the feedback which were not included in Hyatt's work. It is possible that these would not have been noticed had the first coding of the data adopted a more deductive approach and used Hyatt's classification system in the first instance. Specifically it was noticed that Hyatt coded his feedback according to comments made, yet much of the in-text feedback provided on the assignments analysed for this thesis was not comment-based. For example, some feedback constituted a tick or a cross, the underlining or crossing out of text, circling text, or the insertion of punctuation symbols such as exclamation marks. In order that these types of feedback were also included in the in-text analysis the following additions were made to Hyatt's original classification system. The first additions were to include *stylistic corrections* and *stylistic emphasis* to his *stylistic comments* category. *Content-related symbol,* and *content-related emphasis,* were added to his *content-related comments* category and finally *structural correction* was added to his *structural comments* category. As such an amended version of Hyatt's system which included observations from the preliminary analysis were used to analyse the in-text data (see appendix vi).

However, despite adopting and amending Hyatt's extensive classification system, attempts to interpret in-text feedback in line with this framework were occasionally problematic. This was because some feedback comments could be perceived to fit into more than one of the outlined codes. For example, the feedback in Table 5 from a female Chinese control essay could be coded as *developmental comment: alternative* or *content-related comment: negative evaluation* or a *stylistic comment: referencing.*

Feedback Exemplar (Chinese female 1a: Control)	Hyatt's (2005) Category	Abridged Explanation of Hyatt's (2005) Category
'Very assumption based. Any facts/research?'	Developmental comment: alternative	Comment on how improvements could be made to the work and the identification of omissions.
	Content-related comment: negative evaluation	Comment on weaknesses related to appropriateness, accuracy, evidence, clarity and criticality.
	Stylistic comment: referencing	Comment related to the use of academic language supported by appropriate sources

Table 5: Ambiguity of the coding process using Hyatt's (2005) feedback classification system

Such ambiguity is undesirable, but attempts to be consistent in the interpretation were strengthened by, a) data analysis being conducted in the shortest space of time possible, and, b) previous ambiguous feedback examples being reproduced onto another document. Therefore when the researcher found a new ambiguous example it was possible to refer back to this and thus remain consistent in the application of categories. Similar efforts to obtain consistency in data analysis when using a feedback taxonomy have been identified by Orsmond and Merry (2011). The alternative option was to double or triple-code the feedback comments. Kumar and Stracke (2007) took this approach and said that they often did this when there was disagreement between researchers about data categorisation. However, it was considered that this process might serve to misrepresent the number of comments within and across categories with some comments being counted two or three times.

However, it is worthwhile noting that ultimately interpretive judgments about how to classify assignment feedback in relation to Hyatt's classification system fell to me. This evidence of power at play between the researcher and the researched is not unusual and has been much debated in qualitative research (e.g., Kvale, 2008). Despite attempts to adopt a reflexive stance through the use of a critical friend, and the suspension of judgements through bracketing, my background, gender, ethnicity, experiences as a student, and values will have informed this interpretation. In addition my experience as a lecturer with twenty years teaching in Higher Education will undoubtedly have played a role. However, this experience was particularly useful when trying to ascribe meaning to participants' feedback in the absence of any guiding context, and/or when feedback only constituted a short amount of text or a mark (e.g. the underlining of a word). For example, one assignment included the following text with the word 'theory' underlined by the marker, "…which backs up the <u>theory</u>…" It was clear to me that the marker was querying what specific theory the student was referring to and inferring that this needed to be included. Therefore this comment was coded as *developmental comment: alternative* (i.e., the tutor points out omissions in the student's work). Nonetheless, it is also true that this interpretation is ambiguous. It is also worth considering whether this type of feedback is useful to a student who may not understand the markers intent and does not have many years' experience of working within a Higher Education environment to draw upon.

Additionally, it is important to recognise that my interpretation of the meaning of the feedback was also influenced by knowledge of the wider context of the assignment. For example, if a marker had written "Bold statement - needs evidence" several times, later when they just wrote, "Bold statement", it was reasonable to assume that they also believed the statement required evidence even though they did not explicitly state this in the feedback comment. So while the term "Bold statement" considered in isolation might be coded as *stylistic comment: register*, referring to the use of appropriate language, with the preceding context as a guide it remains as *stylistic comment: referencing*.

Ultimately in-text feedback was analysed for differences across gender, ethnicity, and the interactive effects of gender and ethnicity. The following sections demonstrate how the data was assembled to explore between group differences for each analysis.

3.5.8 **Analysis 1: Gender**

To examine between group differences within in-text feedback participants were assigned groups. Three different analyses were conducted for gender since it was considered an important part of the experimental design to attempt to isolate this as a static cue at this stage. If only one analysis of gender had been undertaken half of the participants would have marked an experimental essay supposedly written by a male student and half would have marked an experimental essay supposedly written by a female student. However, participants marked essays where the student name gave clues to a specific ethnicity as well as to gender (e.g., Jagjit Sidhu). Therefore a comparison between participants who marked Jagjit Sidhu and those who marked Natasha Brown would not only be analysing differences in expectancy effects caused by the static cue of gender, but also of ethnicity. Such a design would have laid the project open to criticism previously levelled at other research which overlooked this element of control (e.g., Sprietsma, 2013). Specifically, it would be difficult to claim that differences in feedback were solely attributable to gender since ethnicity, and the interactive effects of gender and ethnicity (the 'double jeopardy' effect [Thomas & Miles, 1995]), may also have played a role.

Consequently Analysis 1.1 consisted of participants whose experimental essay was identified as being written by a male White British student (James Smith) versus participants whose essay was identified as being written by a female White British student (Natasha Brown).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Name
1a	10	Male	Samuel Jones	Male	James Smith
1b	10	Male	Samuel Jones	Female	Natasha Brown

Table 6: Analysis 1.1 – In-text feedback: Gender (White British)

Analysis 1.2 consisted of participants whose experimental essay was identified as being written by a male Asian student (Jagjit Sidhu) versus participants whose essay was identified as being written by a female Asian student (Avinash Puri).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Name
2a	10	Male	Samuel Jones	Male	Jagjit Sidhu
2b	10	Male	Samuel Jones	Female	Avinash Puri

Table 7: Analysis 1.2 – In-text feedback: Gender (Asian)

Analysis 1.3 consisted of participants whose experimental essay was identified as being written by a male Chinese student (Zhi Rong Liu) versus participants whose essay was identified as being written by a female Chinese student (Mei Lin Pang).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Name
3a	10	Male	Samuel Jones	Male	Zhi Rong Liu
3b	10	Male	Samuel Jones	Female	Mei Lin Pang

Table 8: Analysis 1.3 – In-text feedback: Gender (Chinese)

Initially differences in the in-text feedback awarded for the control essay for each analysis were explored to identify whether differences existed when student gender remained the same (i.e., Samuel Jones). Secondly, differences in the in-text feedback awarded for the experimental essay for each analysis were explored to ascertain whether variations existed in feedback provided across female and male students when ethnicity was controlled for. Findings which demonstrated no differences for the control essay, but did show differences within analyses for the experimental essay would provide confidence that differences in the experimental essay were attributable to expectancy effects based on the static cue of student gender and not differences in marking practice.

3.5.9 **Analysis 2: Ethnicity**

Once again to examine between group differences across in-text feedback participants were assigned groups. Two different analyses were conducted with the experimental design once more ensuring that the static cue of ethnicity was isolated thus eliminating the confounding variable of gender.

Analysis 2.1 consisted of participants whose experimental essay was identified as being written by a male student of either White British (James Smith), Asian (Jagjit Sidhu) or Chinese (Zhi Rong Liu) ethnicity.

Marking Group	**No. of Participants**	**Control Essay: Ethnicity**	**Control Essay: Name**	**Experimental Essay: Ethnicity**	**Experimental Essay: Name**
1a)	10	White British	Samuel Jones	White British	James Smith
2a)	10	White British	Samuel Jones	Asian	Jagjit Sidhu
3a)	10	White British	Samuel Jones	Chinese	Zhi Rong Liu

Table 9: Analysis 2.1 – In-text feedback: Ethnicity (Male)

In Analysis 2.2 participants' experimental essays were identified as being written by a female student of either White British (Natasha Brown), Asian (Avinash Puri) or Chinese (Mei Lin Pang) ethnicity.

Marking Group	**No. of Participants**	**Control Essay: Ethnicity**	**Control Essay: Name**	**Experimental Essay: Ethnicity**	**Experimental Essay: Name**
1b)	10	White British	Samuel Jones	White British	Natasha Brown
2b)	10	White British	Samuel Jones	Asian	Avinash Puri
3b)	10	White British	Samuel Jones	Chinese	Mei Lin Pang

Table 10: Analysis 2.2 – In-text feedback: Ethnicity (Female)

Differences in the in-text feedback awarded between groups one to three were first established for the control essay in order to ascertain whether differences existed when inferred student ethnicity remained the same. Next differences in the in-text feedback for the experimental essay for each analysis were explored to establish if variations were present in the feedback markers provided across White British, Asian, or Chinese students. Once again findings which showed no differences between groups in the control essay but did reveal differences in the experimental essay would indicate that differences in the experimental essay were as a result of expectancy effects based on the static cue of student ethnicity and not simply differences in marking practice.

3.5.10 **Analysis 3: Gender and Ethnicity**

The final in-text feedback analyses examined differences pertaining to both gender and ethnicity. Six analyses were conducted to ensure that each gender had been compared to each ethnicity. Analysis 3.1 included participants whose experimental essay was identified as being written by a student who was either male white British (James Smith), or female Asian (Avinash Puri).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Ethnicity	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Ethnicity	Experimental Essay: Name
1a)	10	Male	White British	Samuel Jones	Male	White British	James Smith
2b)	10	Male	White British	Samuel Jones	Female	Asian	Avinash Puri

Table 11: Analysis 3.1 – In-text feedback: Gender and ethnicity (WBM & AF)

Analysis 3.2 consisted of participants whose experimental essay was identified as being written by a student who was either male White British (James Smith) or female Chinese (Mei Lin Pang).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Ethnicity	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Ethnicity	Experimental Essay: Name
1a)	10	Male	White British	Samuel Jones	Male	White British	James Smith
3b)	10	Male	White British	Samuel Jones	Female	Chinese	Mei Lin Pang

Table 12: Analysis 3.2 – In-text feedback: Gender and ethnicity (WBM & CF)

Analysis 3.3 consisted of participants whose experimental essay was identified as being written by a student who was either male Asian (Jagjit Sidhu) or female Chinese (Mei Lin Pang).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Ethnicity	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Ethnicity	Experimental Essay: Name
2a)	10	Male	White British	Samuel Jones	Male	Asian	Jagjit Sidhu
3b)	10	Male	White British	Samuel Jones	Female	Chinese	Mei Lin Pang

Table 13: Analysis 3.3 – In-text feedback: Gender and ethnicity (AM & CF)

Analysis 3.4 involved participants whose experimental essay was identified as being written by a student who was either male Asian (Jagjit Sidhu), or female white British (Natasha Brown).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Ethnicity	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Ethnicity	Experimental Essay: Name
2a)	10	Male	White British	Samuel Jones	Male	Asian	Jagjit Sidhu
1b)	10	Male	White British	Samuel Jones	Female	White British	Natasha Brown

Table 14: Analysis 3.4 – In-text feedback: Gender and ethnicity (AM & WBF)

Analysis 3.5 involved participant groups whose experimental essay was identified as being written by a student who was either male Chinese (Zhi Rong Liu), or female white British (Natasha Brown).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Ethnicity	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Ethnicity	Experimental Essay: Name
3a)	10	Male	White British	Samuel Jones	Male	Chinese	Zhi Rong Liu
1b)	10	Male	White British	Samuel Jones	Female	White British	Natasha Brown

Table 15: Analysis 3.5 – In-text feedback: Gender and ethnicity (CM & WBF)

Analysis 3.6 included participant groups whose experimental essay was identified as being written by a student who was either male Chinese (Zhi Rong Liu), or female Asian (Avinash Puri).

Marking Group	No. of Participants	Control Essay: Gender	Control Essay: Ethnicity	Control Essay: Name	Experimental Essay: Gender	Experimental Essay: Ethnicity	Experimental Essay: Name
3a)	10	Male	White British	Samuel Jones	Male	Chinese	Zhi Rong Liu
2b)	10	Male	White British	Samuel Jones	Female	Asian	Avinash Puri

Table 16: Analysis 3.6 – In-text feedback: Gender and ethnicity (CM &AF)

As before in-text feedback differences provided between groups for the control essay were explored first to establish any differences when perceived student gender and ethnicity were the same (i.e., Samuel Jones). Finally, the in-text feedback for the experimental essay was explored for each analyses to determine whether feedback differed when both gender and ethnicity changed. Once again, results which showed no differences between groups in the control essay yet showed differences between groups in the experimental essay would provide confidence that differences in the experimental essay were attributable to expectancy effects based on the interactive effects of student gender and ethnicity as opposed to differences in marking practice.

While it was considered important to control possible confounding variables in the first two analyses of in-text feedback, the design for the final analysis surrenders this control in an attempt to explore the interactive nature of expectancies. Research investigating how a perceiver's judgement of one piece of information can be dependent upon the information that accompanies it has been scarce, although knowledge of this concept and the related concept of trait centrality (where some traits assume greater importance for how individuals are perceived than others) has been evident since the 1940s (Asch, 1946). More recently researchers have spoken about the 'double jeopardy' phenomenon where individuals who belong to more than one persecuted group (e.g., female and Asian) are at risk of additional expectancy-related biases. Therefore it was considered important to explore whether it was possible to see evidence of these concepts at work within the in-text feedback provided to students.

3.5.11 **Analysis of Summary Feedback**

In order to analyse the summary feedback hierarchical content analysis was chosen. Hierarchical content analysis allows the researcher to identify patterns within the data and explore how these patterns interact in a hierarchical manner. Therefore researchers can compare and contrast what is in the data, divide it into larger and smaller categories, and describe and order the content. Sparkes and Smith (2014) claim that this allows general knowledge about a topic to be developed and can reveal how groups of people behave. Furthermore, it can operate at either a manifest or latent level (Hsieh & Shannon, 2005). Manifest analysis is simply concerned with *what* is in the text (e.g., how much, how many), whereas latent analysis allows the research to move beyond the superficial and explore potential underlying *meanings* and intentions. In order to realise a credible latent analysis the researcher should through the data several times and engage with the process of Epoche or bracketing (Husserl, 1931). These terms refer to a researchers attempt to look before making judgments in order to remove or at least be aware of their own prejudices and viewpoints regarding the phenomenon under investigation. This close connection with the data also means that a hierarchical content analysis allows the researcher to indwell (Maykut & Morehouse, 1994) which involves being immersed in the data and adopting an empathic position. Furthermore, it allows themes to emerge inductively from the data corpus as opposed to being dictated by preconceived frameworks or ideas. As such hierarchical content analysis allows the researcher to use an appropriate method which aligns itself more closely to some key qualitative concepts.

The variety of different types of content analysis, combined with the lack of procedural clarity has led to content analysis being criticised (Tesch, 1990). These criticisms make it increasingly important that researchers using any form of content analysis are clear about the what, why, and how of their procedures. A lack of transparency in procedures makes research evaluation and comparison difficult (Attride-Sterling, 2001) and, as previously addressed has implications for trustworthiness and credibility (Biddle et al., 2005). The procedural guidelines adopted here were those identified by Sparkes and Smith (2014). These were as follows; a) immersion, b) search for, identify, and label themes, c) connecting and ordering themes, d) cross-checking, e) confirmation, f) produce a table (see appendix vii). Coding the data involved organising similar data into categories, comparing and contrasting data in the categories to connect quotes with similar meanings, attaching labels or tags to pieces of the data to identify them as meaningful, grouping together smaller units of similar data (as identified by the tags) into sub-themes, and finally, creating higher-order themes which group together and say something meaningful about the content of several subthemes.

4 RESULTS

4.1 In-Text Feedback

In line with Hyatt's (2005) amended classification system, in-text feedback consisted of six categories, ten subcategories, and thirty-seven further subcategories. The discussion that follows will only consider the categories that attracted the most in-text feedback, although the results for all categories are represented in the tables and diagrams available in the appendices (see appendix viii). A total of fifty-seven out of sixty participants (95%) provided in-text feedback. The three participants who failed to did so for both the control and experimental essay. Therefore there did not appear to be any bias in terms of choosing to provide feedback to one type of name over another.

This chapter first presents the findings from the gender, ethnicity, and gender and ethnicity analyses for in-text feedback in relation to expectancy effects. Only findings from the experimental condition are reported here, although control condition analyses are available in the appendices (see appendix ix) and referred to throughout for means of comparison. The strongest claims for the existence of expectancy effects are based upon finding no between-groups differences in feedback practices in the control condition (when groups marked the same essay written by a student with the same name) but finding differences in the experimental condition (when groups marked the same essay written by students with different names). These types of differences provide confidence that expectancy effects based on knowledge of the student name have impacted on the feedback provided as opposed to just marker variability. As such between-group differences in the experimental condition will only be reported on when negligible or no differences existed in the control condition.

Between-group differences refer to differences observed between markers in one group and markers in other group and reflect findings where experimental control has been maintained (e.g. comparing between groups in the control and then between groups in the experimental condition). Within group differences refer to changes observed within the same group of markers from the control condition to the experimental condition. It was not the primary concern of this research to compare within groups and across conditions. Nonetheless, when analysing the results it became apparent that there were often large differences in the feedback provided within marking groups. If experimental design and control was to be maintained these results would simply be attributed to marker variability and dismissed. However, the change in some marking groups' feedback provision was so distinct that on occasions it has been included too.

Nonetheless, it is acknowledged that findings pertaining to within group changes across conditions do not hold the same weight as those pertaining to between group differences within conditions since the experimental control has been compromised.

Nevertheless, it is of interest that the same group of markers could demonstrate such varying in-text feedback behaviours when moving from marking a WBM (Samuel Jones) in the control condition to an AF (Avinash Puri) in the experimental condition for example. However, it is important to acknowledge that the student name in the control condition was always that of a WBM. Therefore, if differences were noted between the WBM (control) and the AF (experimental) it was not possible to maintain control over the variable of ethnicity as was the case in the experimental condition where Avinash Puri was compared to Jadjit Sidhu. This makes it impossible to ascertain whether these differences are gender or ethnicity based, or a combination of the two. Moreover, is important to remember that the control condition essay and the experimental condition essay were different. Therefore the feedback provided was most likely to have been reflective of the essay content rather than the student name. Cognisant of this, when differences were noted within a specific marking group corresponding differences were sought in the remaining marking groups to examine whether those differences were ubiquitous and therefore more reflective of the essay content than expectancy effects. In recognition of these limitations, when these findings are reported they will not be compared to the findings from previous analyses since they are not always comparing the same variables. Nonetheless, it was considered that these differences in feedback provision also indicated expectancy effects in operation within marking groups. Admittedly it is difficult to draw firm conclusions using the within marking group data in comparison to the between marking groups data and therefore when differences are discussed it will be made explicit where these emanated from.

Throughout this chapter references will be made to comment, corrective, emphasis and symbol-based feedback. It is important to understand the distinctions between these types. Comment-based feedback is the most substantial form of feedback and provides the best opportunity for students to learn. It can serve developmental and educative functions and be phrased in a positive, negative or neutral way. Corrective feedback amends mistakes without explaining why the correction was necessary. As well as having potentially negative connotations it therefore includes an element of ambiguity. Emphasis-based feedback often simply highlights something (by underlining some text for example). This feedback contains the highest level of ambiguity since the tutor’s motivation is not explicit and the action needed is likely to be unclear to the student. As such it has a limited impact on learning. Finally, symbol-based feedback refers largely to the provision of ticks and crosses. This type of feedback has been identified by students as commonplace but ineffective (Price et al., 2010). While a student might know they have got something right or wrong, they might not be aware what this is. Without this knowledge they are less likely to be able to correct mistakes or repeat good work. Therefore when reading this chapter it is important to be mindful that gaining large amounts of some types of feedback (e.g., emphasis-based feedback) might not necessarily advantage the student, particularly if they are provided at the exclusion of other more meaningful feedback.

Feedback for all in-text feedback analyses will now be discussed in relation to the results tables (see appendix viii).

4.2 In-Text Feedback Analysis 1: Perceived Student Gender

4.2.1 Analysis 1.1: White British Male vs. White British Female

This analysis compared the in-text feedback of participants placed into Group 1a or Group 1b. All participants first marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 1a's experimental essay was labelled as being written by a WBM (James Smith) and Group 1b's experimental essay was labelled as being written by a WBF (Natasha Brown). The content of the essay did not change, only the student name.

Analysis 1.1 Gender: White British Male Vs. White British Female (Experimental)		
	WBM (Group 1a)	**WBF (Group 1b)**
Phatic Feedback	0%	0%
Developmental Feedback	22%	16%
Structural Feedback	5%	3%
Stylistic Feedback	53%	51%
Content-related Feedback	20%	30%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 17: Analysis 1.1: In-text feedback classifications for WBM versus WBF (Experimental)

The experimental essay generated different numbers of feedback contributions, with Group 1a (WBM) scoring 343, and Group 1b (WBF) scoring 219. It was evident that total feedback contributions had increased slightly for Group 1a (WBM) in the experimental condition, but decreased for Group 1b (WBF). However, since there were differences evident in the control condition it is likely that these differences can be accounted for by marker variability and not gender-based expectancy effects.

Patterns within the data showed that stylistic feedback was the main type of feedback provided by both groups, scoring 53% for Group 1a (WBM) and 51% for Group 1b (WBF). Subcategories for stylistic feedback showed that *stylistic comments* dominated, followed by *stylistic corrections* and *stylistic emphasis*. However, despite these similarities there were also some differences at further subcategory level within stylistic feedback. For example, while feedback scores related to *punctuation* were comparable at *stylistic comment* and *stylistic correction* levels there were vast differences in *stylistic emphasis*. Specifically, 71% of emphasis-based feedback related to *punctuation* for Group 1a (WBM), whereas this was only 7% for Group 1b (WBF). Although there was a difference of 10% between groups in the control condition this increase reflects a difference that is unlikely to be attributed to marker variability (especially since differences remained comparable with the control condition for other aspects of punctuation-related feedback). Therefore expectancy effects related to student gender might be accountable for such differences.

Another emphasis-based element of stylistic feedback which generated differences was *syntax/word order/grammar*. Scores were 7% for Group 1a (WBM), but reached 36% for Group 1b (WBF). Given that scores were comparable in the control condition (including for *stylistic comments* and *stylistic corrections* related to this category) it can be assumed that markers were generally marking similarly. However, when markers in Group 1b assumed the work was authored by a WBF student they emphasised errors in this domain almost thirty percent more. It therefore seems possible that these differences can be attributed to expectancy effects on the basis of student gender.

The further subcategory related to *stylistic comment: referencing/citation/ quotation/ bibliography* also generated some differences. Group 1a (WBM) received 58% of their total *stylistic comment* scores in this domain, whereas Group 1b (WBF) only received 30%. This amounted to the essay with the WBM name gaining sixty-two comments about referencing on their work whereas the WBF name only gained fourteen. Although there were between group differences in the control condition, this only amounted to 11%, with the difference in the number of comments being three. Given that referencing is a key skill in higher education and provides evidence as to the academic integrity of the work this inequity would seem to disadvantage WBFs.

Content-related feedback gained the next highest scores, comprising 20% of the total feedback for Group 1a (WBM) and 30% for Group 1b (WBF). Although this difference may seem large it is unlikely that it can be attributed to expectancy effects linked to gender since the control essay also revealed similar percentage differences across groups. Therefore it seems likely that these differences can be explained as differences in marking practices. However, when the subcategories were examined for differences in the make-up of content-related feedback it was evident that Group 1a (WBM) attracted more comments in the *positive evaluation* category (56%) than Group 1b (WBF) who only gained 43%. This amounted to Group 1a (WBM) receiving twenty-seven positive comments on their work, while Group 1b (WBF) only received fifteen. Additionally, Group 1a (WBM) gained fewer comments in the *negative evaluation* category (40%) than Group 1b (WBF) who gained 57%. Since such differences were not present in the control condition it is likely that these differences can be explained by expectancy effects related to student gender. Feedback related to *content-related symbols* was 100% positive for both groups.

While percentage scores related to *developmental comments* were almost identical for the control essay, the experimental essay showed that more developmental feedback was provided for Group 1a (WBM, 22%) than for Group 1b (WBF, 16%). When the subcategories were explored for differences the data showed that Group 1a (WBM) received less feedback (42%) in the *developmental comment: alternative* subcategory than Group 1b (WBF, 58%). These differences were not apparent in the control essay. However, it is important to note that because Group 1b (WBF) got a lower percentage score for *developmental comments* overall, their 58% score for the *alternatives* subcategory only translated to twenty-one comments on their work. Therefore although the percentage score for *alternatives* was lower at 42% for Group 1a (WBM) they actually received thirty-two comments which outweighed those received by their female counterparts.

It is important to reflect upon how these percentages translate to number of comments on the assignment. This is because students are not likely to categorise tutors responses and analyse the percentages for each type of feedback they receive. Rather they are likely to use more intuitive skills to get a sense of the overall feedback landscape and the message it conveys. As such the number of comments tutors provide on specific aspects of the work is likely to be much more impactful for students, and will therefore shape their interpretation of their feedback more than percentages would. Consequently, these analyses will often translate percentage scores to numbers of comments.

The *alternatives* subcategory as described by Hyatt (2005) includes feedback where the tutor offers alternatives, suggestions or identifies omissions in the work. In practice, the assignments analysed for this thesis were dominated by the identification of omissions as opposed to the offering of alternatives or suggestions. The identification of omissions was not considered sufficient to be coded as a *negative evaluation* in the *content-related comments* category because often the feedback was a statement and not a judgement (e.g. *"You could have included x"* as opposed to *"This overview is poor"*). Nonetheless, comments in the *alternatives* category often had a more negative tone than a developmental one. This is important to consider when the data showed that assignments bearing a WBM name attracted significantly more feedback identifying what they had failed to include in their work than assignments bearing a WBF name.

Nonetheless, *alternative* feedback is still considered developmental, and therefore if students can interpret it in a positively and not let it negatively impact their motivation and self-efficacy (Hattie & Timperley, 2007; Van Dinther et al., 2011; Nash et al., 2015) it could benefit them in future assignments. Therefore the efficacy of this type of feedback to positively develop students will be dependent upon their interpretation of it. One argument for this result is that WBMs are advantaged by this feedback whereas WBFs have lost a valuable learning opportunity. Perhaps the provision of more of this style of feedback to WBMs illustrated a gender-based bias on the part of the markers that male students are better equipped to cope with this type of feedback than their female counterparts. The counter argument is that WBMs are disadvantaged by receiving negatively phrased developmental feedback instead of more positively phrased examples.

Group 1a (WBM) also received more than double the amount of feedback in the *developmental comments*: *reflective questions* subcategory (45% versus 19%). This amounted to Group 1a (WBM) receiving thirty-four reflective comments on their work whereas Group 1b (WBF) received only seven. While there was a 15% difference between groups for the control condition it was more pronounced for the experimental essays at 26%. Furthermore, it is important to note that the 15% difference in the control condition only equated to a difference of two comments across marking groups. These findings therefore indicate that expectancy effects related to student gender have played a role. *Reflective questions* include feedback where the tutor asks a question for the student to reflect upon (Hyatt, 2005). Interestingly this feedback was provided to assignments bearing WBM names more than twice as often as it was to essays bearing WBF names. Given that reflective questions are principally used to stimulate thought and challenge intellectual capacity, it is important that all students have equal opportunities to learn from such feedback.

Developmental comments: future scored 0% for Group 1a (WBM) and only 0.5% for Group 1b (WBF).

Structural comments attracted similar percentages across groups (Group 1a = 5% and Group 1b = 3%) and *sentence level* comments prevailed over *discourse level* comments for both groups.

In summation, Analysis 1.1 demonstrated that Group 1b (WBF) gained less positive feedback, more negative feedback and less developmental feedback than Group 1a (WBM). Additionally, within the developmental category, WBF students were asked fewer *reflective questions* (which stimulate thought and challenge intellectual capacity). WBMs were provided with more *alternative* comments (which largely highlighted omissions in the work, but also provided them with a learning opportunity), and on balance their feedback was much more positive and developmentally oriented. Given that according to Hyatt's classification system, developmental feedback is provided with the "intention of aiding the student with subsequent work in relation to the current assignment", WBFs are disadvantaged when they are not provided with an equitable chance to do this alongside their WBM counterparts. Furthermore, the combination of receiving less developmental feedback, alongside less positive feedback, and more negative feedback only serves to intensify the problem and demonstrates that WBFs received what might be considered an unhappy triad of feedback.

4.2.2 **Analysis 1.2: Asian Male vs. Asian Female**

This analysis compared the in-text feedback of participants in Group 2a or Group 2b. All participants first marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 2a's experimental essay was labelled as being written by an AM (Jagjit Sidhu) and Group 2b's experimental essay was labelled as being written by an AF (Avinash Puri). The content of the essay did not change, only the student name.

Analysis 1.2 Gender: Asian Male Vs. Asian Female (Experimental)		
	AM (Group 2a)	**AF (Group 2b)**
Phatic Feedback	0%	1%
Developmental Feedback	15%	22%
Structural Feedback	3%	3%
Stylistic Feedback	56%	36%
Content-related Feedback	26%	39%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 18: Analysis 1.2: In-text feedback classifications for AM versus AF (Experimental)

The total number of feedback contributions across groups was different with Group 2a (AM) providing 322 and Group 2b (AF) only 242. It was evident that total feedback contributions had increased slightly for Group 2a (AM) in the experimental condition, but decreased for Group 2b

(AF). However, differences were also evident in the control condition thus making these inconsequential.

Although stylistic feedback was the highest scoring category for Group 2a (AM) at 56% it was only the second highest scoring category for Group 2b (AF) at 36%. Although the percentage of feedback contributions related to stylistic feedback was also different in the control condition (demonstrating a 15% difference between groups), both groups still awarded the most feedback to stylistic elements of the work at this stage. Subcategories of stylistic feedback showed that for Group 2a (AM) *stylistic comments* dominated, followed by *stylistic corrections* and *stylistic emphasis*, whereas for Group 2b (AF) *stylistic corrections* dominated and were followed by *stylistic comments* and *stylistic emphasis.* While there were differences present in the amount of comments received for each subcategory in the experimental condition these differences were also present in the control condition and therefore there is no evidence that at subcategory level stylistic feedback altered in accordance with the gender of the student.

Nevertheless, when the further subcategories of stylistic feedback were examined there were some differences worth reporting. For example, Group 2b (AF) gained 72% of their corrective feedback in the domain of *punctuation*, whereas for Group 2a (AM) this was only 50%. However, when this was translated into the number of corrections actually made on the work Group 2a (AM) gained forty corrections whereas Group 2b (AF) gained thirty-four. Furthermore, there were differences apparent in this area within the control condition making it difficult to establish the presence of expectancy effects in this instance. Nonetheless, there were some interesting within group changes across conditions. When markers in both groups perceived themselves to be marking an essay authored by a WBM (in the control condition) their desire to correct punctuation was substantially lower (Group 2a = 36% and Group 2b = 21%) and these percentages also correlated with a lower amount of corrections on the work. Therefore both sets of markers increased their corrective feedback on *punctuation* (Group 2a increased by 14% and Group 2b by 51%) when the perceived author of the work was Asian instead of White British which may indicate expectancy effects operating on the basis of ethnicity.

Additional within groups differences across conditions existed at further subcategory level within *stylistic correction: referencing/citation/quotation/bibliography* feedback. These differences did not exist between groups in the experimental condition with only a small 3% difference apparent between Group 2a (AM) and Group 2b (AF). However, when within groups differences were observed markers in Group 2b had offered 44% of corrective feedback on referencing (when marking a WBM name in the control condition) which amounted to twenty-three corrections on the essay. However, in the experimental condition the same set of markers, now marking an essay assigned with an AF name only offered 4% of their feedback in this manner, which consisted of only two corrections. Admittedly, this interpretation of the data is once again within groups and across conditions which makes it difficult to claim with certainty that these results are a consequence of expectancy effects related to gender. However, it does illustrate how when moving from marking a WBM to an AF feedback behaviour altered significantly.

Furthermore, these results might indicate a lack of expectancy effects in operation on the basis of gender when both students are Asian. The differences seen across conditions might demonstrate that effects related to gender only become visible when the ethnicity of the student also changes, so that comparing a WBM to an AF generated expectancies which were not present when an AM was compared to an AF. Such findings lend credence to Asch's (1946) work on the interactive effects of expectancies, such that the interpretation of one piece of information can be altered according to the other information presented alongside it. In this example, maybe only small changes were observed in the experimental condition because when ethnicity remained stable, the gender of the student became less important to the perceiver. Later analyses compare the interactive effects of gender and ethnicity.

Content-related feedback gained the second highest feedback contribution scores for Group 2a (AM) at 26% and the highest score for Group 2b (AF) at 39%. This percentage differential was not present in the control condition where the difference between groups was just 2%. Furthermore, whereas in the control condition the subcategory *content-related comments* had dominated over *content-related symbols* for both groups this was different for the experimental condition. Specifically, Group 2a (AM) continued to follow the identified pattern, but Group 2b (AF) received less feedback related to content (12% versus 19%). This percentage difference amounted to Group 2b (AF) receiving only half the comments of their male counterparts (thirty comments versus sixty). Interestingly Group 2b (AF) did receive more symbol-based feedback (26% versus 8%) than Group 2a (AM). Although there was a difference of 5% between groups in the control condition the difference in the experimental condition was 18% and thus points towards the activation of expectancy effects on the basis of student gender. While this symbol-based feedback was 100% positive, it is debatable how much students can learn from such feedback in comparison to more substantial comment-based feedback. As has been noted previously, seeing a tick or a cross would provide a student with some indication as to the worth of what they had written, but it would not serve to explain why what they had written was useful or not and therefore only surface learning can ensue. This data therefore points to the fact that female

Asian students had fewer opportunities to learn something meaningful from their feedback than their AM counterparts. Previous research shows that students perceive symbols such as ticks and crosses as unhelpful (Price et al. 2010) and instead require "… an explanation of how to improve" (Ferguson, 2011, p.56). It may also support Jampol's (2014) findings that feedback provided to female authors was more compassionate (hence more ticks) but less accurate than that provided to males (hence fewer comments).

Further subcategories of content-related feedback showed that Group 2a (AM) received more comments in the *positive evaluation* category (58%) than Group 2b (AF) who received 43%. However, similar differences were also seen between groups in the control condition and therefore can only be attributed to differences in marking practices between groups.

In terms of *negative evaluation* comments, Group 2a (AM) received 37% whereas Group 2b (AF) received 57%. While there was an 11% difference in this category in the control condition, this almost doubled in the experimental condition. Nonetheless, when the number of comments were examined it became clear that Group 2a (AMs) received twenty-two negative comments on their work while Group 2b (AF) received seventeen. Therefore the larger percentage score did not actually translate to females receiving more negative *content-related comments*. Group 2b (AF) also received more negative comments in the control condition despite their percentage score in this domain being lower than Group 2a (AM). Therefore it is unlikely that expectancy effects related to student gender have played a role in this instance.

Nonetheless, markers propensity to provide more positive feedback and less negative feedback changed extensively across conditions. For example, markers in both groups did not show an inclination to provide positive *content-related comments* when marking the control essay (perceived to have been written by a WBM student). However, when they perceived themselves to be marking the work of an Asian student (either male or female) their behaviour changed significantly. Markers in Group 2a (AM) provided 33% more and markers in Group 2b (AF) provided 35% more than in the previous condition. There were concomitant reductions in the amount of negative feedback provided across conditions too. Of course it is possible that the second essay warranted more positive comments than the first essay, although given that these essays were chosen because they were of a similar academic standard (lower second class), this makes this interpretation less likely. Therefore taken together these results more likely indicate that expectancy effects related to ethnicity rather than gender were a catalyst for the changes in feedback behaviour across marking groups, since when only the gender of the student changed and ethnicity remained constant (in the experimental condition) smaller changes were witnessed.

Content-related symbols were once more overwhelmingly positive for both groups at 96% and 100%.

Developmental feedback comprised the third largest feedback component for both groups scoring 15% for Groups 2a (AM) and 22% for Group 2b (AF). These differences in the total percentage of feedback contributions were comparable to the control condition. When the subcategories were examined (see Table 19) it became evident that despite receiving less overall developmental feedback Group 2a (AM) received more feedback related to *alternatives* than Group 2b (AF), scoring 62% and 49% respectively. These percentages also translated to Group 2a (AM) receiving more comments on their work. These differences were not apparent in the control essay and support the results from Analysis 1.1 which demonstrated that although WBF names attracted a higher percentage of feedback in the *alternatives* subcategory it was WBM names that received substantially more comments on their work. Nonetheless, Group 2a (AM) received less feedback (30% versus 43%) and fewer comments (fourteen versus twenty-three) related to *reflective questions*. This difference was almost twice as big as that of the control condition and suggests that expectancy effects have potentially played a role. Additionally, this result reverses the trend of Analysis 1.1 where WBF essays attracted over 50% fewer comments than WBM essays on *reflective questions*.

	Developmental Comment: alternative	**Developmental Comment: future**	**Developmental Comment: reflective question**	**Developmental Comment: informational comment**
Group 2a Experimental (AM)	62%	0%	30%	9%
Group 2b Experimental (AF)	49%	0%	43%	8%

Table 19: Analysis 1.2: Developmental feedback differences across subcategories for AMs and AFs

The differences between developmental feedback subcategories was much smaller for AMs versus AFs than for WBMs versus WBFs. Nonetheless, the data demonstrated that assignments in Group 2a (AM) attracted the negative feedback previously identified as synonymous with the *alternative* subcategory more often. Furthermore, Group 2a (AM) failed to receive feedback related to *reflective questions*, a feedback type which can provide the impetus for deep learning and critical thinking. There were no *developmental comments* related to *future* for either group.

Structural feedback gained identical percentages across groups at 3%. *Sentence level* comments dominated over *discourse level* comments for both groups.

So for this analysis, Group 2b (AF) gained less *content-related comment* feedback and more *content-related symbol* feedback which might prevent deeper learning. Although initially findings suggested that Group 2b (AF) received more negative feedback than their male peers (in support of Analysis 1.1 where WBF students received more negative feedback) the number of comments made on the work suggested this not to be the case. In contrast to Analysis 1.1, assignments bearing an AM name (Group 2a) received more *alternative* comments and fewer *reflective questions* than assignments bearing an AF name (Group 2b). This indicates that Group 2a (AM) received more feedback which highlighted omissions in the work and less feedback to inspire intellectual stimulation.

4.2.3 **Analysis 1.3: Chinese Male vs. Chinese Female**

This analysis compared the in-text feedback of participants in Group 3a or Group 3b. All participants first marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 3a's experimental essay was labelled as being written by a CM (Zhi Rong Liu) and Group 3b's experimental essay was labelled as being written by a CF (Mei Lin Pang). The content of the essay did not change, only the student name.

Analysis 1.3 Gender: Chinese Male Vs Chinese Female (Experimental)		
	CM (Group 3a)	**CF (Group 3b)**
Phatic Feedback	0%	0%
Developmental Feedback	10%	16%
Structural Feedback	1%	3%
Stylistic Feedback	52%	50%
Content-related Feedback	37%	31%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 20: Analysis 1.3: In-text feedback classifications for CM versus CF (Experimental)

Feedback contributions for the experimental condition also differed across groups; Group 3a (n=233) and Group 3b (n=326). It was evident that total feedback contributions had reduced for both groups. However, since there were differences evident in the control condition it is likely that these differences can be accounted for by marker variability and not gender-based expectancy effects.

Feedback related to stylistic components of the work received the highest scores as was also the case in the control condition. Furthermore, results were almost identical across groups with Group 3a (CM) gaining 52% of this type of feedback and Group 3b (CF) gaining 50%. Since the scores for stylistic feedback were also identical within the control condition this seemed to indicate that the gender of the student had no influence over marking practices in this instance. This was further borne out when subcategories were examined. *Stylistic comments*, *stylistic corrections*, and *stylistic emphasis* were ranked in the same order across groups and attracted similar percentage scores for each subcategory, a finding that was also evident in the control condition.

However, closer inspection of the data did reveal some differences. Specifically, under the further subheading *stylistic corrections*: *referencing/citation/quotation/bibliography*, Group 3a (CM) scored 33% and Group 3b (CF) scored 17%, reflecting a 16% difference between groups and translating to males receiving over a third more corrections on their work. Although there was a difference in the control condition this only amounted to 6% between groups. Therefore, although some of this difference might be attributed to variations in marking practices there is also evidence that expectancy effects related to student gender may have played a role.

Similarly when exploring the *stylistic emphasis: presentation* subcategory there was a large score differential between Group 3a (CM, 40%) and Group 3b (CF, 0%) demonstrating that assignments presented as being written by a male student attracted more feedback on presentation. However, this large percentage difference only equated to Group 3a (CM) receiving six pieces of emphasis-based presentational feedback on their work versus Group 3b (CF) receiving zero. There was also a difference between groups in the control condition (Group 3a =17% and Group 3b = 0%) with Group 3a (CM) receiving four comments and Group 3b (CF) receiving zero. It therefore appears as if the markers that constitute Group 3a (CM) do have a tendency to emphasise presentational issues when marking and therefore these differences can only be attributed to differences in marking practices.

Content-related feedback attracted the second highest number of feedback contributions for both groups (Group 3a = 37% and Group 3b = 31%). Similar percentage differences were identifiable across groups in the control condition. *Content-related comments* also dominated over *content-related symbols* for both groups, a pattern also replicated in the control condition. Feedback related to *positive evaluative comments* and *symbols* and *negative evaluative comments* and *symbols* were very similar across groups in both the experimental and control conditions suggesting no indication of expectancy effects at work in this instance.

Scores for *developmental comments* differed by 6% across groups. The same percentage differential was found in the control condition. The majority of *developmental comments* related to *alternatives,* then *reflective questions*, and finally *informational comments*. This order was also true of the control condition. There were differences in scores for each of these subcategories, but since these differences were duplicated in the control condition these can be attributed to differences in marking practices between groups. There were no *developmental comments* related to *future* for either group.

Structural feedback once more attracted the fewest number of feedback contributions with Group 3a (CM) scoring 1% and Group 3b (CF) scoring 3%. Scores in the control condition were 4% across both groups. Although reflecting on a category which only constituted a small amount of the overall feedback picture might be considered overly meticulous, differences emerged within the experimental condition when the subcategories were examined.

Specifically, 100% of the *structural comments* for Group 3a (CM) were at a *sentence level*. This group therefore did not receive any *structural comment* feedback at *discourse level*. This is compared to Group 3b (CF) who received 45% of their *structural comment* feedback at *sentence level* and 55% at *discourse level*. Since these differences were absent from the control condition it remains likely that such differences result from expectancy effects and not marking practices. Furthermore, Group 3a (CM) only received three comments on structure overall, whereas Group 3b (CF) received eleven. In the control condition, when marking work presented as written by a WBM, markers in Group 3a (CM) had shown a willingness to provide feedback on structural issues providing nine comments, but when marking work presented as written by a CM their propensity to provide such feedback reduced.

In sum, this analysis revealed fewer gender-based differences than previous analyses. Additionally, one of the differences surrounded what is arguably a less significant type of feedback i.e., *stylistic corrections*: *referencing/citation/quotation/bibliography*. Nonetheless, it is interesting that when markers perceived the essay to be written by a male student there was a greater need to correct mistakes related to referencing than when the same essay, containing identical errors, was thought to have been written by a female student. This was the first analyses to demonstrate differences in structural feedback, and while this constituted a small percentage of the overall feedback picture it remained true that (Group 3b (CF) received more structural feedback that Group 3a (CM).

4.3 In-Text Feedback Analysis 2: Perceived Student Ethnicity

4.3.1 Analysis 2.1: Male Ethnicities

This analysis compared the in-text feedback of participants placed into Group 1a, Group 2a and Group 3a. All participants first marked the control essay presented as being written by Samuel

Jones and then an experimental essay. Group 1a's experimental essay was labelled as being written by a WBM (James Smith), Group 2a's was labelled as written by an AM (Jagjit Sidhu), and Group 3a's was labelled as written by a CM (Zhi Rong Liu). The content of the essay did not change, only the student name.

Analysis 2.1: Ethnicity **White British Male Vs. Asian Male Vs. Chinese Male Experimental**			
	White British Male (Group 1a)	**AM (Group 2a)**	**CM (Group 3a)**
Phatic Feedback	0%	0%	0%
Developmental Feedback	22%	15%	10%
Structural Feedback	5%	3%	1%
Stylistic Feedback	53%	56%	52%
Content-related Feedback	20%	26%	37%
Administrative Feedback	0%	0%	0%
Ambiguous Feedback	0%	0%	0%

Table 21: Analysis 2.1: In-text feedback classifications for male ethnicities (Experimental)

Total feedback contributions once more generated different numbers for each group. Group 1a (n=343), Group 2a (n=322) and Group 3a (n=233). Total feedback contributions also varied in the control condition and therefore these differences can be accounted for by marker variability and not ethnicity-based expectancy effects.

As with the control condition, stylistic feedback overshadowed other types of feedback for all groups. Percentage scores were similar across groups with Group 1a (WBM) totalling 53%, Group 2a (AM) scoring 56%, and Group 3a (CM) reaching 52%. Furthermore, for all groups *stylistic comments* were ranked above *stylistic corrections* which outranked *stylistic emphasis.*

However, despite the apparent similarities within stylistic feedback, an examination of the further subcategories exposed some differences. For example, *stylistic comment: punctuation* only scored between 0-2% across groups in the control condition, but in the experimental condition the range was between 1-9%. Group 1a (WBM) scored 1%, Group 2a (AM) scored 6%, and Group 3a (CM) scored 9%. Corrective feedback scores on *punctuation* were much higher; Group 1a (WBM) scored 46%, Group 2a (AM) scored 50%, and Group 3a (CM) scored 60%. Comparable differences were witnessed in the control condition for corrections however. Nonetheless, when *stylistic emphasis: punctuation* was examined there were huge differences between groups. Group 1a (WBM) received the most feedback in this further subcategory with 71% of their emphasis-based feedback being related to *punctuation*, followed by Group 3a (CM) at 40% and then Group 2a (AM) at 18%. This constituted a 53% difference across groups in the experimental condition compared to a 14% difference in the control condition.

	Stylistic Comment: Punctuation	**Stylistic Correction: Punctuation**	**Stylistic emphasis: Punctuation**
Group 1a Experimental (WBM)	1%	46%	71%
Group 2a Experimental (AM)	6%	50%	18%
Group 3a Experimental (CM)	9%	60%	40%

Table 22: Analysis 2:1 -Stylistic Feedback differences in punctuation across male ethnic groups

As the table above indicates, markers were more likely to comment on the punctuation-related elements of a students work if they had an Asian or a Chinese name than if they had a White British name. Students with a White British name were more likely to receive only emphasis-based feedback on *punctuation* perhaps due to a perception that they are more likely to comprehend this type of feedback than their Asian or Chinese counterparts. These results indicated that although differences in marking practices were observed in the control condition for stylistic feedback related to *punctuation* these differences were accentuated further in the experimental condition. This suggests that that expectancy effects related to ethnicity have played a role in influencing the patterns of feedback provided by the participants.

Differences were also found in the further subcategories of stylistic feedback related to *syntax/ word order/grammar* (see Table 23). However, since there were also fairly large discrepancies within the control condition for many of these further subcategories it is difficult to claim with certainty that the differences observed later in the experimental condition were as a result of expectancy effects. Nonetheless, while Group 1a (WBM) and Group 2a (AM) attracted similar scores for *stylistic comments* and *corrections* related to *syntax/word order/grammar*, Group 3a (CM) barely registered any feedback for these elements. In fact, Group 3a (CM) failed to attract a single comment on their academic writing skills in this subdomain. Furthermore, they were only provided with a handful of corrections (6%) and no feedback at all on *stylistic emphasis*.

	Stylistic comment: syntax/word order/grammar	**Stylistic correction: syntax/word order/grammar**	**Stylistic emphasis: syntax/word order/grammar**
Group 1a Experimental (White British male)	7%	26%	7%
Group 2a Experimental (Asian male)	11%	21%	0%
Group 3a Experimental (Chinese male)	0%	6%	0%

Table 23: Analysis 2.1: Stylistic feedback differences in *syntax/word order/grammar* across male ethnicities

Stylistic comments are broadly identified in Hyatt's (2005) feedback classification system as comments that, '… consider the use and presentation of academic language within the assignment'. Comments, corrections and emphases specifically related to *syntax/word order/grammar* concern feedback which lets the student know that their sentence construction is either good or requires improvement. While Group 1a (WBM) and Group 2a (AM) attracted comparable amount of *stylistic comments* and *corrections*, Group 3a (CM) did not, therefore perhaps preventing this student from learning how to improve their academic writing and indicating that expectancy effects related to ethnicity have played a role here.

Stylistic comments: *referencing/citation/quotation/bibliography* attracted different amounts of feedback in the control condition as well as the experimental condition making it difficult to infer that these discrepancies amounted to anything other than variations in marking practice. However, when *stylistic corrections: referencing/citation/quotation/bibliography* are examined instead of *stylistic comments,* it is apparent that this further subcategory only resulted in a difference of 10% across groups in the control condition, whereas in the experimental condition the difference amounted to 32%. Once again, it was Group 3a (CM) which was influential in creating such a difference. Group 1a (WBM) scored 7%, Group 2a (AM) scored 1%, and Group 3a (CM) scored 33%. Students with a Chinese name are therefore much more likely to obtain feedback connected to referencing-related issues than students with either a White British or Asian name.

Further differences were observable when stylistic feedback related to *presentation* were analysed. While there were only very small differences in the control condition for *stylistic correction: presentation* and *stylistic emphasis: presentation*, there were much larger differences in the experimental condition. Specifically, of the corrective feedback provided for Group 1a (WBM), 7% related to presentational issues, for Group 2a (AM) this was 15% and for Group 3a (CM) 0%. Feedback related to *stylistic emphasis: presentation* saw even bigger differences with Group 1a (WBM) gaining 0%, Group 2a (AM) 27% and Group 3a (CM) 40%.

Overall, the White British name (Group 1a) received the least feedback on presentational issues (see Table 24), but where feedback was provided it was meaningful rather than simply being emphasis-based (where attention was drawn towards an issue without providing any explanation or correction). Essays bearing an Asian name (Group 2a) attracted both corrective and emphasis-based presentational feedback thus providing students with both an opportunity to learn from their presentational mistakes and to have them identified for them. However, essays bearing a Chinese name (Group 3a) received a large amount of emphasis-based feedback and no corrective feedback whatsoever. The Chinese name (Group 3a) also received the lowest amount of feedback comments related to *presentation* and therefore solely received more ambiguous presentational feedback (e.g., emphasis-based) and not explicit feedback about how to improve this aspect of their work. These differences also suggest that expectancy effects related to student ethnicity are in operation.

	Stylistic comment: presentation	**Stylistic correction: presentation**	**Stylistic emphasis: presentation**
Group 1a Experimental (White British male)	11%	7%	0%
Group 2a Experimental (Asian male)	17%	15%	27%
Group 3a Experimental (Chinese male)	7%	0%	40%

Table 24: Analysis 2.1: Presentational feedback differences across male ethnicities

Content-related feedback attracted the second highest number of overall feedback contributions for Groups 2a (AM) and 3a (CM), but only the third highest for Group 1a (WBM). This pattern was replicated in the control condition. While there were differences in overall content-related feedback scores within the control condition these were more pronounced in the experimental condition. Group 1a (WBM) gained the least content-related feedback at 20%, followed by Group 2a (AM) at 26% and the Group 3a (CM) at 36%. The difference in overall content-related feedback did not emanate from the *content-related comments* categories which were comparable across both groups and conditions. Neither were there noteworthy differences in the make-up of those comments in terms of positive and *negative evaluation*. The *content-related symbols* subcategory did provide some anomalies however. There was a difference of 5% between groups in the control condition with Group 1a (WBM) and Group 2a (AM) getting 3% of their total feedback in this form, whereas Group 3a (CM) got 8%. However, these anomalies were magnified in the experimental condition where Group 1a (WBM) got 6% of their feedback in this form, Group 2a (AM) gained 8% and Group 3a (CM) received 21% of their overall feedback in this way. All of the symbol-based feedback was positive, but put into context Group 1a (WBM) gained twenty ticks on their assignment, Group 2a (AM) gained twenty-three, and Group 3a (CM) received forty-eight.

Therefore the increased overall percentage scores for content-related feedback for Group 3a (CM) were almost entirely comprised of symbol-based feedback as opposed to comment-based feedback. While all of this feedback was positive (and therefore took the form of ticks on the work instead of crosses) it is worth being mindful of the previous arguments which have queried how useful such feedback is. Alternatively, since some research has identified that many students either do not engage with feedback at all or do so superficially (Weaver, 2006), scanning through an assignment and seeing it littered with ticks might act to enhance self-confidence.

Percentage scores related to overall *developmental comments* varied between groups for both the experimental and the control conditions, making it likely that these differences can simply be explained as variations in marking practice. This notwithstanding, exploration of the subcategories within *developmental comments* did reveal some differences (see Table 25).

At percentage level Group 2a (AM) and Group 3a (CM), gained more developmental feedback related to *alternatives* than Group 1a (WBM). However, since Group 1a (White gained a higher percentage of developmental feedback overall, the amount of comments they received (thirty-two) related to *alternatives* actually surpassed those of either Group 2a (AM, twenty-nine comments) or Group 3a (CM, thirteen comments). Given the earlier discussion about the negative tone of feedback found within the *alternative* subcategory it is noteworthy that the WBM name captured considerably more of this type of feedback. Nonetheless, there is also a developmental quality to this type of feedback which Hyatt (2005) claims should help students with subsequent work. Therefore it could also be argued that Group 3a (CM) have been disadvantaged as a result of receiving less of this type of feedback. Although there were also differences in the control condition when examining the amount of *alternative* comments these differences were larger in the experimental condition, pointing to evidence that expectancy effects on the basis of student ethnicity may have played a role alongside marker variability.

	Developmental comment: alternative	**Developmental comment: future**	**Developmental comment: reflective question**	**Developmental comment: informational comment**
Group 1a Experimental (White British name)	42%	0%	45%	13%
Group 2a Experimental (Asian name)	62%	0%	30%	9%
Group 3a Experimental (Chinese name)	57%	0%	35%	9%

Table 25: Analysis 2.1: Developmental feedback differences across male ethnicities

Group 1a (WBM) attracted more feedback in the *reflective questions* subcategory (45%) than either their AM (30%) or CM (35%) counterparts. Therefore WBMs received more *developmental comments* on their work in the *alternative* subcategory which might be interpreted as either positive or negatives and were also provided with more feedback aligned with stimulating thought and extending learning. However, there were similar differences apparent in the subcategory in the control condition and therefore marker variability is the most likely explanation for these.

Within group differences were also of interest here since markers in Group 2a (AM) and Group 3a (CM) did not show a predisposition to provide a lack of feedback in *reflective questions* in the control condition (when the essay bore a White British name). In fact, in this condition their *reflective questions* feedback outscored that of Group 1a (WBM). Nevertheless, when the student name changed from White British (in the control condition) to either Asian or Chinese (in the experimental condition), their propensity to provide reflective feedback dropped by up to 24% percent. Similarly, there did not appear to be a predisposition for markers in Group 3a (CM) to provide high levels of *alternative* based feedback in the control condition. Yet in the experimental condition, when marking a Chinese name as opposed to a White British name the inclination to include such comments increased by over 20%. Taken together these results point to changes in markers behaviour across conditions (e.g. markers in Group 2a and 3a provided less *reflective questions* in the experimental condition). It is possible that expectancy effects related to ethnicity were the main catalyst for the changes in feedback behaviour across conditions, since markers in Group 1a (WBM) who marked a WBM name in the control condition and again in the experimental condition marked more consistently than those markers in Groups 2a (AM) and 3a (CM) whose student name changed from a WBM to either an Asian or CM. There were no *developmental comments* related to *future* for either group.

Structural comments gained relatively similar scores across groups with no great disparity from those obtained across groups within the control condition. *Sentence level* comments prevailed over *discourse level* comments for all groups.

In summary, there were several differences evident in the stylistic feedback awarded. Asian and Chinese names were more likely to receive comments related to *punctuation* than the White British name who only received emphasis-based feedback in this domain. White British and Asian names received more comment and corrective-based feedback on syntax/word order/grammar in comparison to the Chinese name who did not attract a single comment on their writing skills in this area alongside receiving only minimal corrective and emphasis-based feedback. Conversely it was the Chinese name who attracted the most feedback linked to *referencing/citation/ quotation/bibliography* when compared to White British and Asian names. In terms of presentational feedback, the White British name gained the least, but it was more substantial in nature, since it was comment-based. Asian names attracted both corrective and emphasis-based feedback whereas the Chinese name gained the lowest number of both comments and corrections related to *presentation* instead mainly receiving emphasis-based feedback.

In terms of content-related feedback the amount of feedback was comparable in terms of the comments received and in terms of the negative or positive tone of the feedback. However, once again the CM name provided evidence of expectancy effects, this time related to *content-related symbols* where he gained more than double the amount of ticks on his assignment compared to either of the other male ethnicities.

The CM also gained fewest comments in the developmental feedback subcategory *alternatives.* The amount of comments provided to the WBM just eclipsed those received by the AM.

4.3.2 **Analysis 2.2: Female Ethnicities**

This analysis compared the in-text feedback of participants placed into Group 1b, Group 2b, or Group 3b. All participants first marked the control essay presented as being written by Samuel Jones and then an experimental essay. Group 1b's experimental essay was labelled as being written by a WBF (Natasha Brown), Group 2b's experimental essay was labelled as an AF (Avinash Puri), and Group 3b's experimental essay was labelled as a CF (Mei Lin Pang). The content of the essay did not change, only the student name.

Analysis 2.2: Ethnicity White British Female Vs. Asian Female Vs. Chinese Female Experimental			
	WBF (Group 1b)	**AF (Group 2b)**	**CF (Group 3b)**
Phatic Feedback	0%	1%	0%
Developmental Feedback	16%	22%	16%
Structural Feedback	3%	23%	4%
Stylistic Feedback	57%	46%	66%
Content-related Feedback	25%	23%	17%
Administrative Feedback	0%	0%	0%
Ambiguous Feedback	0%	2%	0%

Table 26: Analysis 2.2: In-text feedback classifications for female ethnicities (Experimental)

Total feedback contributions were different once again: Group 1b (WBF, n=219), Group 2b (AF, n=242), and Group 3b (CF, n=326). However, since there were differences evident in the control condition it is likely that these differences can be accounted for by marker variability and not expectancy effects in relation to ethnicity.

Unusually scores related to stylistic feedback only dominated the overall feedback contributions for Groups 1b (51%, WBF) and 3b (50%, CF). The score for Group 2b (AF) was 36%. Overall scores for stylistic feedback had also varied in the control condition, although the category had maintained status as the highest scoring category across all groups. In contrast to the control condition, *stylistic comments* only dominated over *stylistic corrections* and *stylistic emphasis* for Group 3b (CF). Group 1b (WBF) and Group 2b (AF) gained more corrective feedback than comment-based feedback. *Stylistic emphasis* was the lowest scoring subcategory for each group. The differences between groups in the experimental condition were comparable with those in the control condition for each subcategory therefore indicating that any differences were simply indicative of different marking practices.

However, once again when the further subcategories that made up stylistic feedback were examined some patterns emerged. For example, *stylistic comments* related to *referencing/ citation/quotation/bibliography* gained similar percentage scores across Group 1b (WBF) and Group 2b (AF) in the control condition (45% versus 48%) and this translated to a similar amount of comments. However, in the experimental condition the difference between groups increased from 3% to 34%, with Group 1b (WBF) scoring 30% and Group 2b (AF) scoring 64% which translated to over a third more comments. Given that such differences were not observable in the control condition it can be assumed that these differences were as a result of expectancy effects related to the ethnicity of the student.

Another further subcategory of stylistic feedback which demonstrated differences was *stylistic correction: punctuation*. Here scores varied between groups in both the control and experimental conditions therefore suggesting that marker variability as opposed to expectancy effects could account for such differences. However it was of interest that there was a rising profile across all groups for this type of feedback in the experimental condition. But whereas scores for Group 1b (WBF) increased by 19% and Group 3b (CF) by 27%, Group 2b (AF) showed an increase of 51%. Therefore markers in all groups perceived the experimental essay to need more corrective feedback for punctuation-related elements of the work. However, despite the essays containing identical content, when the student name was that of an AF there was a perception that many more errors of this type existed in the work.

The total scores for overall content-related feedback were 30% for Group 1b (WBF), 39% for Group 2b (AF), and 31% for Group 3b (CF) demonstrating that student work bearing an Asian name gained the most content-related feedback. This was different to Analysis 2.1 where the Chinese male received the most feedback. However, when the subcategories were examined it was clear that the balance of the feedback for Group 2b (AF) was heavily weighted in favour of *content-related symbol* feedback as opposed to *content-related comments*. The other groups both received feedback which was fairly evenly split between comments and symbols (see Table 27). These differences in *content-related symbol* feedback did not exist in the control condition and therefore suggest expectancy effects related to ethnicity have played a role here. Though interesting these findings do not match those from Analysis 2.1 where the CM received the most *content-related symbol* feedback. The merits of receiving comment-based feedback as opposed to symbol-based feedback for learning outcomes have been detailed previously.

	Content-related Comments	**Content-related Symbols**
Group 1b Experimental (White British female)	16%	14%
Group 2b Experimental (Asian female)	12%	26%
Group 3b Experimental (Chinese female)	16%	15%

Table 27: Analysis 2.2: Content-related feedback differences across female ethnicities

The positive and negative tone of the *content-related comments* and symbols were evaluated. It was noticeable that there were between group differences in both the control and experimental conditions. While the experimental essay achieved much higher percentage scores for positive comments across all ethnic groups, it is noteworthy that Group 3b (CF) were provided with 18% more positive comments than Group 1b (WBF) and Group 2b (AF), a difference which was only 8% in the control condition. This translated to Group 3b (CF) receiving thirty-one positive comments on their work in relation to fifteen for Group 1b (WBF) and thirteen for Group 2b (AF). Group 3a (CM) did also receive the highest percentage of positively oriented *content-related comments* in Analysis 2.1 although these were not considered indicative of expectancy effects at work in this instance because scores in the control group were also highly variable. As part of the same analysis Group 3a (CM) did receive the most positive *content-related symbol* based feedback in the form of ticks.

The percentages for developmental feedback scores were comparable across groups with the differences paralleling those in the control condition. However, although subcategories did not demonstrate any meaningful differences within the experimental condition, *developmental comment*: *reflective questions* did highlight an interesting trend. Despite scores for this subcategory having been variable in the control condition, markers within Group 1b had demonstrated a propensity to provide extensive feedback in this domain (57%), outscoring the other groups. However, in the experimental condition Group 1b (WBF) posted the lowest score for this type of feedback (19%). While Group 2b (AF) and Group 3b (CF) remained within 6% of their control condition percentages, Group 1b (WBF) recorded a decrease of 38%. It is unlikely that this difference can be attributed to expectancy effects related to ethnicity (given that this group marked an essay presented as being written by a White British student on both occasions), but it might instead serve to highlight how when ethnicity remains stable gender might become more visible to the marker and impacts expectancies more significantly than ethnicity. This supports the findings from Analysis 1.1 where between groups analysis also showed that WBF scored lower on reflective questions than their male counterparts. Furthermore, this contention once again lends support to Asch's (1946) concept of trait centrality with gender being the central trait in this instance. Alternatively, it simply resurrects the arguments surrounding the reliability of marking practice.

Only Group 1b (WBF) gained any feedback on *developmental comments* related to *future*, although these only amounted to 3% of the total amount of developmental feedback. The findings here did not support those found in Analysis 2.1 which demonstrated that although Group 1a (WBM) gained less developmental feedback related to *alternatives* as a percentage, this equated to them receiving more comments on the subcategory in their work.

Structural level feedback was comparable with the control condition showing minimal differences across groups and type of feedback (e.g., sentence level and discourse level).

In sum, AF names received more *stylistic comments*: *referencing*, and more *stylistic correction: punctuation* than other groups despite scoring the lowest for overall feedback contributions related to stylistic feedback. This perhaps demonstrates that if we truly want to understand feedback and the messages it provides to students then it is important to move beyond a superficial assessment of feedback types and adopt a more meticulous approach. AFs also gained the highest amount of Content-related symbol-based feedback, while Group 3b (CF) were given more *content-related comments: positive evaluation*.

4.4 In-Text feedback Analysis 3: Perceived Student Gender and Ethnicity

To avoid replication, the final analyses only compared groups which had not been compared as part of the earlier analyses. Previous gender-based analyses have kept ethnicity constant, so that for example, James Smith was compared to Natasha Brown, but not to Avinash Puri or Mei Lin Pang. Previous ethnicity-based analyses compared ethnicity but kept gender constant thus comparing James Smith with Jadgit Sidhu and Zhi Rong Liu for example, but not James Smith with either Avinash Puri or Mei Lin Pang.

The following analyses therefore compare:

- WBM and AF (James Smith vs. Avinash Puri)

- WBM and CF (James Smith vs. Mei Lin Pang)
- AM and CF (Jagjit Sidhu vs. Mei Lin Pang)
- AM and WBF (Jagjit Sidhu vs. Natasha Brown)
- CM and WBF (Zhi Rong Liu vs. Natasha Brown)
- CM and AF (Zhi Rong Liu vs. Avinash Puri)

The results of groups that have been analysed as part of the earlier analyses will be summarised where appropriate in order to provide a complete picture of the data set.

4.4.1 Analysis 3.1: White British Male vs. Asian Female

This analysis compared the in-text feedback of participants in Group 1a or Group 2b. All participants first marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 1a's experimental essay was labelled as being written by a WBM (James Smith) and Group 2b's experimental essay was labelled as being written by an AF (Avinash Puri). The content of the essay did not change, only the student name.

Analysis 3.1 Perceived Student Gender & Ethnicity White British Male Vs. Asian Female (Experimental)		
	WBM (Group 1a)	**AF (Group 2b)**
Phatic Feedback	0%	1%
Developmental Feedback	22%	22%
Structural Feedback	5%	3%
Stylistic Feedback	53%	36%
Content-related Feedback	20%	39%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 28: Analysis 3.1: In-text feedback classifications for WBM and AF (Experimental)

Total feedback contributions were vastly different with Group 1a (WBM) scoring 343, and Group 2b (AF) scoring 242. These differences were replicated in the control condition. Stylistic feedback was the dominant feedback type for both groups and resulted in similar between-group differences as in the control condition. Examination of the subcategories for stylistic feedback illustrated that although there were differences between the percentages assigned to stylistic comments, corrections and emphasis these differences were comparable with the differences in the control condition. Therefore these differences can only be attributed to differences in marking practice and not to expectancy effects related to the interactive effects of gender and ethnicity.

Nonetheless, a more in-depth exploration of the further subcategories of stylistic feedback revealed some differences in the domain of punctuation. While there were small between-group differences with regard to *stylistic correction: punctuation* and *stylistic emphasis: punctuation* in the control condition, markers in Group 2b (AF) did not show a propensity to correct punctuation at this stage. In fact they were more than 50% less likely to provide corrective feedback specifically on *punctuation* than markers in Group 1a (WBM). However, in the experimental condition, when marking a female Asian name, Group 2b (AF) were far more likely to provide corrective feedback on punctuation than their opposing group, with their score rising from 21% in the control condition to 72% in the experimental. Whether or not the second essay required more corrective feedback on punctuation is unclear, although scores in this domain for Group 1a (WBM) remained constant across conditions indicating that perhaps this was not the case. Nonetheless it is difficult to know whether these variations reflect expectancy effects at work or whether marking practice within Group 2b (AF) was simply highly variable.

Another difference was visible in stylistic feedback related to the *referencing/citation/quotation /bibliography* category. Once again differences did exist in the control condition, therefore making claims of expectancy effects in the experimental condition difficult. However, although percentage scores for *stylistic comments* and *stylistic corrections* related to referencing were comparable between essays, Group 2b (AF) received far more *stylistic emphasis-based* feedback on this than their WBM counterparts (see Table 29). Nonetheless, it is also important to note that because Group 1a (WBM) received a greater percentage of stylistic feedback in comparison to Group 2b (AF) their 21% share of feedback in this domain consisted of three comments. Group 2b (AF) received a much higher percentage at 67% but this was only made up of four comments therefore suggesting that there would be very little difference in the feedback students received.

	White British Male (Group 1a)		**AF (Group 2b)**	
Referencing/citation/ quotation/bibliography	Control	Experimental	Control	Experimental
Stylistic Comments	34%	58%	48%	64%
Stylistic Corrections	12%	7%	44%	4%
Stylistic Emphasis	19%	21%	13%	67%

Table 29: Analysis 3.1: Stylistic feedback differences between WBM and AF across conditions

Markers in Group 2b (AF) had not shown a tendency to provide *stylistic emphasis-based* feedback in the control essay. Their inclination to highlight such errors increased when the essay was thought to have been written by a female Asian student. Of course it is possible that the quality of the second essay justified more feedback in this area, but no comparable percentage increase was seen across conditions for Group 1a (WBM).

Previous analyses in this chapter have demonstrated that essays bearing an AF name attracted more *stylistic corrections*: *punctuation* specifically in relation to Group 1b (WBF) and Group 3b (CF). This trend continued with Group 2b (AF) also gaining more feedback on this than Group 1a (WBM). Previous analyses also revealed that AFs received more *stylistic comments* related to *referencing/citation/quotation/bibliography* than other female groups. While no differences in that area were found here Group 2b (AF) did receive more *stylistic emphasis-based* feedback on referencing-related issues than Group 1a (WBMs).

AF students (Group 2b) were also provided with almost 20% more content-related feedback than WBMs (Group 1a). Differences within the control condition for content-related feedback amounted to 10%. Exploration of further subcategories revealed that although Group 1a (WBM) and Group 2b (AF) received similar amounts of *content-related comment* feedback (14% versus 12%) AFs were provided with significantly higher amounts of *content-related symbol* feedback (26% versus 6%). These percentages translated to Group 2b (AF) having received sixty-four ticks on their work while Group 1a (WBM) received twenty. Since similar differences were not present in the control condition (where differences in scores for *content-related symbol* feedback totalled 5% and only 8 ticks) these differences can be attributed to expectancy effects on the basis of student gender and ethnicity.

The same pattern was evident in Analysis 2.2 where the balance of content-related feedback for (Group 2b) AFs heavily favoured symbol-based feedback over comments. In comparison Group 1b (WBF) and Group 3b (CF) received a balance of comment and symbol-based feedback. Another group that has received similar feedback in terms of *content-related symbols* over comments is Group 3a (CM). Analysis 2.1 demonstrated that work perceived to have been written by a CM gained more content-related feedback than their WBM or AM counterparts, but the differences were largely symbol related.

Developmental feedback scored 22% for both groups. Subcategories of *reflective questions* and *alternatives* dominated this category. *Informational comments* scored ≤ 3% for each group and there were no *developmental comments: future* for either group.

Structural feedback attracted similar percentages (Group 1a = 5% and Group 2b = 3%) and were dominated by *sentence level* as opposed to *discourse level* comments.

4.4.2 **Analysis 3.2: White British Male vs. Chinese Female**

This analysis compared the in-text feedback of participants in Group 1a or Group 3b. All participants marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 1a's experimental essay was labelled as being written by a WBM student (James Smith) and Group 3b's experimental essay was labelled as being written by a CF (Mei Lin Pang). The content of the essay did not change, only the student name.

Analysis 3.2 Perceived Student Gender & Ethnicity White British Male Vs. Chinese Female (Experimental)		
	WBM (Group 1a)	**CF (Group 3b)**
Phatic Feedback	0%	0%
Developmental Feedback	22%	16%
Structural Feedback	5%	3%
Stylistic Feedback	53%	50%
Content-related Feedback	20%	31%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 30: Analysis 3.2: In-text feedback classifications for WBM and CF (Experimental)

The total number of feedback contributions across groups was only marginally different; Group 1a (n=343) and Group 3b (n=326). As was the case in the control condition, stylistic feedback was most dominant scoring 53% for Group 1a (WBM), and 50% for Group 3b (CF). Subcategories for stylistic feedback showed that *stylistic comments* dominated, followed by *stylistic corrections* and *stylistic emphasis*. Although there were differences in the percentages awarded to each category, these differences were similar to and in the same direction as differences identified in the control condition.

Nonetheless, a more detailed examination of the further subcategories revealed some differences. Percentages related to the further subcategories within *stylistic comments* were comparable, but there were differences evident within *stylistic corrections*. Specifically, *stylistic corrections: punctuation* amounted to 63% of the corrective feedback for Group 3b (CF), but to only 46% of the corrective feedback for Group 1a (WBM). There were differences evident in this type of feedback in the control condition too, although these were smaller than those in the experimental condition. Nonetheless, this does make it difficult to claim that the changes in the experimental condition were as a result expectancy effects related to student gender and ethnicity. Markers in Group 3b (CF) did not show an inclination to correct *punctuation* mistakes in the control condition when they perceived themselves to be marking the work of a WBM. In fact they provided 9% less corrective feedback than Group 1a. However, in the experimental condition, when under the impression that they were marking the work of a CF Group 3b increased their feedback in this domain by 27%.

However, Group 1a (WBM) did receive feedback in the area of *punctuation*, but this was in the form of emphasis-based entries (such as the circling of a mistake) rather than a correction. Indeed Group 1a (WBM) received 40% more *stylistic emphasis-based* feedback on their *punctuation* than Group 3b (CF), a difference that was far larger than that of the control condition (11%). The perceptions that might underpin these different practices are unknown. They may relate to a perceived need to provide more substantial feedback to CFs because they are less familiar with the requirements of academic writing. Similarly, they may only provide emphasis-based feedback to WBMs because they believe this group should be able to understand and interpret what the marker means more proficiently. Or, it might be that CFs are considered more industrious than WBMs and will therefore take the time to digest more comprehensive feedback in order to improve. Nevertheless, corrective feedback is generally considered more useful than emphasis-based feedback since the level of ambiguity is reduced and therefore CFs seem to have been advantaged in relation to WBMs in this domain.

Group 3b (CF) were also provided with more corrective feedback in relation to *referencing/ citation/quotation/bibliography* (17% versus 7%), a difference that was not evident in the control condition. However, Group 1a (WBM) were provided with 16% more corrective feedback than Group 3b (CF) in the domain of *syntax/word order/grammar.* These scores were almost identical between groups in the control condition. Such inconsistent results may therefore reflect the lack of reliability in marking processes rather than expectancy effects at work.

In terms of content-related feedback, Group 3b (CF) gained 11% more than Group 1a (WBM). However Group 3b (CF) were also prone to providing more of this type of feedback in the control condition, outscoring Group 1a (WBM) by 4% in this instance. Nonetheless, there was a bigger difference in the experimental condition. Moreover, although scores were similar between groups for *content-related comments* (Group 1a = 14%, Group 3b = 16%) CFs gained more symbol-related feedback on their work (Group 1a = 6%, Group 3b = 15%). The symbol-related feedback was 100% positive for both groups. Although there were differences between groups for this domain in the control condition too, these were much smaller at 4%. Consequently this finding echoes that of analysis 3.1 and 2.1 whereby a cursory glance at the data revealed that AFs and CMs received more content-related feedback. However, further examination revealed that the increased amount of feedback received was only composed of symbols and therefore lacked credibility as a form of feedback that would fulfil an educative function.

Percentage scores for developmental feedback were identical in the control condition, but varied by 6% in the experimental condition in favour of Group 1a (WBM). Scores for the further subcategories were comparable, with the exception of *developmental comment: information* where Group 1a (WBM) scored 13% and Group 3b (CF) scored 4%. There was only a 3% difference in the same direction in the control condition. Informational comments are intended to offer "… the student additional academic insight into the topic under discussion" (Hyatt, 2005). This insight is provided with the intention of "… aiding the student with subsequent work in relation to the current assignment". While a between group difference in the experimental condition of 9% may not seem huge, when it is considered that Group 1a (WBM) was provided with ten such comments on their assignment while Group 3b (CF) was provided with two it is clear that WBMs were given more opportunities to develop their academic insight and submit stronger future assignments.

Structural comments attracted similar percentages across groups (Group 1a = 5% and Group 3b = 3%). Sentence Level comments prevailed for Group 1a (WBM) whereas Discourse Level comments prevailed for Group 3b (CF). These patterns were replicated in the control condition.

4.4.3 **Analysis 3.3: Asian Male vs. Chinese Female**

This analysis compared the in-text feedback of participants placed into Group 2a or Group 3b. All participants first marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 2a's experimental essay was labelled as being written by an AM (Jagjit Sidhu) and Group 3b's experimental essay was labelled as being written by a CF (Mei Lin Pang). The content of the essay did not change, only the student name.

Analysis 3.3 Perceived Student Gender & Ethnicity Asian Male Vs. Chinese Female (Experimental)		
	AM (Group 2a)	**CF (Group 3b)**
Phatic Feedback	0%	0%
Developmental Feedback	15%	16%
Structural Feedback	3%	3%
Stylistic Feedback	56%	50%
Content-related Feedback	26%	31%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 31: Analysis 3.3: In-text feedback classifications for AM and CF (Experimental)

Overall feedback contributions for the experimental group were very similar; Group 2a (n = 322) and Group 3b (n=326). The weighting of feedback contributions across categories mirrored those of the control condition with stylistic feedback being followed by content-related feedback, developmental feedback and finally structural feedback.

Stylistic feedback scores were similar and accounted for 56% of total feedback for Group 2a (AM) and 50% for Group 3b (CF). Both groups also provided a consistent amount of feedback in this category in the control condition therefore indicating similar marking practices and a lack of expectancy effects at work. Markers scores for this category remained within 6% of each other in the experimental condition, markers in Group 3b (CF) who had shown a propensity to provide higher levels of stylistic feedback when marking a male White British name (outscoring markers in Group 2a) provided 16% less feedback in this domain when they perceived themselves to be marking a CFs work (and were outscored by Group 2a). There was also a reduction in stylistic feedback scores of 5% for Group 2a (AM). So the markers feedback was consistent in the control condition and the experimental condition for this category but they still provided less feedback on stylistic elements of the work when it was considered to have been written by either an AM or a CF.

Examination of the subheadings for stylistic feedback revealed that they followed the pattern of the control condition in that *stylistic comments* outweighed *stylistic corrections* and *stylistic emphasis*. However, interrogation of the data did reveal some interesting differences at further subcategory level. Specifically these related to *stylistic corrections: punctuation* where Group 2a (AM) scored 50% and Group 3b (CF) scored 63% reflecting a 13% difference between groups. Although this percentage difference did not translate to a huge difference in the number of corrections provided on the work there was no difference at all in percentages or corrections in the control condition. Furthermore, when *stylistic emphasis: punctuation* was considered, Group 2a (AM) scored 18% and Group 3b scored 31%. As well as there being only a 3% difference between groups in the control condition the percentage difference in the experimental condition equated to Group 2a (AM) receiving one emphasis-based piece of feedback on their work whereas Group 3b (CF) gained ten. Therefore it seems likely that expectancy effects related to student gender and ethnicity may have played a role here with the CF student gaining more corrective and emphasis-based feedback on *punctuation* compared to the AM despite the essays being identical. This finding partly supports findings in Analysis 2.1 where CMs were also provided with more corrective feedback on *punctuation* than WBMs and AMs. However Analysis 2.2 found that although Group 3b (CF) gained more corrective feedback related to *punctuation* than Group 1b (WBF) it was Group 2b (AF) who dominated this category.

Exploration of the further subcategories related to *syntax/word order/grammar* revealed differences in feedback too. Specifically, Group 2a (AM) received 11% more corrective feedback than Group 3b (CF). However this difference was also present in the control condition and as such can be attributed to differences in marking practices. Nonetheless, it is noteworthy to consider within group feedback differences. Markers in both groups showed an inclination to correct mistakes related to *syntax/word order/grammar* when marking the control essay (perceived to have been written by a WBM) and provided Group 2a (AM) with 49% of their stylistic feedback in this domain and Group 3b (CF) with 39%. Yet when marking essays presumed to have been written either by an AM or CF in the experimental condition this inclination significantly reduced (Group 2a = 21% and Group 3b = 10%).

Placed into context, in the control essay Group 2a received 29 corrections on *syntax/word order/grammar* and Group 3b received 28. In the experimental conditions Group 2a (AM) received 17 corrections and Group 3b (CF) received only 6. Furthermore, there was also a reduction on emphasis-based feedback for *syntax/word order/grammar* in the experimental condition with control scores of 17% (Group 2a) and 18% (Group 3b) decreasing to 0% and 6% respectively. Of course, it is possible that the experimental essay simply contained fewer *syntax/word order/grammatical* related errors than the control essay, and that the feedback reflects reality rather than expectancy effects at work, but nonetheless it is interesting that markers feedback styles seemed to change significantly when either the gender or ethnicity or both the gender and ethnicity of the student changed from that of a WBM. Since differences in feedback practice were apparent in both the control and experimental conditions, it cannot be claimed with certainty that expectancy effects are in operation because both sets of markers marked differently from each other across both conditions, but the direction of these differences (e.g. providing more or less feedback in specific areas) did match when they marked work written by a non-WBM.

This lack of *stylistic correction* feedback on *syntax/word order/grammar* for Group 3b (CF) supports the findings from Analysis 2.1 where Group 3a (CMs) were found to receive less feedback than AMs and WBMs in this domain (in fact this finding extended to CMs gaining less feedback in terms of comments, corrections and emphasis-based feedback)

Similarly, when exploring the further subcategories for *presentation* it is clear that Group 2a (AM) received more feedback than Group 3b (CF) across all aspects; comments, corrections and emphasis. Markers in Group 2a (AM) did show an inclination to provide more feedback in these areas in the control condition too, but the increased differences seen in the experimental condition might also indicate expectancy effects in operation. This finding largely supports the results of Analysis 2.1 which also showed that Group 3a (CM) gained the lowest number of comment and corrective feedback on presentational issues compared to their WBM and AM counterparts.

Content-related feedback gained the second highest number of feedback contributions for both groups, scoring 26% for Group 2a (AM) and 31% for Group 3b (CF). Similar percentage differences were observed across groups in the control condition. Although *content-related comments* dominated over *content-related symbols* for Group 2a (AM) there was an almost equal split between comments (16%) and symbols (15%) for Group 3b (CF). These patterns were replicated in the control condition. Feedback pertaining to Positive Evaluative Comments and Symbols and Negative Evaluative Comments and Symbols were similar across groups in the experimental and control conditions suggesting no indication of expectancy effects at work in this instance. These findings do not replicate those of Analysis 2.2 where Group 3b (CF) were provided with more *content-related comments: positive evaluation* than their WBF and AF counterparts.

Developmental feedback scores represented as a total of overall feedback contributions were almost identical across groups (Group 2a = 15% and Group 3b 16%). The same pattern was evident in the control condition. The majority of *developmental comments* related to *alternatives*, then *reflective questions,* and finally *informational comments*. There were no *developmental comments* related to *future* for either group. Many differences in scores across subcategories were also reflected in the control condition and are therefore likely to be as a result of differences in marking practices between groups. However, the subcategory *developmental comments: reflective questions* did reveal that Group 2a (AM) received 14% less feedback in this domain in comparison to Group 3b (CFs). This translated to Group 2a (AM) receiving fourteen r*eflective questions* on their work whereas Group 3b (CF) received twenty-three. These differences were not apparent in the control condition which only yielded a difference of 3% and four comments. Therefore such differences can be attributed to expectancy effects stimulated by the interactive effects of gender and ethnicity. These results support Analysis 2.1 where Group 2a (AM) were also seen to receive the lowest amount of feedback in this area compared to their WBM and CM counterparts. Nonetheless, it is also evident that AMs received more *developmental comments* related to *alternatives* than CFs which may be argued to mitigate the need for them to receive alternative types of developmental feedback too. However, these percentage differences only translated to AMs receiving two additional comments on their work and since feedback related to *alternatives* attracted different scores in the control condition, the difference here is likely to be a result of variance in marking practices.

Structural feedback attracted the same score for both groups (3%) and these scores were comparable with those in the control condition. However, when *structural feedback: discourse* level and *structural feedback: sentence level* were examined it was apparent that these only resulted in a difference of 12% across groups in the control condition, whereas in the experimental condition the difference amounted to 35%. Specifically, Group 3b (CF) gained 35% more comments on structure at discourse level than Group 2a (AM), whereas Group 2a (AM) gained 35% more comments on structure at sentence level than Group 3b (CF). Such inconsistent results make it difficult to make judgements pertaining to the presence of expectancy effects beyond noting that markers were closer in the feedback provided when they considered they were marking a WBM student's work, than when they thought they were marking either an AM or a CF.

4.4.4 **Analysis 3.4: Asian Male vs. White British Female**

This analysis compared the in-text feedback of participants in Group 2a and Group 1b. All participants first marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 2a's experimental essay was labelled as being written by an AM (Jagjit Sidhu) and Group 1b's experimental essay was labelled as being written by a WBF (Natasha Brown). The content of the essay did not change, only the student name.

Analysis 3.4 Perceived Student Gender & Ethnicity Asian Male Vs. White British Female (Experimental)		
	AM (Group 2a)	**WBF (Group 1b)**
Phatic Feedback	0%	0%
Developmental Feedback	15%	16%
Structural Feedback	3%	3%
Stylistic Feedback	56%	51%
Content-related Feedback	26%	30%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 32: Analysis 3.4: In-text feedback classifications for AM and WBF (Experimental)

Total feedback contributions generated different numbers for each group; Group 2a (n=322) and Group 1b (n=219). As with the control condition stylistic feedback outranked all other feedback categories for both groups with the percentage differences also echoing those in the control condition (Group 2a = 56%, Group 1b = 51%). Group 2a (AM) mirrored the rankings identified in the control condition for the subheadings of stylistic feedback (i.e. *stylistic comments* dominated over *stylistic corrections* which dominated over *stylistic emphasis*). However, markers in Group 1b

(WBF) provided more corrections than comments on stylistic elements of the work, while *stylistic emphasis* remained the lowest scoring.

There were also some interesting differences recorded at further subcategory level for stylistic feedback (see Table 32). For example, Group 1b (WBF) were provided with more comment-based, corrective, and emphasis-based feedback than their Group 2a (AM) on *syntax/word order/grammar*. Differences in the amount of feedback provided across these further subcategories were also evident in the control condition however, perhaps illustrating that marker variability as opposed to expectancy effects were responsible for these.

	Asian Male (Group 2a)		**White British Female (Group 1b)**	
Syntax/word order/grammar	Control	Experimental	Control	Experimental
Comments	36	11	16	17
Corrections	49	21	41	30
Emphasis	17	0	39	36

Table 33: Analysis 3.4: Syntax/word order/grammar feedback differences between AMs and WBFs across conditions

Markers in Group 2a (AM) showed a propensity towards providing *syntax/word order/grammar* based feedback in the control condition when they perceived themselves to be marking a WBM students work, but this inclination dissipated when they perceived themselves to be marking the work of an AM. Put into context Group 2a (AM) made thirty-one comments, twenty-nine corrections and five emphases in this domain when marking what was labelled as a WBM's work in the control condition and only ten comments, seventeen corrections and zero emphases when marking the essay labelled as being written by an AM in the experimental condition. Of course, it is feasible that the experimental essay might simply have contained fewer *syntax/word order/grammar* errors than the control essay, although feedback for Group 1b (WBF) remained relatively consistent across conditions indicating that this was not the case. Analysis 2.1 found that AMs received comparable amounts of feedback in this area to WBMs but that CMs received less feedback. There were no differences reported for these further subcategories in Analysis 2.2 which compared female ethnicities.

Content-related feedback received the second highest number of feedback contributions for both groups. This was also true for the control condition. While there were differences in the scores for content-related feedback these differences were comparable with those seen in the control condition. There were also no significant differences between groups at subcategory level for this type of feedback, with comment and symbol based comments generating comparable differences with the control condition. Observation of the further subcategories for comments related to either *positive evaluation* or *negative evaluation* only demonstrated differences comparable with the control condition therefore suggesting the marker variability was the most likely cause for these. However, once again it is curious that positive evaluative comments increased in the experimental condition for both groups by a substantial margin. Group 2a (AM) increased from 25% positive comments in the control condition to 58% in the experimental and Group 1b (WBF) increased from 9% in the control condition to 43% in the experimental. Placed in context this meant that Group 2a (AM) gained 15 positive comments on their assignment in the control condition, but 35 such comments in the experimental. Similarly Group 1b (WBF) gained 4 positive comments in the control condition and 15 in the experimental. As has been suggested previously it might be that the control essay simply warranted fewer positive comments than the experimental essay. In order to crosscheck this, trends within all other groups were examined and it was found that irrespective of the gender and ethnicity of the student the experimental essay always gained much higher percentages on *content-related comment: positive evaluation* feedback. Therefore it appears as though it was the academic quality of the control essay and not the perceived gender and ethnicity of the author that generated lower scores in this instance. This is despite both essays being judged to be of lower second class standard.

Developmental feedback generated similar percentage scores across groups (Group 2a = 15%, Group 1b = 16%). These scores were comparable with those in the control condition, therefore suggesting expectancy effects had not played a role. This notwithstanding, there were some differences when the subcategories for *developmental comments* were examined. Specifically, Group 1b (WBF) gained 11% fewer comments on *reflective questions* than Group 2a (AM) and only half the amount of comments. This was more surprising because markers in Group 1b had shown a disposition towards providing this type of feedback in the control condition providing 57% of their developmental feedback in this way, versus providing only 19% in the same way in the experimental condition. Given that markers in this group were marking essays with names implying the same ethnicity (White British) it would appear that any differences here came about as a result of the gender of the student. This contention is supported by the results from Analysis 1.1 (male versus female White British) which found that Group 1a (male name) did receive more than double the amount of feedback on *reflective questions* than Group 1b (WBF). Reflective feedback is claimed to stimulate critical thinking and learning which '… affects future action' (Ghaye, Danai, Cuthbert & Dennis, 1996, p.2) and therefore might house some of the learning benefits of feedforward.

Another point of note is that both Group 2a (AM) and Group 1b (WBF) were provided with fewer *reflective questions* than either group in the control condition. Although the differences between groups in each condition were comparable, therefore suggesting a consistency to the marking process, there is also a clear trend towards providing fewer reflective comments when the student is perceived to be either an AM or a WBF as opposed to a WBM. In order to check whether the experimental essays might simply have warranted fewer *reflective questions* results from all groups were studied. With the exception of Group 3a (CM) all experimental groups scored within a few percent of each other for this type of feedback, therefore suggesting that expectancy effects related to the gender and ethnicity of the student have played a role in the type of feedback they are awarded.

In contrast Group 1b (WBF) gained 10% more developmental feedback for the *informational comments* further subcategory, than their Group 2a (AM) counterparts. This difference was not apparent in the control condition and therefore can be attributed to expectancy effects on the basis of gender and ethnicity. White British students of both genders appear to have gained more *informational comments* on their work than some other groups since Group 1a (WBM) also received more *developmental comments* of this nature in Analysis 2.1 when compared to Group 3a (CM). Informational comments are defined by Hyatt (2005) as providing a direct comment on a related issue in order to stimulate additional academic insight on a topic. As such, it appears that although *informational comments* may also stimulate critical thinking they have a more autocratic tone than some other forms of developmental feedback and perhaps only allow students to reflect on the topic at hand as opposed to encouraging broader reflection. WBFs were therefore provided with a more autocratic and less reflective type of feedback than their AM counterparts.

Finally, for *developmental comments* there was only 1 comment related to *future* and this was provided to Group 1b (WBF) and amounted to 3% of feedback in the developmental category.

Structural comments received identical scores across groups and *sentence level* comments prevailed over *discourse level* comments for both groups. There were differences in the amount of sentence versus discourse level comments provided but these differences were also present in the control condition

Group 2a (AM) gained 13% less feedback at *discourse level* and 13% more at *sentence level* compared to Group 1b (WBF). This pattern was replicated when the control condition for Group 2a (AMs) was observed too. Given that comparable changes were not seen across conditions for

Group 1b (WBF) it appears that perhaps it is the ethnicity of the student that has instigated these feedback changes as opposed to the gender (since the control essay was also written by a student with a White British name, but they were a male student).

4.4.5 **Analysis 3.5: Chinese Male vs. White British Female**

This analysis compared the in-text feedback of participants in Group 3a and Group 1b. All participants marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 3a's experimental essay was labelled as being written by a CM (Zhi Rong Liu) and Group 1b's experimental essay was labelled written by a WBF (Natasha Brown). The content of the essay did not change, only the student name.

Analysis 3.5 Perceived Student Gender & Ethnicity Chinese Male Vs. White British Female (Experimental)		
	CM (Group 3a)	**WBF (Group 1b)**
Phatic Feedback	0%	0%
Developmental Feedback	10%	16%
Structural Feedback	1%	3%
Stylistic Feedback	52%	51%
Content-related Feedback	37%	30%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 34: Analysis 3.5: In-text feedback classifications for CM and WBF (Experimental)

Total feedback contributions were different across groups; Group 3a (n=233) and Group 1b (n=219). Stylistic feedback contributions far outweighed all other feedback categories totalling 53% for Group 3a (CM) and 51% for Group 1b (WBF), a finding which echoed the control condition. Group 3a (CM) mirrored the rankings identified in the control condition for stylistic feedback subheadings (i.e. *stylistic comments* dominated over *stylistic corrections* which dominated over *stylistic emphasis*). However, more corrections than comments were provided on the stylistic elements of the work supposedly written by a WBF (Group 1b) while *stylistic emphasis* remained the lowest scoring. Nonetheless, despite this different ranking percentage scores across subcategory level did remain fairly constant between groups.

When the further subcategories of stylistic feedback were scrutinised it became evident that there were some interesting differences to explore. For example, Group 3a (CM) received the most feedback on *punctuation* across all categories (*comments, corrections and emphasis*) than Group 1b (WBF). Emphasis-based feedback saw the biggest discrepancy between groups with Group 3a (CM) received 33% more feedback in this domain. Given that the between group differences in the control condition amounted to only 4% it can be assumed that these differences were as a result of expectancy effects related to student gender and ethnicity and not a result of differences in marking practices. When the results of this analysis are compared alongside Analysis 2.1 it is evident that CMs received more feedback on punctuation-related elements of their work than their White British and AM counterparts as opposed to more meaningful elements of their work which enhance metacognitive skills. The present analyses has demonstrated that this remains the case when the work of a CM was compared to a WBF student.

A further subcategory of stylistic feedback which also captured attention related to *syntax/word order/grammar*. While comment-based and corrective feedback related to this domain was comparable between groups within the control condition, there were much larger differences in the experimental condition suggesting that these differences can be attributed to expectancy effects on the basis of student gender and ethnicity. Group 3a (CM) gained far less feedback in each of these areas than their Group 1b (WBF) counterparts (see Table 35)

	Stylistic comment: syntax/word order/grammar	**Stylistic correction: syntax/word order/grammar**	**Stylistic emphasis: syntax/word order/grammar**
Group 3a: Chinese male: Experimental	0	6	0
Group 1b: White British female: Experimental	17	30	36

Table 35: Analysis 3.5 – Syntax/word order/grammar feedback differences between CMs and WBFs

These percentages translated to Group 1b (WBF) having gained 8 comments and 15 corrections on elements of their writing style, whereas Group 3a (CM) gained zero comments and three corrections. These findings echoed those of Analysis 2.1 where CMs (Group 3a) also gained far less feedback in these domain than Group 1a (WBM) and Group 2a (AM). There were no differences reported in Analysis 1.3 which compared Group 3a (CM) with Group 3b (CFs). There were also vast differences with relation to the emphasis-based feedback each group provided, but since these differences were also present in the control condition they can only be attributed to different marking practices at work across groups.

The *referencing/citation/quotation/bibliography* elements of feedback also provided some interesting patterns at further subcategory level. Specifically, Group 3a (CM) were provided with more comment-based (60% versus 30%) and corrective feedback (33% versus 6%) than Group 1b (WBF). While there were some between group differences in the control condition these only amounted to 11% for comment-based and 3% for corrective feedback whereas the differences in the experimental condition were 29% and 27% respectively. This finding demonstrates that although there might be an element of marker variability at work, greater differences existed when the student name changed and therefore expectancy effects are likely to have played the more dominant role. This finding supports those of Analysis 2.1 which demonstrated that students with a Chinese name were more likely to obtain corrective feedback connected to referencing-related issues than students with either a White British or Asian name. Interestingly Analysis 1.3 also provided evidence of the same pattern when a CM name (Group 3a) was compared to a CF name (Group 3b).

Content-related feedback received the second highest number of feedback contributions across both groups. This was also the case for the control condition. While Group 1b (WBF) received more *content-related comments* than Group 3a (CM) this trend was reversed for *content-related symbols*. While the differences between groups were smaller in the control condition than in the experimental condition the changes were still relatively small and therefore are likely to reflect variations in marking practice as opposed to expectancy effects. Nonetheless, it is useful to note that markers in Group 3a halved the amount of *content-related comments* they provided from the control to the experimental condition (when moving from marking the work of a WBM to a CM), and almost tripled the amount of *content-related symbols* they provided (forty-eight ticks compared to twenty-two). It is possible that this might be explained by markers investing more effort into providing comment-based feedback for WBM students because of an expectancy that they might understand the feedback better and be more likely to act on it.

A closer look at the make-up of the *content-related comments* demonstrated that although Group 1b (WBF) received more of these comments it was also true that a larger amount of these comments constituted a *negative evaluation* of the work (57%) as compared to Group 3a (CM) at 32%. However, since markers in Group 1b also showed a propensity to provide more negative feedback in the control condition this is more likely to reflect a predisposition of these markers to provide such feedback irrespective of the name on the assignment.

Scores for developmental feedback were as follows; Group 3a = 8%, Group 1b =15%. Similar differences were observed in the control condition therefore suggesting expectancy effects had not played a role. Nonetheless, there were subcategories which generated different scores across groups. For example, Group 3a (CM) gained 35% of their developmental feedback in the form of *reflective questions* whereas Group 1b (WBF) only gained 19% of feedback in this domain. Scores in the control condition were only two percent apart, therefore indicating that this difference is a result of expectancy effects related to students' gender and ethnicity and not variations in marking practice.

Interestingly, Group 1b (WBF) were also found to gain the lowest amount of feedback in terms of *reflective questions* in Analysis 1.1 versus Group 1a (WBM) and in Analysis 2.2 in comparison to their Group 2b (AF) and Group 3b (CF) counterparts. Importantly however, when the amount of comments were cross checked with percentages it became evident that despite Group 3a (CM) gaining 35% of their feedback in this form this only amounted to them being asked eight *reflective questions*. This was only one more than Group 1b (WBF) were asked despite this form of feedback only making up sixteen percent less of their total developmental feedback.

In a similar vein, percentage scores looked almost identical across the *developmental comment: alternative* category for both groups and yet in practice CMs only received 13 comments offering them an alternative compared to WBFs being offered 21 such comments. A similar pattern was observed in Analysis 2.2 where Group 1a (WBM) received a lower percentage score in the *developmental comment: alternative* subcategory than Group 2a (AM) and Group 3a (CM) but, similar to the Group 1b (WBF) outcome above, this actually translated to them receiving more comments on their work with Group 3a (CM) once more receiving the lowest amount of this type of feedback in terms of comments. There were no reported differences in Analysis 1.3 which compared Group 3a (CM) with Group 3b (CF).

There were also disparities between the amount of feedback provided in terms of *informational comments* (Group 3a = 9% and Group 1b = 19). This difference was also evident in the number of comments provided for each group (Group 3a = 2 and Group 1b = 7). These differences were not apparent in the control condition. The provision of more informational feedback for WBFs was also noted in analysis 3.4 (AM vs. WBF). Group 1a (WBMs) have also been provided with more of this type of feedback than both Group 3a (CM) and Group 3b (CFs) as was shown in Analysis 2.2 and Analysis 3.2 thus suggesting that White British names generate more of this type of feedback generally. Finally, *developmental comments*: *future* only generated one comment across groups. This was provided to Group 1b (WBF) and amounted to 3% of feedback in the developmental category.

Structural comments were similar across groups (Group 3a = 1% and Group 1b = 3%). There were no differences reported in the control condition. *Sentence level* comments dominated over *discourse level* comments for both groups which was also true for the control condition.

4.4.6 **Analysis 3.6: Chinese Male vs. Asian Female**

The final analysis compared the in-text feedback of participants in Group 3a and Group 2b. All participants marked the control essay presented as being written by Samuel Jones and then the experimental essay. Group 3a's experimental essay was labelled as being written by a CM (Zhi Rong Liu) and Group 2b's experimental essay was labelled as being written by an AF (Avinash Puri). The content of the essay did not change, only the student name.

Analysis 3.6 Perceived Student Gender & Ethnicity Chinese Male Vs. Asian Female (Experimental)		
	CM (Group 3a)	**AF (Group 2b)**
Phatic Feedback	0%	1%
Developmental Feedback	10%	22%
Structural Feedback	1%	3%
Stylistic Feedback	52%	36%
Content-related Feedback	37%	39%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Table 36: Analysis 3.6: In-text feedback classifications for CM and AF (Experimental)

Total feedback contributions were marginally different across marking groups; Group 3a (n=233) and Group 2b (n=242). Stylistic feedback prevailed for Group 3a (CM) and comprised 52% of total feedback contributions. However, this category was the second most popular type of feedback for Group 2b (AF) who scored 36%. Interestingly, Group 2b (AF) were the only group throughout all analyses for whom Stylistic feedback failed to dominate. It is also true however that even in the control essay (while supposedly marking a WBM student's work) participants in Group 2b showed a weaker propensity than all other groups to provide this type of feedback (although it still remained the most dominant category in the control condition).

Examination of the subcategories for stylistic feedback also revealed differences in rankings. While Group 3a (CM) followed the common pattern (and the pattern of the control condition) where *stylistic comments* outscored *stylistic corrections* which outscored *stylistic emphasis*, Group 2b (AF) scored highest for *stylistic corrections* followed by *stylistic comments* and then *stylistic emphasis*. Nonetheless, it is important to note that there was only a small difference between the number of *stylistic comments* versus *stylistic corrections* for Group 3a (CM) too.

	Overall stylistic feedback	**Stylistic comment**	**Stylistic correction**	**Stylistic emphasis**
Group 3a Experimental (Chinese male)	52%	25%	21%	6%
Group 2b Experimental (Asian female)	36%	14%	19%	3%

Table 37: Analysis 3.6 -Stylistic feedback differences between CM and AF

What perhaps is more telling is how those percentages translated to the narrative of feedback students were provided with. For example, both groups gained similar amounts of corrective feedback on the stylistic elements of their assignment. This amounted to forty-eight corrections for Group 3a (CM) and forty-seven for Group 2b (AF). However, the patterns were quite different for *stylistic comments*. In this instance Group 3a (CM) gained fifty-nine comments on their work and Group 2b (AF) only received thirty-three. As has previously been identified, comment-based feedback is most likely to provide students with an opportunity to understand the message the marker is trying to convey since it is less ambiguous than corrective or emphasis-based feedback. As such it is also the type of feedback most likely to enhance learning and produce higher quality future work if it is acted upon. Nonetheless, the differences between groups for each subcategory were also comparable with those found in the control condition therefore indicating that the differences found were simply as a result of variations in marking practice.

Observation of the further subcategories for stylistic feedback identified some interesting patterns. Group 2b (AF) gained 12% more feedback on *stylistic correction: punctuation* than Group 3a (CM). However, there were also differences in this domain in the control condition which makes it difficult to make a judgement that expectancy effects were at work in the experimental condition. Nonetheless, it is perhaps noteworthy that when the student name changed from a WBM (in the control condition) to an AF (Group 2b, in the experimental condition), the increase in corrective feedback from these markers totalled 51%. While there was also an increase from markers in Group 3a (CM) of 17% this only amounted to a difference of three corrections on the work, whereas for AFs the corrections increased from eleven corrections to thirty-four. It is of course possible that the experimental essay warranted more of this type of feedback. Indeed with the exception of Group 1a (WBM) the experimental essay did gain more feedback in this domain for all other groups. Nonetheless, markers in Group 2b (AF) were only half as likely to provide such feedback in the control condition as markers in Group 3a, but then outscored them when the student name had changed. It is therefore possible that marker variability and expectancy effects played a role here.

Feedback on the further subcategories related to *referencing/citation/quotation/bibliography* also varied, but since there were also vast differences noted in the control condition these changes may simply be accounted for by variations in marking practices across groups. Nonetheless, when focused on *stylistic correction: referencing/citation/quotation/bibliography,* markers in Group 3a (CM) moved from scoring 5% in the control condition to 33% in the experimental condition. This supports the findings from Analyses 1.3, 2.1, and 3.5 demonstrating that that Group 3a (CM) received more corrective feedback on *referencing/citation/quotation/ bibliography* than any other group. In a reverse trend markers in Group 2b (AF) moved from scoring 44% in the control condition to just 4% in the experimental condition.

Percentage scores for *stylistic emphasis: referencing/citation/quotation/bibliography* also saw big differences between groups with Group 3a (CM) scoring 20% and Group 2b (AF) scoring 67%. There was a between groups score differential of 23% in the control group too, demonstrating that markers did not mark consistently in this area even when the student name remained the same. Therefore it is unlikely that these differences can be attributed to expectancy effects and are more likely explained by differences in marking practice.

Importantly, although the differences in percentages for Group 3a (CM) and Group 2b (AF) was large in the experimental condition these disparate percentages did not translate to equivalent differences in the number of feedback entries on the assignment, with the 47% difference in scores only equating to Group 2b (AF) receiving one more feedback entry on the assignment. Analysis 3.1 demonstrated that Group 2b (AF) did receive more emphasis-based feedback in this domain than Group 1a (WBMs) although generally they were found to receive more comment-based feedback in this area than other females (Analysis 2.2).

With regard to *stylistic emphasis: presentation* Group 3a (CM) scored 40% whereas Group 2b (AF) scored 0%. While it is true that scores were also different in the control condition there was only a 17% variation. Nonetheless, this still makes it difficult to claim that the differences in the experimental condition are as a result of expectancy effects on the basis of student gender and ethnicity. What is interesting however is the change in feedback behaviour between marking the control essay and the experimental essay. Specifically, the propensity of markers in Group 3a to provide this type of feedback doubled when the name changed from a WBM to a CM. Group 3a (CM) also scored 7% for *stylistic comment: presentation* whereas Group 2b (AF) scored 0%. Taken together this amounted to Group 3a (CM) being provided with ten pieces of presentation related feedback on their work while Group 2b (AF) received zero. While it is debatable how much feedback on presentation can impact on grades, it is still important to provide equitable opportunities for students to develop in all areas of their work.

This result supported those of Analysis 2.1 which found that Group 3a (CM) received more *stylistic emphasis-based* feedback on presentation than both other males groups. There were no differences when compared to Group 3b (CFs) or Group 1b (WBFs). There were no reports of Group 2b (AF) receiving less emphasis-based presentational feedback than any other groups.

Differences were also observed when the further subcategories related to *syntax/word order/ grammar* were examined. Specifically Group 2b (AF) gained more feedback across comments, corrections and emphasis than Group 3a (CM). However, differences across groups were also found in the control condition which makes it difficult to claim that the differences found in the experimental condition were as a result of expectancy effects. Nonetheless, it is compelling that while markers in Group 3a (CM) demonstrated a propensity to provide feedback on *syntax/word order/grammar* when marking the first essay, supposedly written by a WBM student, their desire to do so all but disappeared when marking work identified as authored by a CM (see Table 38). It is true that while percentages remained more stable across conditions for markers in Group 2b (AF) these percentages translated into fewer comments, corrections and emphases being made on the work in the experimental condition. Nonetheless, there was a greater reduction in feedback provided on the work for Group 3a (CM) which received forty-eight pieces of feedback in this area in the control condition and only three in the experimental condition. Other analyses suggest that Group 2b (AF) received comparable amounts of feedback in this area when compared with any other group and therefore this pattern only holds true against Group 3a (CM). However CMs did appear to receive limited feedback in this domain overall. Analysis 2.1 showed that they received far less stylistic comments, stylistic corrections and stylistic emphasis on *syntax/word order/grammar* elements of their work than any other male group and this pattern was replicated in Analysis 3.5 which found the same pattern evident when Group 3a (CM) were compared with Group 1b (WBF).

	Chinese Male (Group 3a)		**Asian Female (Group 2b)**	
Syntax/word order/grammar	Control	Experimental	Control	Experimental
Comments	17%	0%	11%	12%
Corrections	46%	6%	27%	17%
Emphasis	9%	0%	13%	17%

Table 38: Analysis 3.6: Syntax/word order/grammar feedback differences between CM and AF across conditions

Taken together, these analyses suggest that marking practices often alter significantly across assignments. Although at times it is difficult to claim that the catalyst for these differences are related to expectancy effects linked to student gender and ethnicity there are at times large fluctuations in marking styles which must raise questions as to the reliability of marking as a whole.

The total scores for overall content-related feedback were similar; Group 3a = 37%, Group 2b = 39%. These similarities were echoed in the control condition. *Content-related symbols* outranked *content-related comments* for both groups which was different from the control condition where *content-related comments* prevailed. Nonetheless, the differences evident for both comments and symbols were only slightly greater than those found in the control condition and therefore show little evidence of marker variability of expectancy effects in operation.

Despite the similarities observed at category and subcategory levels there were some differences apparent when further subcategories were scrutinised. Of particular interest were the *content-related comments* pertaining to either *positive evaluation* or *negative evaluation*. Group 3a (CM) gained more positively evaluative comments (68%) and less negatively evaluative comments than Group 2b (AF) (43%). This amounted to work perceived to have been written by a CM being awarded twenty-five positive comments, while work perceived to have been written by an AF was awarded thirteen. However, there were between group differences apparent in the control conditions too making causality difficult to establish.

Marking practices changed drastically across conditions. Markers in Group 3a moved from awarding 30% positively evaluative comments in the control condition when they perceived themselves to be marking the work of a WBM to 68% in the experimental when they perceived themselves to be marking the work of a CM. Similarly, markers in Group 2b moved from 8% when they perceived themselves to be marking the work of a WBM to 43% when they perceived themselves to be marking the work of an AF. Group 3a (CM) therefore gained more positively oriented *content-related comments* than Group 2b (AFs), similar amounts to Group 3b (CFs), Group 1a (WBM) and Group 2a (AM) and fewer than when compared to Group 1b (WBF).

Developmental feedback scores varied across both conditions, indicating that any changes were as a result of marker variability rather than expectancy effects related to student gender and/or ethnicity. Scores at subcategory level were also disparate, but since this was also the case during the control condition these scores are not worthy of further interrogation.

Structural comments gained similar percentages across groups (Group 3a = 1% and Group 2b = 3%) and *sentence level* comments prevailed over *discourse level* comments for both groups.

4.5 Summary of In-text Feedback Results

The following summary illustrates how findings from the nine previous analyses were patterned in relation to Hyatt's (2005) amended feedback classification system. Only findings related to feedback categories where between group differences were found are summarised here. These findings provide the strongest argument for the existence of expectancy effects on the basis of either student gender, student ethnicity, or student gender and ethnicity since the experimental design attempted to control for marker variability.

Interpretation of feedback findings will be minimal throughout this chapter since a more substantive analysis will be included in the discussion.

4.5.1 Developmental Feedback

According to Hyatt (2005) developmental feedback provided on student work includes, "… comments made by the tutor with the intention of aiding the student with subsequent work in relation to the current assignment". Developmental feedback included four subcategories. Three are reported here since the fourth failed to attract any feedback comments.

- *Alternative Comments:* The tutor offers alternative suggestions or points out omissions in the work
- *Reflective Questions*: The tutor poses a question for the student to consider
- *Informational Comments*: The tutor comments on a related topic in order to provide additional insight.

Differences for each of these areas will now be considered.

4.5.1.1 Gender Differences

In terms of overall percentage scores on developmental feedback the results were contradictory. WBFs received less than WBMs, but AFs received more than AMs, with the Chinese gender analysis revealing no differences.

In terms of the subcategories of developmental feedback, the *alternatives* subcategory revealed that in percentage terms WBFs had outscored their male counterparts. However WBFs actually gained fewer comments on their work (twenty-one versus thirty-two), and therefore WBMs were provided with more opportunities to use such feedback. Similarly AFs received fewer comments and received a lower percentage score on *alternatives* than AMs. It is important to acknowledge that although *alternative* feedback was considered by Hyatt (2005) to be developmental in nature, throughout these analyses feedback categorised in this way often included judgmental comments which highlighted omissions in the work. The impact of judgmental and negatively-oriented feedback has been reported to reduce student motivation and self-efficacy (Hattie & Timperley, 2007; Van Dinther et al., 2011; Nash et al., 2015). This notwithstanding, the developmental aspect of this type of feedback means that if student can use this feedback in a facilitative way it has the capacity to feedforward into subsequent work and provide useful learning opportunities. Therefore White British and AMs were provided with feedback which if interpreted positively offered them a learning opportunity which was absent for White British and AFs.

Another subcategory which highlighted some differences was *reflective questions*. WBMs received more than double the amount of feedback than their female counterparts (45% versus 19%). In contrast however, AMs received less feedback in this domain than their female counterparts (30% versus 43%).

The gender results therefore present a mixed picture. Although WBFs have fared badly the landscape for Asian students was less clear. AMs were advantaged by receiving more *alternative* comments than their female peers, whereas AFs were asked more *reflective questions* than AMs. No differences were reported at subcategory levels for Chinese groups.

4.5.1.2 **Ethnicity Differences**

Comparison of the male ethnicities only illustrated a difference within the *alternative* subcategory. In purely percentage terms AMs gained the highest score. However, since WBMs attracted a far higher percentage for developmental feedback overall, they actually received the highest number of *alternative* comments (thirty-two), compared to AMs (twenty-nine comments) and CMs (thirteen comments). Therefore WBMs have been shown to be advantaged by this type of feedback in relation to their WBF peers and their male Asian and Chinese counterparts. There were no differences found for *alternative* feedback within the comparison of female ethnicities.

4.5.1.3 **Gender and Ethnicity Differences**

The sole difference found across the remaining six analyses for *alternative* comments related to the CM versus WBF analysis. Although percentage scores were almost identical, CMs received fewer comments (thirteen versus twenty-one) on their work. The same was true for the CM when compared to other male ethnicities, demonstrating that they were disadvantaged in relation to a male student with a White British or Asian name and a student with a WBF name.

More differences were found in the domain of *reflective questions*. Firstly, the AM gained 14% less feedback and nine fewer comments designed to provoke reflection than the CF. This supports the results from the ethnicity analyses where the AM also received the lowest amount of feedback in comparison to their male peers. No differences were reported between AMs and AFs, but the result was reversed when they were compared to WBFs. Specifically, WBFs gained 11% less feedback here than the AM and attracted only half the amount of questions. This demonstrated that with the exception of the WBF the AM name attracted the least reflective feedback. The position of the WBF was compounded further when compared with the CM. In this instance the WBF scored 19% and CM 35%. However the amount of comments provided on the work only showed a small difference in favour of the CM. Nonetheless, this finding reveals that the WBF received the lowest amount of developmental feedback related to *reflective questions* than any other group.

Three analyses examining gender and ethnicity found disparities in the area of *informational comments*. The WBM had already been reported to gain more feedback in this domain than the CM as part of an earlier analysis and now also outscored the CF by 10%. In addition the WBF was provided with 10% more comments than AMs and 9% more than CMs. No between group differences were reported for the other analyses, but it appears that White British students of both genders are advantaged by receiving more of this type of comment in relation to specific groups. For WBF students this might serve to mitigate for the lack of feedback they have received in other developmental domains (i.e. *reflective questions*). No such mitigating factors exist for CM student however who gained less developmental feedback overall than their other male counterparts and gained low scores for *informational comments* and the lowest score for *alternatives.*

4.5.2 **Structural Feedback**

Structural feedback refers to the organisation of the assignment. Hyatt (2005) considers that it has two subcategories;

- Discourse Level: Comments that consider structure at a macro level and how specific sections fit together to form a coherent whole
- Sentence Level: Comments that consider structure at a micro level in terms of the organisation of individual sentences and how they link to other sentences.

There were minimal comments on structure across the 120 essays and only one gender-based between groups difference. This difference will now be discussed.

4.5.2.1 Gender Differences

CMs gained less overall feedback on structure than CFs. They received three comments overall whereas CFs gained eleven. Additionally 100% of the male feedback related to Sentence Level issues whereas the female gained a near even split between both *sentence* and *discourse level* feedback. Therefore the CF was aided by the receipt of more feedback designed to help organise her assignment. No differences were reported for any other analyses.

4.5.3 Stylistic Feedback

Stylistic feedback gained by far the most feedback on the 120 assignments and therefore revealed numerous between group differences. Hyatt (2005) considered stylistic feedback to reflect, "… comments which consider the use and presentation of academic language". He then included six further subcategories which addressed various aspects of academic language. Only the following four received feedback;

- *Punctuation*
- *Syntax/word order/grammar*
- *Referencing/citation/quotation/bibliography*
- *Presentation*

Each of these further subcategories was further categorised according to whether markers had made a comment on the issue (*punctuation: comment*), corrected the issue (*punctuation: correction*), or simply emphasised the issue by way of circling some text for example (*punctuation: emphasis*). As such, previous explanations of the diminishing educative functions of comment-based, corrective and emphasis-based feedback types need to be borne in mind again here.

4.5.4 Stylistic Comments, Corrections and Emphasis

Scores gained for the subcategories of *stylistic comments, stylistic corrections* and *stylistic emphasis* were comparable across all analyses with the exception of CM versus AF. The CM followed the pattern for every other group which was that comments dominated over corrections which dominated over emphasis. Nonetheless, the AF was the only group who gained a higher percentage for *stylistic corrections* than for *stylistic comments* and finally *stylistic emphasis*. This translated to the CM gaining fifty-nine comments on their work regarding aspects of their academic language skills, compared to just thirty-three for the AF. Clearly given the difference in quality between comment-based corrective and emphasis-driven feedback this potentially disadvantaged the AF student in comparison to her CM counterpart. Unfortunately no other between group differences can be included here since there was too much variation in marking practice in the control groups to allow confidence that differences in the experimental group were a result of expectancy effects. Consequently interactive expectancy effects can only be reported between the CM and AF group.

This notwithstanding, it was actually at further subcategory levels (i.e., *punctuation*; *syntax/word order/grammar; referencing/citation/quotation/bibliography* and *presentation*) that differences in the data became apparent. This demonstrated that if the feedback landscape is to be understood more comprehensively it is important to move beyond the superficial to interrogate the data more rigorously. With this in mind the differences generated for each further subcategory of stylistic feedback will now be discussed.

4.5.4.1 **Gender Differences: Punctuation**

The first gender specific difference in stylistic feedback related to the domain of *punctuation*. While WBMs and females gained comparable feedback scores for both comments and corrections there were substantial differences linked to emphasis-based feedback on *punctuation*. These differences amounted to 71% of emphasis-based feedback awarded to the WBM being punctuation-specific with only 7% for WBFs. There were no differences found for the other gender analyses.

4.5.4.2 **Ethnicity Differences: Punctuation**

The patterning of punctuation-related feedback for male ethnicities was interesting. Specifically, Asian or CM names were more likely to receive comment-based or corrective feedback on punctuation elements of their work whereas WBMs received less of these in place of the receipt of more emphasis-based feedback. CMs also outscored AMs in terms of comment and corrective-based feedback. Given that the worth of different feedback types has been previously discussed it would seem misguided to assume that the WBM has been advantaged by such a superficial form of feedback. Instead this finding indicates that the CM name has been provided with a better developmental opportunity.

In terms of female ethnicities it was difficult to draw such firm conclusions on punctuation-related feedback since differences did exist in both the control and experimental conditions. Usually this would therefore preclude such findings being reported here. Nonetheless, an interesting pattern was observed for punctuation-related corrections. All markers perceived the experimental essay to contain more punctuation-related errors than the control essay. Markers in the WBF essay group increased their score by 19%, markers in the CF group increased theirs by

27%, but markers for the AF increased theirs by 51% demonstrating that when markers perceived the work to be authored by an AF punctuation errors were corrected far more often.

While it is difficult to compare all aspects of these findings across ethnic groups, it appears that Chinese and Asian students attract more significant feedback on punctuation-related issues than White British students do. This is particularly true for WBMs.

4.5.4.3 **Gender and Ethnicity Differences: Punctuation**

Three analyses showed differences for *punctuation*. CFs were provided with more corrective feedback on *punctuation* than WBMs (63% versus 46%), but WBMs dominated on emphasis-based feedback, exceeding the CF score by 40%. No differences were reported between WBMs and AFs. The dominance of emphasis-based feedback on *punctuation* feedback for WBMs was also found when they were compared to other male ethnicities and WBFs. This indicates that the WBM name stimulated expectancy effects on the basis of ethnicity, gender and the interactive effects of gender and ethnicity with the exception of AFs.

In addition to gaining more corrective feedback than WBMs, CFs also gained a higher percentages of both corrective (63% versus 50%) and emphasis-based feedback (31% versus 18%) compared to the AM. No differences were found when the CM and CF were compared, illustrating that no gender bias existed when ethnicity remained the same. However changing both gender and ethnicity demonstrated differences that can be attributed to the interactive effects of gender and ethnicity when compared to White British or AMs. This finding is partly at variance with results from the female ethnicity comparison where CFs were only provided with more corrective feedback than WBFs. This suggests that interactive expectancy effects hold true for CFs across all groups aside from AFs.

When the CM was compared to the WBF they received more feedback in every category (comments, corrections and emphasis). A 33% difference in the emphasis-based domain was the biggest discrepancy between groups. No differences were reported when CMs were compared to Chinese or AFs. Interactive expectancy effects therefore only revealed themselves when the CM was compared to a WBF. This notwithstanding, results from previous analyses where gender was controlled also showed that the CM was provided with more corrective feedback than Asian and WBMs suggesting that when compared to students of these ethnic groups ethnicity alone is sufficient to trigger an expectancy effect.

Taken together these results suggest although WBMs received more emphasis-based feedback than any other group it was the Chinese names of both genders that consistently attracted more useable forms of feedback on *punctuation*. Student ethnicity (as triggered by the name) stimulated an expectancy effect which lead markers to, a) notice more punctuation errors, and b) feel compelled to correct them. That both Chinese names had the same impact on markers behaviours could indicate that either gender was irrelevant, (since both Chinese names gained more corrective feedback) or that markers could not distinguish between genders upon viewing the Chinese names. Either way it appears as though the interactive effects of gender and ethnicity have created the expectancy effect in this instance.

4.5.4.4 **Gender Differences: Syntax/word order/grammar**

Gender differences related to *syntax/word order/grammar* were only apparent between the WBM and female and this was solely in terms of emphasis level feedback. Only 7% of total emphasis-based feedback fell into this domain for the male student, whereas for the female student this score was 36%. Since this was the only group for which a gender comparison was possible we cannot claim that these gender effects are wide-ranging. Furthermore, they only relate to what might be considered a less substantive type of feedback (i.e. emphasis-based).

4.5.4.5 **Ethnicity Differences: Syntax/word order/grammar**

There were some interesting differences found for *syntax/word order/grammar* within the analysis for male ethnicities. WBM and Asian names attracted similar levels of feedback for comments (7% versus 11%) and corrections (26% versus 21%) whereas the male Chinese name attracted 0% for comments and only 7% for corrections. This failure to attract a single comment on their academic writing skills was not compensated for by other types of feedback in this domain since only three corrections were made and there was no emphasis-based feedback. Since no differences were recorded between the CM and CF names it is possible that gender is less likely to stimulate expectancy effects for students with Chinese names than ethnicity is. No differences were found in the analysis of female ethnicities indicating that female students from different ethnic groups did not trigger the same expectancy effects for this type of feedback as male students did.

4.5.4.6 **Gender and Ethnicity Differences: Syntax/word order/grammar**

Differences for gender and ethnicity were only observed for two groups. The AM gained more correction-based feedback on their academic writing skills than the CF (21% versus 10%). Unfortunately however no differences were found between CFs and CMs or between CF and other female ethnicities and therefore no comparative data is available to place this finding into context. It does however demonstrate that both CM and female names gained less feedback in this area than AMs and therefore when compared to an AM name expectancy effects were operational on the basis of gender and ethnicity.

Nonetheless, when the CM was compared to the WBF they gained less feedback across the subcategories of comments (0% versus 17%), corrections (6% versus 30%), and emphasis (0% versus 36%). The CM is therefore disadvantaged when compared to other male ethnicities and female White British names, but not to Chinese or AF names. These findings lend support to the concept of trait centrality where some information about a target is more influential than other pieces of information (Asch, 1946). When the CM is compared against other male names ethnicity is the central trait that triggers the expectancy effect and gender is secondary. However, when the CM is compared to Asian or CFs, ethnicity moves to the background and gender becomes the central trait as competing biases have vied for prime position. The fact that only the WBF outscored the CM in this domain may indicate that WBFs are protected from expectancy effects related to gender by their privileged White British status, whereas Asian and CFs are susceptible to the 'double jeopardy' effect (Thomas & Miles, 1995).

4.5.4.7 **Gender Differences: Referencing/citation/quotation/bibliography**

The first gender difference in this category showed that the WBM gained a higher percentage of referencing-related comments than the WBF (58% versus 30%). This translated to the male gaining sixty-two comments on his work and the female gaining fourteen. The second difference related to corrections instead of comments, but demonstrated that 33% of the CM's corrective feedback pertained to referencing issues, whereas for the CFs this was 17%. There are therefore gender related expectancy effects in operation here that potentially disadvantage White British and CFs. There were no differences reported between AMs and females.

4.5.4.8 **Ethnicity Differences: Referencing/citation/quotation/bibliography**

Although no differences were reported for male ethnicities at comment level, there were differences in the level of corrective feedback issued. Specifically the WBM scored 7%, the AM 1%, and the CM 33%. Given that the CM also outscored the CF on corrective feedback in the earlier analysis it is evident that the activation of expectancy effects were triggered by both the gender of the CM and his ethnicity. Unfortunately it was difficult to draw such firm conclusions across all female ethnicities since differences did exist in both the control and experimental conditions. For example, although CFs did gain many more comments on referencing than either White British or AFs, markers in this group had already demonstrated a high propensity to mark in this way in the control condition. Consequently results are only credible for the White British and AF since their scores in the control condition were comparable. In this instance AFs gained 64% of their *stylistic comments* on referencing-related issues whereas the WBF only gained 30%. These results indicate that CMs are advantaged in comparison to other males and AFs are advantaged in comparison to other female groups on the basis of comment-based and corrective feedback for referencing-related issues.

4.5.4.9 **Gender and Ethnicity: Referencing/citation/quotation/bibliography**

In relation to analyses exploring interactive effects differences were found in three analyses. Although there were no differences at comment and correction levels for AFs versus WBMs, AFs did gain higher scores at emphasis level for referencing-related feedback. The AF therefore gained more feedback on referencing-related issues when compared to both the WBM and female groups. No differences were found between the AF and either the Asian or CM. Admittedly, the differences related to WBMs only referred to emphasis-based feedback, but it does indicate that expectancy effects have operated here on the basis of both gender and gender and ethnicity.

Differences were also found between the CM and the WBF. The CM received higher percentage scores for both comment (60% versus 30%) and corrective feedback (33% versus 6%). This continues a pattern for the CM since he also gained more corrective feedback than other male ethnic groups and the CF. Therefore the CM gained more corrective feedback when compared to every other group for which comparisons were available. This demonstrated that expectancy effects for a male Chinese name were triggered by gender, ethnicity and the interactive effects of gender and ethnicity. It has been less easy to determine the expectancy effects in operation for the CF since fewer comparisons were available. However, although she scored lower than the CM for corrective feedback she did outscore WBMs (17% versus 7%). It is therefore evident that the expectancy effects found for the CF were provoked by gender and the interactive effects of gender and ethnicity. Interestingly most of the interactive effects found for this feedback category were found when Chinese names were compared to White British names showing that participants marking student work from either of these ethnic backgrounds might be more prone to expectancy effects.

Taking the results for referencing-related issues together it was clear that students with a Chinese name were much more likely to obtain feedback connected to referencing-related issues than students with either a White British or Asian name.

4.5.4.10 **Gender Differences: Presentation**

No gender differences were found for presentational issues for any group.

4.5.4.11 **Ethnicity Differences: Presentation**

Although no differences were found between female ethnicities there were vast differences observed for the male ethnicities. White British and AMs received comparable amounts of *comment-based and corrective presentational* feedback whereas the CM scored much lower in each further subcategory. However the largest difference was found for emphasis-based feedback where the trend reversed and CMs received a score of 40%, AMs received 27% and the WBM received 0%. Overall therefore the WBM gained the least feedback on presentational issues, but where feedback was provided it was more substantial. The AM received a balanced profile of comments, corrections and emphases, whereas the CM principally received less valuable and more ambiguous emphasis-based feedback. This findings suggest that expectancy effects have been generated here on the basis of student ethnicity when participants marked essays bearing male names.

4.5.4.12 **Gender and Ethnicity Differences: Presentation**

Only two of the six analyses generated any notable differences. Firstly, the AM outscored the CF on all aspects of presentational feedback (comments 17% versus 9%, corrections, 15% versus 2% and emphasis, 27% versus 0%). The second analyses found that the CM received more emphasis-based feedback than the AF (40% versus 0%). Consequently the AM is more likely to be advantaged by the presentational feedback he received since it contained *comments* and *corrections*. Although the CM did gain more feedback than the AF this only pertained to *emphasis-based* feedback.

Nonetheless, in these analyses it was the male names which gain more feedback on presentation. This propensity to award more presentational feedback to male names cannot be compared to the results of the gender-based analyses since differences were not reported for either Chinese names or Asian names. This lack of reporting does not definitively claim that between group differences did not exist, just that when differences already existed in the control condition any resultant changes in the experimental condition make later claims of difference unreliable. This suggests that gender alone was not sufficient to trigger expectancy effects in the marking process and may not be the central factor driving feedback practice in this instance. Instead there is some evidence that the interactive effects of gender and ethnicity can stimulate changes for Asian and Chinese names.

4.5.5 **Content-Related Feedback**

Hyatt (2005) maintained that content-related feedback referred to whether the content of the assignment was appropriate and/or accurate. This category was made up of two subcategories; *content-related comments* and *content-related symbols*. Comments referred to any feedback which included a text-based comment on content. Symbols were added as part of the critical amendments made to the framework since it was apparent that many markers added ticks to work when they considered the content to be good, or crosses when they considered it to be incorrect. Previous commentary about the value of ticks on student work will need to be taken into consideration again here. Both comments and symbols could either be positive, negative or non-evaluative in nature and these were reflected through the further subcategories below.

- *Positive Evaluation:* Comments on the strengths of the work are noted, particularly in relation to theory, literature, the ability to construct argument and reflection.
- *Negative Evaluation:* Comments on the weaknesses are noted which include problems with those areas cited above, as well as problems linked to the provision of evidence, a lack of clarity, the need for clarification, and a lack of criticality.
- *Non-Evaluative Summary:* Comments offer a summary of the assignment.

4.5.5.1 **Gender Differences**

WBMs and females gained comparable levels of content-related feedback and this was distributed evenly between comment-based and symbol-based feedback. Interestingly AFs gained substantially more content-related feedback than AMs (39% versus 26%). Given the previously outlined discussion about the usability and educative function of symbol-based feedback however it is important to acknowledge that the higher percentage gained by the AF was only made up of *content-related symbol* feedback as opposed to *content-related comments*. In fact, the AF only gained half the *content-related comments* of the AM (thirty versus sixty). This indicates that the AF was deprived of an opportunity to move beyond surface level learning for aspects of her essay related to content. There were no differences reported in content-related feedback for CMs versus females.

Exploration of further subcategories also revealed some differences. For example, the WBM gained more positive feedback comments than the WBF (56% versus 43%) which translated to the male gaining twenty-seven positive comments on their work, while the female only received fifteen. In terms of negative feedback the WBF gained more negative comments (57% to 40%) than their male peers. There were no gender differences reported for the Asian or Chinese names.

4.5.5.2 **Ethnicity Differences**

Comparison of male ethnicities found that the CM received the most overall content-related feedback at 36%, followed by AMs at 26% and WBMs at 20%. While initially this suggests that the CM was in the most privileged position and the WBM in the least, closer inspection of subcategories revealed that in terms of *content-related comments* male ethnicities all scored similarly and no differences were evident for the further subcategories related to the positive and negative nature of those comments. Instead the increased percentage score for CMs was generated exclusively from *content-related symbols* where they gained 21% of their overall feedback, compared to AMs who scored 8% and WBMs who scored 6%. Examination of further subcategories showed that the symbol-related feedback was 100% positive for all male ethnic groups and translated to the WBM having twenty ticks on their assignment, the AM having twenty-three, and the CM having forty-eight. Nonetheless, positive interpretations of the additional feedback provided to the CM must be interpreted with caution since higher overall content-related feedback scores do not always consist of useable or developmental feedback. This underlines the importance of examining the data at subcategory and further subcategory levels because overall percentages can often obscure the more compelling patterns in the data.

Comparison of female ethnicities also revealed some differences. AFs gained the most overall content-related feedback at 39%, followed by the CF at 31% and the WBF at 30%. This finding confirms that both WBMs and females have gained the least amount of overall content-related feedback in comparison to their Asian and Chinese peers. However, once again examination of the subcategories reveals a more complex picture. While the White British and CFs overall content-related feedback was evenly distributed between comments and symbols it has already been acknowledged as part of the gender results that the AFs was not. In fact, despite a higher overall percentage she scored the lowest for c*ontent-related comments* when compared to other female ethnicities. This finding demonstrates that the AF is disadvantaged not only in relation to other AMs but also in relation to the female ethnicities included here. It also potentially changes the impact of the finding for WBMs and females when it is considered that despite receiving the lowest amount of overall content-related feedback both groups received comparable amounts of comments to other ethnic groups. The only area they scored lower in was the receipt of symbol-based feedback (in the form of ticks on their work).

Nonetheless, the feedback landscape shifts again somewhat when the findings for the further subcategories are considered. The subcategory of *content-related comments* generated data for two of its further subcategories; *positive evaluation* and *negative evaluation*. Analysis revealed that CFs were provided with 18% more positive comments than either White British or AFs. This was evidenced through the CF receiving thirty-one positive comments on her work compared to the WBF who received fifteen and the AF who received thirteen. Therefore overall AFs received the least *content-related comments* and received fewer positive comments on their work than other female ethnicities. WBFs generated a balanced profile of comments versus symbols but received the second lowest score for positive comments after their Asian peers. CFs fared best since they gained a balanced profile of comments and symbols and their comments were more positively oriented. Unfortunately a comparison of this further subcategory data with the male ethnicities is not possible since no differences were reported.

4.5.5.3 **Gender and Ethnicity Differences**

Only two analyses revealed any differences in content-related feedback. The first related to WBM versus AF and illustrated that AFs received more overall content-related feedback (39% versus 20%). However, once examination of subcategories and further subcategories was completed it was once more clear that the additional feedback received by AFs was solely comprised of symbols (26% versus 65%) and they received far fewer *content-related comments* (thirty comments versus forty-eight). This result adds weight to the findings from previous analyses which have already shown that AFs gained less useful types of content-related feedback than both their male Asian peers, and other female ethnicities. Since no comparisons were available between AMs and CMs it can be concluded that AFs fared worse than every other group in this domain. In addition they also gained fewer positive comments than other female groups.

The second analysis demonstrated a similar patterns when comparing WBMs with CFs. Although scores were comparable (showing a difference of only 2%). CFs gained more symbol-related feedback than their WBM peers (15% versus 6%). As was the case with the previous analysis, these symbols were 100% positive. This finding supports that of the CM who also gained more symbol-based feedback than his male peers. Nonetheless, it is important to distinguish between the results for the AF and for the CM and CF. Although these groups all gained more symbol-based feedback this issue has more potential impact for the AF since they did not gain comparable amounts of *content-related comments* which was the case for the CM and female. Where students received similar amounts of comments *and* additional ticks on their work the ticks can simply be viewed as a nice addition to the overall feedback picture, but where they have been used as a substitute for the provision of more substantive feedback in the form of comments this becomes problematic for student learning.

There were no differences found between AMs and CFs, AMs and WBFs, AFs and CMs and CMs and WBFs.

4.6 Summary Feedback

This chapter will now present the findings from the hierarchical content analysis conducted on the summary feedback.

Participants were asked to mark assignments in line with their current marking practices. This led to many failing to provide a summary feedback statement. Specifically, 50 essays out of 120 did not contain any summary feedback. This figure is alarming and indicates the variability present in marking practices in some UK-based HEIs. However, although there were small variations across different genders and ethnicities regarding the missing feedback no pattern could be discerned which would point to expectancy effects being the cause.

As previously identified in the methodology chapter, the hierarchical content analysis was conducted following procedures outlined by Sparkes and Smith (2014). Overall 370 raw data themes were found; 198 in the control essays and 172 in the experimental essays. These coalesced into eleven sub-themes which were subsequently categorised into four higher-order themes; *Descriptive Feedback, Autocratic Feedback, Supportive Feedba*ck and *Developmental Feedback*. Table 39 illustrates the content of these themes and their corresponding sub-themes. Examples of the type of comments that were included in each sub-theme are also provided.

Higher-order themes	Sub-themes	Exemplars
Descriptive Feedback	Positive Description	"Good structure with a well-developed and balanced argument".
	Negative Description	"On occasion you use 10 words when 3 would do"
	Neutral Description	"You have initially identified the aim of your argument and have then presented discussions of different sides of this debate"
Autocratic Feedback	Instructional	"ensure that your first sentence picks up clearly from the preceding one and has a concluding statement which ideally relates to the essay title"
	Directive	"develop your style of writing"
Supportive Feedback	Encouragement	"This was a promising piece of work which demonstrated your

		developing power of critical discussion.”
	Praise	“Overall, well done”
Developmental Feedback	Future-oriented	“you will need to start referring to the primary literature in future essays”
	Advice	“Don’t be afraid to have your own say, you have the critical capacity to defend your own position”
	Informational	“Therapeutic use exemptions are already in place! Bit of an oversight here, would have been worth doing a bit more reading on this section.”
	Resource-based	“Make sure you check University handbooks for citing sources.”

Table 39: Hierarchical content analysis results for summary feedback

Results will be analysed and discussed in relation to differences found in the feedback provided on the basis of student gender, ethnicity, and gender and ethnicity. This will include details on the number of comments provided for each of the themes and sub themes. However, it is recognised that these numbers are quite small and only represent one way of examining the presence of expectancy effects. Therefore, more meaningful interpretations of the data will be sought through exploration of the feedback content. This will include but is not limited to, consideration of the tone of feedback, and noticing whether specific groups attract negative comments that revolve around specific issues. The primary focus here will be on whether the feedback demonstrates expectancy effects at work. The secondary concern is how any differences in feedback content may impact the learning experience of the student.

It was not always possible to compare sub-themes because groups did not generate any comments. For example, no male ethnicities attracted any feedback for the *Autocratic: Directive* sub-theme in the experimental condition and therefore it was impossible to conduct any gender or ethnicity comparisons. Furthermore, even if data were available for each gender and ethnicity it has only been discussed here if the groups being compared had provided similar feedback in the control condition (when they both marked work identified as being written by a WBM), but then provided different feedback in the experimental condition when they marked the same assignment as each other but with a different name. As was the case with the in-text analyses, the only exception to that is when differences were noted within groups and across conditions that also signalled expectancy effects were in operation. Despite the previously discussed loss of experimental control that such comparisons include they were considered useful to include.

However, since the level of confidence for expectancy effects being the catalyst for these changes is lower when comparing within groups the reader will be informed when inferences are drawn from this data.

The tables below detail the number of comments provided for each higher order and sub-theme according to student gender and ethnicity. Results for the control condition are presented first followed by the experimental condition.

Number of Comments for Summary Feedback Themes (Control Condition)						
All Groups Marked a WBM Name	**Group 1a**	**Group 2a**	**Group 3a**	**Group 1b**	**Group 2b**	**Group 3b**
Descriptive Feedback						
Positive Description	4	11	3	11	4	7
Negative Description	16	22	7	13	6	11
Neutral Description	1	1	1	1	0	0
Total No. of Descriptive Comments	*23*	*34*	*11*	*25*	*10*	*18*
Autocratic Feedback						
Instructional	7	11	3	6	6	6
Directive	0	2	2	2	2	2
Total No. of Autocratic Comments	*7*	*13*	*5*	*8*	*8*	*8*
Supportive Feedback						
Encouragement	0	2	0	1	1	1
Praise	2	0	0	0	0	0
Total No. of Supportive Comments	*2*	*2*	*0*	*1*	*1*	*1*
Developmental Feedback						
Future-oriented	0	1	1	0	0	0
Advice	3	3	4	4	4	6
Informational	1	7	0	2	0	0
Resource-based	0	0	0	0	2	0
Total No. of Developmental Comments	*4*	*11*	*5*	*6*	*6*	*6*

Table 40: Number of comments for summary feedback themes (control condition)

Number of Comments for Summary Feedback Themes (Experimental Condition)						
	WBM (Group 1a)	**Asian Male (Group 2a)**	**CM Group 3a)**	**WBF (Group 1b)**	**Asian Female (Group 2b)**	**CF (Group 3b)**
Descriptive Feedback						
Positive Description	9	13	14	15	14	11
Negative Description	7	13	2	8	7	10
Neutral Description	1	1	0	0	1	0
Total No. of Descriptive Comments	*17*	*27*	*16*	*25*	*22*	*21*
Autocratic Feedback						
Instructional	3	6	2	4	4	0
Directive	0	0	0	1	3	1
Total No. of Autocratic Comments	*3*	*6*	*2*	*5*	*7*	*1*
Supportive Feedback						
Encouragement	0	1	1	0	0	3
Praise	1	3	0	5	1	0
Total No. of Supportive Comments	*1*	*4*	*1*	*5*	*1*	*3*
Developmental Feedback						
Future-oriented	0	0	0	0	0	0
Advice	1	4	3	1	3	2
Informational	0	4	2	3	1	0
Resource-based	0	0	1	1	0	0
Total No. of Developmental Comments	*1*	*8*	*5*	*5*	*4*	*2*

Table 41: Number of comments for summary feedback themes (experimental condition)

4.6.1 Summary Feedback Differences: Gender

The different number of comments provided to WBF versus WBMs for *Positive Descriptive* feedback cannot be interpreted as illustrative of expectancy effects at work since the number of comments also varied in the control condition. However, it was noticeable that there was a difference in the content of the positive feedback provided. WBF essays gained more *Positive Descriptive* feedback related to developing argument in their work than WBMs. The comments WBMs received were also less extensive, for example, “valid arguments backed up”, and, “some good arguments”. Conversely, WBFs received comments such as, “Overall you provide evidence to support your arguments using good introductory and concluding sentences that add clarity”, and, “You have developed your arguments with good reasoning and the text flows well”. The greater context provided in these example serves to provide a richer learning opportunity for the WBF student since it aligns their feedback with an example of how they managed to embed argument. This makes it easier for them to replicate this good practice in future assignments.

Additionally, WBFs were provided with more *Supportive Feedback* than WBMs. This was most apparent within the subtheme related to *Praise* where males received one comment and females received five. The comment on the male essay simply read “Good attempt” whereas the feedback for the females included “I appreciate your attempt to answer the question”, “I liked that you took a position in your conclusion”. The importance of praiseworthy feedback particularly for first year students is supported by the work of Shields (2015) who demonstrated that student interpretations of feedback are closely linked to their beliefs about themselves as learners. On the basis of these results therefore, WBF students are more likely to believe in their potential and have confidence in writing future assignments than WBMs.

The primary aim of the thesis was to identify and discuss between group differences within conditions rather than comparing within groups across conditions. Nonetheless, there are times when other comparisons can be insightful. For example, participants in Group 2a (AM), 2b (AF), 3a (CM), or 3b (CF) did not provide any *Positive Descriptive* feedback on structure when marking the control essay (identified as being written by a WBM), but when participants thought they were marking the work of a non-White British student in the experimental condition comments were provided in this domain (AM, n=3; AF, n=4; CM, n=3; CF, n=3). For example, “…you have set out a clear structure with an introduction, main body and a clear conclusion which acts as an effective summary of your views”, and “Your structure & progression is good”. Furthermore, participants in Groups 3a (CM) and 3b (CF) did not make positive comments about writing style when they thought the author was a WBM in the control condition, but they made several such comments when they perceived themselves to be marking the work of either a CM (three comments) or female (two comments) in the experimental condition. For example, “very clear and easy to read”, and “Your language and sentence structure was easy to follow and helped in making your point”. These differences in feedback content point to ethnicity-based expectancy effects operating within specific groups of markers. Furthermore, if such feedback is perceived as desirable to the students or as having some educative function then the concept of sympathetic marking may be relevant here (Shay, 2008). Sympathetic marking pertains to, “…a situation where inferences about student identity may result in a more generous assessment” (Shay, 2008, p.161). Although Shay’s research was grade rather than feedback-focused, she reported that non-anonymised assessment had the potential to influence markers to apply a positive bias to students for whom English was their second language. In practice this lead to these students being given 'the benefit of the doubt' when grades were applied. Perhaps the same can be said here in terms of feedback. Students with non-British names were provided with feedback on the more elementary aspects of academic writing perhaps due to a perception that they might not have had the same opportunities as students with a British name to understand the necessity of such criteria.

Summary feedback differences related to *Negative Descriptive* feedback were also observed. For example, Group 3a (CM) only received two negative comments whereas Group 3b (CF) received ten. Admittedly, there were also between group differences in the control condition which amounted to four comments, and it is also true markers in Group 3a did show a propensity to provide a low amount of negative feedback here too. However, in addition to the low number of negative comments provided to the CM in the experimental condition, it was the content of these comments that was interesting. CMs only received comments related to referencing style and quality of sources. However, when the essay was labelled as being written by a CF it also attracted feedback on issues of structure, writing style, clarity of argument and the quality of concluding comments. Furthermore, the feedback comments were more detailed. For example, "However, you mix the issue of 'social drugs' with 'performance enhancing drugs' which although both related to the question, does little to clarify the strength of your argument", and "I don't think you convey a convincing path to the conclusion you reach so that it can be seen to be clearly and well-based."

Who is advantaged through the receipt of more or less *Negative Descriptive* feedback is open to debate. Perhaps receiving fewer negative comments which are limited in scope serves to both protect the student's self-esteem (Shields, 2015) and help them to perceive the required improvements as manageable. In this case the CM may be in the most fortunate position. However, there is evidence to suggest that negative feedback can be motivational in nature (Pitt & Norton, 2016), and importantly that failure to provide negative feedback impacts learning opportunities and stymies development particularly for ethnic minority groups (Harber, 1998; Croft & Schmader, 2012).

Expectancy effects can therefore be considered to be in operation in the following ways regarding gender. WBFs were provided with more *Positive Descriptive* comments related to argument and more praise-related feedback than WBMs. CFs were provided with more Negative Descriptive feedback which pertained to a wider range of issues than CMs. Furthermore, across conditions and within group differences were noted for the following areas; non-White British students received more positive comments on the structure of their work and the CM gained more positive feedback on writing style.

4.6.2 **Summary Feedback Differences: Ethnicity**

Although there were no between group differences for *Positive Descriptive* feedback, it was once again interesting to note two within group changes from the control to the experimental condition. First, markers in Group 3a (CM) only provided three positive comments when they thought themselves to be marking the work of a WBM in the control condition. However in the experimental condition when marking the work of a CM the same group of markers provided fourteen positive comments. The same pattern was true for markers in Group 2b (AF) who provided four *Positive Descriptive* feedback comments in the control condition, but fourteen in the experimental condition when they perceived themselves to be marking the work of an AF. While generally the second essay did receive more *Positive Descriptive* comments than the control essay, this increase was far larger than that of the other groups. Furthermore, for both the CM and the AF the types of positive feedback they were provided with was different. In the control condition markers commented on evidence of wider reading, key concepts being addressed, the relevance of the work, and an ability to answer the question. When they perceived themselves to be marking the work of a CM or AF however, they commented on a broader range of issues such as a clear writing style, the structure of the work, the style and range of references, the ability of the student to construct argument, their level of understanding and the ability to write a well-balanced conclusion.

There was a difference for *Instructional Autocratic* feedback among the female ethnicities. When the essay was presented as being written by either a White British or AF they were awarded four instructional comments each. However, when the same essay was presented as being written by a CF markers did not provide a single comment. This suggests that expectancy effects have played a role on the basis of student ethnicity. This can be further supported by the fact that markers in Group 3b (CF) did not shy away from providing *Instructional Autocratic* feedback in the control condition when they considered themselves to be marking the work of a WBM. In fact, their scores in this condition were on par with markers in both other groups. Whether this finding represents an advantage or a disadvantage for the CF student is debatable. Purely autocratic feedback represents what McLean et al. refer to as, "…feedback as telling" (2015, p.925). How such feedback is received by students will depend upon a range of individual and cultural factors, not least whether the students wants to be passive recipient in the process. This notwithstanding, *Instructional Autocratic* feedback was differentiated from *Directive Autocratic* feedback in the content analysis because it did fulfil an educative function as well as a command. An *Instructional Autocratic* feedback comment about structure might state

> "Part of this relates to your paragraph structure - ensure that your first sentence picks up clearly from the preceding one and has a concluding statement which ideally relates to the essay title. You also need to use one paragraph to develop one point" (Female Asian, Experimental)

Whereas a *Directive Autocratic* feedback comment on the same issue might simply state "Work on essay structure" (Female Asian, Control).

Supportive Feedback comments for both *Encouragement* and *Praise* were quite limited for both genders and across both conditions. However, AMs did receive more supportive feedback (four comments) than WBMs (one comment). Furthermore, the feedback provided for the AM was sometimes more detailed. For example, "on the basis of this with some attention to my comments you should be competent of improving further" and "Good attempt to summarise a complex argument" as opposed to the comment for the WBM which simply stated "Good attempt".

Similar differences for *Supportive Feedback* were witnessed for female ethnicities too. Scores were the same for all ethnicities in the control condition with each group receiving a single comment. In the experimental condition WBFs gained five comments, CFs gained three and AF only received one. While the comment for the AF rather blandly stated that it was a "nice essay to read", White British and CF students were given more specific feedback which alluded to their competencies at more advanced academic skills. For example, "This was a promising piece of work which demonstrated your developing power of critical discussion" (CF), "you have the makings of a balanced, argued essay" (CF), and "I liked that you took a position in your conclusion" (WBF).

While the differences for *Supportive Feedback* for both male and female ethnicities were small it is possible that these can be attributed to expectancy effects on the basis of ethnicity since the scores in the control condition did not indicate any marker variability in this domain.

There were no discernible differences at sub-theme level for either male or female ethnicities in relation to aspects of *Developmental Feedback*. However, overall CFs gained less *Developmental Feedback* than other female ethnicities. In the control condition each group of markers provided six developmental comments. In the experimental condition WBFs gained five comments, AFs gained four and CFs only gained two. The two comments for the CF fell into the *Developmental*

Feedback: Advice sub-theme and said “I think your work would benefit from wider reading” and “Check grammar and spelling – proof read”. White British and AFs also received similar advice-based feedback, but they benefited from also receiving comments coded as *Developmental Feedback: Information* which provided additional information on how they might improve their work. Therefore while the numbers are small, White British and AFs were provided with more developmental feedback opportunities than their CF counterparts.

4.6.3 **Summary Feedback Differences: Gender and Ethnicity**

In terms of positively phrased *Descriptive Feedback* there were differences apparent in the following groups. AFs gained five more comments in this domain than WBMs. However, the feedback content was largely similar consisting of comments related to structure, writing style, the ability to construct argument, and provision of a strong conclusions.

AMs gained fewer comments than WBFs (thirteen vs. seventeen). Although there were more similarities than differences in feedback content, it was apparent that AMs attracted more comments related to structure (four versus one) and WBFs attracted more comments related to writing style (six versus three).

The only noteworthy differences for negatively phrased *Descriptive Feedback* related once more to the CM. In the control condition scores were similar between the CM and the AF (seven versus six), but in the experimental condition the CM only attracted two comments while the AF gained seven. As has been previously discussed in the gender analyses, the content of the CMs feedback pertained solely to referencing issues. In contrast, the AF gained a wider range of feedback related to clarity of argument, and writing style. Similar differences were also apparent when the CM was compared to the WBF. Although the differences in the number of comments received cannot be considered reliable for this comparison (since there were also differences evident in the control condition), it was apparent that feedback content was also more varied for the WBF and also related to writing style, clarity of argument, and making assumptive comments.

Neither male nor female Chinese names received any praise-related *Supportive Feedback*. Differences were observed between the CF and the AM who received three comments and the CM and the WBF who received six. Previous analyses have demonstrated that the WBF also outscored their male counterparts in relation to praise-related *Supportive Feedback*. The results for the CF are tempered slightly by the fact that this essay gained three comments in the *Supportive Feedback: Encouragement* sub-theme and therefore the work was not totally devoid of supportive comments. In contrast, the CM only gained one supportive comment across both sub-themes. While once again the numbers are small, they do indicate expectancy effects at work on the basis of student gender and ethnicity in this instance. Given the important role that positive and supportive feedback has been documented to play in enhancing student confidence and motivation (e.g. Hyland, 1998; Pitt & Norton, 2016) it is a pity that some student names appeared to attract more of this type of feedback than others.

4.6.4 **Summary of Summary Feedback Results**

Gender differences which indicated expectancy effects at work were as follows. The *Positive Descriptive* feedback received by WBFs contained comments pertaining to more high-level academic skills than those provided to WBMs. Additionally, WBFs gained more *Supportive Feedback* comments related to *Praise* than their male counterparts. CMs received fewer *Negative Descriptive* feedback comments than CFs and these comments were narrow in range, solely consisting of feedback on academic writing skills.

Furthermore, within groups and across conditions differences demonstrated that when markers provided feedback for non-White British students they commented positively on aspects of structure. For Chinese students of both genders this also extended to them receiving positive comments on writing style. Both these types of feedback were absent when participants marked work perceived to have been written by a WBM.

Ethnicity-related differences which demonstrated expectancy effects in operation were visible for the CFs who gained less *Instructional Autocratic* feedback than their female peers, failing to attract a single comment in this area. CFs also attracted less *Developmental Feedback* overall than White British and AFs.

AMs gained more *Supportive Feedback* than WBMs and the feedback was more detailed. Conversely, AFs gained the lowest amount of *Supportive Feedback* among the female ethnicities and their comments were more generic in nature than those provided to their female counterparts.

Additionally, within groups and across conditions differences revealed that markers who went onto mark either a CM or an AF provided more *Positive Descriptive* comments for these students than they did for the WBM in the control condition. Furthermore, a broader range of positive were identified for these students.

Gender and ethnicity differences showed that the AM who had previously received more *Supportive Feedback* than the WBM now received less *Positive Descriptive* feedback than the

WBF. The AF who had previously received less *Supportive Feedback* than both other female ethnicities did receive more *Positive Descriptive* feedback than the WBM.

In terms of *Negative Descriptive* feedback the CM gained fewer comments and less variety of comments than the AF and fewer and less varied comments than the WBF. Furthermore, no essay bearing a Chinese name attracted any praise-related *Supportive Feedback.* Differences were apparent when the CM was compared to the WBF and when the CF was compared to the AM.

What became apparent through analysing the summary feedback was that groups of markers tended to mark very differently from each other even when provided with marking criteria. It was clear that marking groups included some participants who always commented on the ability to construct argument or writing style for example, which underlined arguments related to markers having their own agendas and not marking in a social vacuum (Price, 2005; Shay, 2005, 2008; Tuck, 2012). The repercussions of such variability for this thesis were that expectancy effects related to student gender, ethnicity, and gender and ethnicity were rarely evident in the summary feedback comments. The vast differences in the feedback provided in the control condition (when the student name remained the same) frequently determined that any differences in the experimental condition could be attributed to expectancy expects. Instead differences in the experimental condition were most likely to be reflective of differences in marking practice.

4.7 Reflective Comments

At the end of each essay participants were asked what factors had influenced their perceptions of the work. A range of comments were made which indicated a variety of expectancies at work. For example, "…what I expected from an 'average' L4 student.", "The style of writing and clearly laid out introduction suggested the student's essay would be of a high quality", and "The previous essay enabled me to form a better judgement of this essay. Order definitely played a role". However, it is comments that pertain to gender or ethnicity that are of most interest here since these would indicate that, a) markers have consciously processed the student name, and b) this may have played a role in the feedback provided.

In the control condition one participant explicitly stated that their perception of the work had been influenced by the student's gender, "… knowing it was a male student influenced my perception (not expecting it to be as 'polished' as a female student essay)." Another participant made specific reference to the student's name in relation to the characteristics they perceived an individual with such a name to have.

> "The student's name. Samuel (being a biblical name) immediately (to me that is) gives the preconception that the student is diligent and likely to produce a good piece of work".

In the experimental condition three participants made reference to either the male or female Chinese name in relation to their expectations about the quality of the student work. In relation to the male name one stated, "Chinese name – I expected there to be more issues with the academic writing, spelling and grammar etc." In reference to the female name, one participant noted "INTERNATIONAL STUDENT: Impressed with English writing skills". Another explained,

> Students name. Being of a non-British origin, one immediately makes assumptions on the quality of written English in the essay, when in fact it is arguably better than in the preceding student's work.

These results demonstrate evidence of halo effects at work since markers have used knowledge about the trait of an individual (e.g. Chinese) and then inferred other ambiguous information about them as a result of their cognitive biases. These results therefore support earlier research into halo effects within educational settings (Dennis, 2007; Forgas, 2011; Malouff et al., 2013).

While these comments only account for five participants out of sixty, they do serve to illustrate that markers bring their own frames of reference to the marking process and that marking is a ultimately a social practice (Price, 2005; Shay, 2005, 2008; Tuck, 2012). Shay (2008) noted that judgement-making practices required in marking were both habitual and deeply internalised. Therefore tutors engage in marking processes in a purposeful but not always conscious way. She further noted that this lack of conscious awareness regarding categorical thinking meant that markers are often unable to articulate how they come to make certain judgements over others. In light of Shay's (2008) work it therefore seems likely that more than 5 participants will have noticed and been influenced by seeing the student name on the assignments. They may not have been consciously aware that this played a role in the feedback they provided, but this does not guarantee that it did not. Furthermore, even if more participants did recognise and process differences in their marking, the potential for social desirability to play a role in their responses was quite high given the sensitive nature of the topic area. Given these challenges perhaps gaining 5 explicit responses outlining that expectancy effects related to the student name played a role was more revealing than first thought.

5 GENERAL DISCUSSION

The purpose of this thesis was to explore expectancy effects and bias within marking processes. Using a novel method, it specifically sought to explore whether the marker knowing the student name would bias the in-text and/or summary feedback provided. In response to NUS claims (1999, 2008, 2012), that University marking processes were biased according to student gender and ethnicity the following research questions were explored.

Do expectancy effects as primed through knowledge of these characteristics impact upon feedback in a way that suggests biased practice:

i) Student gender
ii) Student ethnicity
iii) Student gender and ethnicity

Traditionally, bias in marking has been explored through grades rather than feedback. Student gender has been the main lens through which marking bias has been examined (e.g. Bradley, 1984; Goddard-Spear, 1984; Newstead & Dennis, 1990; Dennis & Newstead, 1994; Hinnerich et al., 2011; Breda and Ly, 2014; Krawczyk, 2018), while research related to ethnic bias has been both meagre and largely school-based (e.g. Ouazad, 2009; Van Ewijk, 2010; Kiss, 2013; Sprietsma, 2013; Hinton & Higson, 2017). Furthermore, very little research has explored the interactive effects of gender and ethnicity (e.g. female *and* Asian), despite the challenges of multiple category memberships first being identified over seventy years ago (Ashe, 1946). Furthermore, much of the research conducted in the area of bias in marking has lacked experimental rigour (including not using real teachers to mark work and not using authentic assessments); has not controlled for confounding variables such as marker variability; and has not supplied marking criteria to participants.

Additionally, educational research more broadly has been criticised for omitting to include a sound theoretical underpinning (Tight, 2004; 2014; Shay 2009), and this criticism holds true for research on marking bias. Often researchers have made unsubstantiated claims that expectancies have biased marks without explaining expectancy effects or detailing how and why those effects might have arisen (e.g. Newstead & Dennis, 1990; Dennis et al., 1996). Given that May contends:

> The idea of theory, or the ability to explain and understand the findings of research within a conceptual framework that makes ‘sense’ of the data, is the mark of a mature discipline whose aim is the systematic study of particular phenomena (2001, p.29).

it would appear that, despite its long history, research on marking bias has yet to reach maturity. Furthermore, given these identified failings it is of concern that this research has been used to place pressure on HEIs to change assessment policy and practice and move towards anonymous marking for all written assignments.

An awareness of the limitations of previous research helped to ensure the originality of this thesis in terms of its focus, experimental design, and theoretical underpinning. It was hoped that the findings would, a) outline the impact of expectancy effects on feedback practice in relation to student gender and ethnicity, b) feed into the anonymous marking debate, and c) impact future policies on anonymous marking.

This chapter will now explore some reflections on the feedback findings more generally with particular attention being paid to the role marker variability played in the process. It will then examine some of the key findings in relation to both gender, ethnicity and the combined effects of gender and ethnicity before commenting on what these results tell us about the presence of feedback bias overall. Results will then be critically interpreted in line with relevant theories and models before examining strategies to address expectancies and implicit bias.

5.1 General reflections on the feedback

Worryingly markers frequently made spelling mistakes, and summary comments were often poorly written and lacked clarity. Additionally, despite the literature being replete with evidence that feedforward is critical for deep learning (Higgins et al., 2002; Hattie & Timperley, 2007; Boud & Malloy, 2013), there was not a single example of this type of feedback throughout the 120 essays. Although recent research has suggested that students recall future-oriented feedback (or feedforward) less well than past-oriented feedback (Nash et al., 2018) this currently remains an isolated finding, and therefore more evidence is required before the merits of feedforward are debunked. Positive feedback comments were also rare for both in-text and summary feedback. Furthermore, when positive feedback was supplied it was largely symbol-based (in the form of ticks), which may stimulate the desired emotional response in a student, but is also ambiguous and lacking in educative function. Given that positive feedback has been described by students as both effective to their learning and motivational (Lizzio et al., 2003; Pereira et al., 2016; Pitt & Norton, 2016), this finding was particularly disappointing.

The tendency for markers to provide indeterminate types of feedback was further evident in the amount of emphasis-based feedback provided (e.g., underlining text, circling a word). While the use of ticks lacks a sound pedagogical underpinning and is undervalued by students (Price et al.

2010; Ferguson, 2011), it does at least inform them they have done something right - even if they remain unsure what that something was. Comparatively, emphasis-based feedback is more confusing, since students might not be sure if the requisite text was identified because it was a good or a bad example.

In summary comments, markers were most inclined to provide descriptive feedback, followed by autocratic and developmental feedback. While evidence of developmental feedback comments was pleasing, since students perceive such feedback to develop their understanding and be future-focused (McLean et al., 2015), the discovery of autocratic feedback was less welcome. Parallels can be drawn between autocratic feedback and the use of imperatives in feedback comments which reinforce the power at play and inhibit learning (Lea & Street, 2000). Furthermore, it chimes with McLean et al.'s research which found that students often experience feedback as being "… about telling, transmitting, being told" (2015, p.925). They identified such feedback to be unidirectional, focused on the present rather than the future, and having an external agency which emphasised the tutor as expert and failed to engage the student in the process.

Stylistic comments constituted the majority of feedback. This included feedback on essay mechanics such as punctuation, grammar, referencing and presentation. This finding supports earlier research (e.g., Stern & Soloman, 2006), but was disheartening since stylistic feedback is a barrier to effective learning and students have being critical of its utility (Ferguson, 2011; Li & Deluca, 2014).

Overall these findings replicate research which has been critical of feedback for its corrective and error-focused nature and its failure to move beyond description and be developmental and future-oriented (Orrell, 2006; Walker, 2009; Orsmond & Merry, 2011). Therefore it falls short of delivering the written feedback that students prefer. These include, feed forward to help with future assignments, personalised feedback, developmental and encouraging feedback, suggestions about where to seek help, explanations about why the grade was appropriate, and encouragement (Lizzio & Wilson, 2008; Ferguson, 2011; Li & DeLuca, 2014; Winstone et al., 2016).

Consequently, the feedback found in this thesis largely failed to match student needs. This poses broader questions about student engagement with and responses to such feedback. The way that students respond to feedback is complex and related to their emotional maturity (Nash et al., 2015; Pitt & Norton, 2016). Given that the essays in this thesis were identified as being those of a first year student, the notion of emotional maturity has increased relevance. When markers were asked what influenced their perceptions of the work, many noted the year of the student, and claimed to have adjusted comments and grades accordingly. Unfortunately however this awareness did not culminate in the provision of the more emotionally supportive feedback necessary for first year students (Poulos, 2008), and therefore paints a negative picture regarding the likelihood of students engaging positively with it. It does however support Read et al.'s (2005) research which highlighted that even when tutors were aware that students might be lacking in self-confidence they did not provide more positive feedback to such students. Furthermore, it is important to recognise that when people enter new situations (such as University), their sense of self-concept is weakened and they are more prone to expectancy effects in the form of self-fulfilling prophecies (Jussim & Harber, 2005). Therefore while it is worth remembering that students' needs are multiple and varied (Hepplestone & Chikawa, 2014), it is also pertinent to consider the vulnerability of first year students to fulfilling the prophetic effects of their feedback. This is especially so given that research has identified first year students as having higher levels of engagement with feedback than their second and third year counterparts (Ali, Rose, & Ahmed, 2015).

5.2 Marker Variability

The assessment pack provided to each participant included marking criteria. The intention was to help standardise the marking process and guide areas of feedback provision (e.g., structure, ability to create argument, quality of examples etc.). In turn this would provide more confidence that any differences in feedback could be attributed to expectancy effects arising as a result of knowledge of the student name. Nonetheless, one of the most startling findings was the huge variability evident in marking practices which made it much more difficult to establish expectancy effects. This indicated that markers may pay little attention to the marking criteria. Instead perhaps they relied upon the tacit knowledge or connoisseurship that has previously been identified as interfering with the efficacy of criteria (Woolf, 1995; Ecclestone, 2001; Knight & Yorke, 2003; Sadler, 2005; Shay 2005, 2008).

Variation in marking became apparent in a number of ways. There were huge in-text feedback differences in the control condition when the student name remained the same. The experimental design necessitated that groups of markers needed to be marking similarly in the control condition (when marking the same essay written by a student with the same name), for any differences between-groups of markers in the experimental condition (when marking the same essay but by different student names to each other), to be attributed to expectancy effects.

Therefore when participants marked differently to each other in the control condition this reduced the opportunities for expectancy effects to be identified in the experimental condition.

Finding variations in in-text feedback within the control condition was compelling because it indicated other influential factors were at work aside from the expectancy effects primed by the student name. Indeed it is necessary to consider the name as an important, but not unique variable and accept that expectancies can be formed and take effect using other information. For instance, the knowledge that the essay was first year work, or the perceived calibre of the University the work was completed for. Evidence of norm-referencing has also been found to impact on the consistency of marking (Shay, 2004; Orrell, 2008; Bloxham, 2009). This occurs when a student's work is judged alongside that of another student instead of on its own merits. Given that participants marked two essays, this was a possibility.

Additionally, there is evidence that marking and feedback can be impacted by marker's motivation. High motivation stimulates more data-driven (and therefore accurate) judgements (Pendry & Macrae, 1996) as long as cognitive load is not too high (Biesanz et al., 2001). Given that participants knew their marking was to be included as part of a research project they may have been highly motivated, processing information and providing feedback in a more meticulous way. On the other hand, the drive to form accurate expectancies is also influenced by the interdependence between the perceiver (marker) and the target (student) (Jussim, 1993, Neuberg & Fiske, 1987). Since participants did not know the students and did not expect future interactions with them, their motivation to invest high cognitive effort might have been lowered.

Furthermore, mood state can influence the information processing strategies perceivers use. Specifically, perceivers in a bad mood are more accurate and process information in a more individuated way than people in a good mood (Bodenhausen, et al., 2001; Forgas, 2007; Forgas & Laham, 2009). Similarly people's circadian rhythms (whether they are morning types or evening types), impact on their processing choices and use of categorical thinking, with more bias evident during their least favourite part of the day (Bodenhausen, 1990). Resultantly, the disparate feedback provided by markers in the control condition could be attributed to a number of variables and these may play as much, if not more of a role than expectancy effects generated by the name.

The variation found in the control condition therefore suggests against viewing anonymous marking as a panacea which would eradicate variability and simultaneously rid the marking process of expectancy effects and bias. Moreover, it helps to explain research findings which have reported that student performance remained unaffected when anonymous marking has been introduced (Owen et al., 2010; Hinton & Higson, 2017; Pitt & Winstone, 2018). While these papers, (and to some extent the findings from this thesis), provide evidence that the student name may not be the catalyst for expectancy effects and bias that some have suggested, it is also important to recognise that bias can manifest itself in other ways. Therefore despite its centrality to the anonymous marking debate, the student name may only be the starting point in the explorations of those who wish to reduce bias. In short, it would be premature to celebrate such findings as evidence that bias does not exist but rather accept that we may simply have been looking in the wrong places.

Another variation apparent within the marking process was that fifty out of 120 participants failed to provide any summary feedback. This made comparisons between certain groups impossible and the subsequent analysis incomplete. Participants were asked to mark, "…in line with current teaching practice", which suggests that almost half of them do not usually provide summary feedback. Given that students assign the most value to written summary and in-text comments this is particularly concerning (Ferguson, 2011). Moreover, when both in-text and summary feedback were provided there were huge inconsistencies in quantity. Summary feedback ranged from two words to 350, and in-text feedback from 219 to 343 per script. Such divergent feedback practices are unsurprising, having being been noted in large scale projects exploring students experiences of feedback (Jessop, El Hakim & Gibbs 2014; Jessop & Tomas 2017). However this variation may well fall short on meeting perceptions of fairness that students have identified as being important (Lizzio & Wilson, 2008).

Finally, this thesis initially aimed to explore expectancy effects between groups and within conditions in an attempt to account for marker variability. For example, to compare markers in Group 1a with markers in Group 2a in the control condition and determine whether they marked the same essay in the same ways when the student name presented to them was the same. Subsequently, if their feedback practices altered in the experimental condition there could be more confidence that this was not a result of marker variability (since this had previously been controlled for), but was a result of expectancy effects based on knowledge of the student name. However, huge variations in practice within marking groups (i.e. Group 1a in the control condition versus Group 1a in the experimental condition) were noticed when analysing the in-text feedback. For instance, a group of markers who had previously shown no inclination to comment on the structure of the work when marking a WBM (in the control condition) made numerous comments on this when marking work of a non-White British student (in the experimental condition).

Nonetheless, when comparing within groups and across conditions the comparisons being drawn are across two different essays which may account more for any feedback differences found than the student name. As such experimental control has been compromised. However, on many occasions it was evident that such changes existed only for one group of markers. To extend the previous example, while the groups of markers who went onto mark an Asian or Chinese name in the experimental condition were suddenly motivated to make numerous comments on the structure of the work, the group of markers who went onto mark the same essay but bearing a White British name in the experimental condition did not. This comparison therefore lends weight to the idea that expectancy effects may have been more responsible for these differences than marker variability. Nonetheless, these unexpected findings do muddy the waters since the likelihood is that marking is prone to both expectancy effects and inconsistency at both inter and intra-individual level.

The variability found in marking processes supports findings from a comprehensive review on feedback practices (Li & Deluca, 2014). Nonetheless, this is far from comforting since it points to there being massive problems with marker variability which may be an even bigger issue than expectancy effects. One appealing explanation for such variability is housed in a growing body of literature which examines marking as a social process (Connell et al. 1992; Mutch, 2003; Shay, 2005; 2008; Tuck, 2012). Proponents of this school of thought would be unsurprised by these divergent marking and feedback practices and would suggest they are simply, "… the inescapable outcome of the multiplicity of perspectives that assessors bring with them" (Shay, 2005, p.665). Using Bourdieu's theory of social practice, Shay (2005) explains that while there is a shared consensus or perceptual framework within the academic community which is used when marking student work, there are also a multitude of different interpretations which exist. Consequently, a "double truth" (Bourdieu & Wacquant, 1992, p.255) exists, where two modes of knowledge, the objective and the subjective, overlap. In other words, markers bring both objective and subjective criteria to the reading of a text. For this reason such scholars are sceptical of the efficacy of objective measures such as paper trails of marking criteria and the calibration of standards, which are considered integral to robust assessment practices. Instead, they believe that we should free ourselves from a longstanding "… collusion with the myth of objectivity" (Shay, 2005, p.676) and only use these measures to stimulate dialogue about assessment and practice.

Crucially however, it is the subjective element of marking and feedback that makes it vulnerable to expectancy effects and subsequent biases. Therefore it is no surprise that those who acknowledge assessment and feedback as social practices have also explicitly acknowledged that people engage in these in a less than conscious way (Shay, 2008). However, although pedagogical research has alluded to the role that implicit processes play in marking processes, and even used expectancy effects as an explanation for these (Dennis et al. 1996; Malouff &Thornsteinsson 2016), it has spectacularly failed to engage with the wealth of social and cognitive psychology literature pertaining to expectancy effects and bias. This seems surprising when such theories and concepts can provide a rich avenue to help researchers understand the role that these effects might play in this social process. While this literature might only offer a lens to help understand part of this process, it would provide the field with a greater awareness of what is happening in the minds of lecturers as they mark student work. In addition, the research provides strategies aimed at reducing expectancies and biases (i.e., stereotype suppression, anti-bias workshops, Implicit Association Tests [IATs]), and therefore the potential to lessen their role within the social process also becomes possible.

5.3 Expectancy Effects

The work of Olson et al (1996) and Warr and Knapper (1968) have been useful to apply a theoretical underpinning to the role that expectancy effects might play in marking and feedback processes. Expectancy formation happens quickly and paves the way for subsequent expectancy effects and bias (Olson et al., 1996). Even when interactions are brief, the judgements made about others can have long-lasting effects (Fiske & Taylor, 1991; Jussim & Harber, 2005). Importantly for this research, principles of expectancy formation do not rely on the perceiver (marker) being in the presence of the target (student) for them to be influential (Olson et al., 1996; Eysenck, 2009). As such the static cue of a student name was an appropriate choice of priming stimulus to activate schemas in the minds of the markers and trigger expectancy processes.

Given the importance of the student name, an attempt was made to gain some measure of its influence on the markers. However, asking an explicit question about the influence of the name might have been ineffectual, since expectancy formation and effects are often deeply internalised or unconscious (Bargh, 1997; Macrae & Bodenhausen, 2000). Therefore even if the student name was influential participants may not be aware of it or be able to articulate its impact. Furthermore, even if category activation is viewed as a conscious process, it is unlikely that participants would admit that the student name had impacted on their marking due to self-presentational concerns (Harber, 1998; 2010). Indeed expectancies are reactive to measurement because people consciously manage their attitudes when it is in their interests to do so (Fiske, 1987). Therefore, in order to ascertain whether the student name had been cognitively processed and had influenced expectancies, a more general question inquired about factors influencing participants' perceptions of the work.

Of the three antecedents to expectancies identified by Olson et al. (1996), two had the potential to influence the expectancy formation process. Since participants had no *direct experience* (prior knowledge) of the target, any expectancies were formed either by *indirect experiences* they might have had with a specific type of student in the past, or *other beliefs* which do not require experience of a target and relate to a set of beliefs perceivers construct for themselves (i.e. this is written by an Asian student. English might be their second language, so I'll give them the benefit of the doubt). Similarly, when considering Warr and Knapper's (1968) model, two of the three information processing sources were available. Participants could use *present stimulus person information* (i.e. name, gender, ethnicity) and *present context information* (i.e. first year undergraduate level assignment), but not *stored stimulus person information*.

Using Olson et al.'s (1996) model, it is clear that participants who commented on the student name had formed their expectancies on the basis of either indirect experience or other beliefs. For these participants the expectancies formed were likely to have certain properties (which serve to determine the expectancy response). The most relevant property here is accessibility, which refers to the ease with which an expectancy comes to mind (Olson et al., 1996). Expectancies which have been recently formed are highly accessible and therefore more likely to be used to make sense of future interactions. Hence the presentation of the student name immediately prior to reading the work fulfilled the accessibility criteria. The property of importance is also worth considering since it indicated that markers who commented on the student name lacked the motivation to form an accurate expectancy of the student's work and therefore invested less cognitive effort in the interaction (because the task was not important enough). Consequently, they are likely to have relied upon schema-driven processing, using the student name to guide their expectancies as opposed to the content of the work itself. According to Olson's et al.'s (1996) framework this makes these markers more prone to expectancy effects and bias.

In accordance with Warr and Knapper's (1968) model, participants who identified that the student name helped form their perceptions possessed certain stable characteristics (e.g., beliefs, attitudes) and temporal states (e.g., mood, motivation). These co-determined that the information (the student name), was processed through the input selector and became a source of information on which decisions were made, inferences drawn, and responses shaped. Participants influenced by the prime of the student name were processing information schematically since they had assigned the student to a category in the early stages of the interaction.

Once an expectancy has been formed the expectancy response follows. The labels differ, with Olson et al. (1996) claiming the categories to be, cognitive, affective, and behavioural and Warr and Knapper (1968) affective, attributive, and expectancy. Despite these differences, evidence of name-based expectancy responses were found in the comments provided by participants. For example, an affective response was evident when one participant claimed they were, "… impressed with the English writing skills" of an essay labelled as written by a Chinese student. Attributive responses refer to perceptions of a target's dispositional characteristics and were evident when a participant noted that they had expected an essay written by a male student to be "… less polished" than that of a female. Another noted that the name Samuel conveyed the expectancy that the student would be "… diligent, and likely to produce a good piece of work" due to the religious connotations associated with it. Furthermore, several participants suggested that Chinese students would struggle to produce good quality academic writing.

Evidence of such beliefs demonstrate halo effects at work. Given the evidence of halo effects upon marking practices (e.g., Dennis, 2007; Forgas, 2011; Malouff et al., 2013) it is very likely that they will have shaped markers feedback. Perceptual and cognitive biases impact how perceivers remember, interpret and explain events in expectancy-confirming ways (Darley & Fazio, 1980; Miller & Turnbull, 1986; Jussim, 1989, 1990, 1993). Furthermore, expectancies have previously influenced the grades awarded on student work (Diederich, 1974; Dennis, 2007; Malouff et al., 2013) with popular student names gaining higher marks (Harari & McDavid, 1973; Erwin & Calev, 1984).

Nonetheless, the number of participants who identified that the student name influenced their perceptions was only eight percent. One explanation for this is that participants were not under normal time pressures to return the work. Category application is most likely to occur when a perceiver lacks time or cognitive capacity to think deeply (and accurately) about others (Bodenhausen et al. 1999; Brewer & Feinstein 1999; Fiske et al., 1999), and judgement then becomes more stereotypic (van Knippenberg et al., 1999). Therefore perhaps in the absence of marking deadlines most participants were inclined to use data-driven processing and relied less on the student name as a trigger for expectancies.

Alternatively, the low percentage might be explained through the concept of dual processing. Fiske and Taylor (1991) would interpret these results as evidence that participants were 'motivated tacticians' who switched between schema and data-driven processing at will according to their motives, needs and goals. Goal states of the perceiver have previously been identified as influencing information processing choices (Blair & Banaji, 1996; Macrae et al. 1997; Spencer et al. 1998). Therefore participants' knowledge that they were marking work as part of a research project might have motivated them to employ strategies which maximised chances of a more accurate outcome. Accordingly, even if participants began processing information schematically and activated a category based on the student name they would be able to avoid applying that category by switching their information processing style to be more data-driven.

Therefore participants had information available to them in addition to the student name which might be similarly expectancy-inducing. Expectancies might have been influenced by marker's mood state or circadian rhythms (Bodenhausen, 1990; Bodenhausen, et al., 2001; Forgas, 2007). There are also a host of personal biases that individuals bring to the process. For example, one marker may be a pedant for clarity of writing while another abhors poor referencing. Consequently participants might still have been relying on expectancies and processing data schematically, but using criteria other than the student name.

However, the waters are further muddied by the unconscious nature of processing (Bargh & Pietromonaco, 1982). Although participants were explicitly asked to reflect upon what had influenced their perceptions of the essay it is possible that the name and ethnicity of the student could have made an impact without their conscious awareness. As Macrae and Bodenhausen remind us;

> Although it may typically seem as if we are consciously directing our own behavior, the reality of the situation is that frequently we are not. Instead, many of our complex social actions have their origin in the impenetrable and silent workings of the unconscious mind (2000, p.107).

Despite using gender and ethnicity which are 'privileged' categories (Fiske et al. 1996) with high activation potential it is therefore likely that many more markers noticed and were influenced by the name of the student. However because this happened unconsciously they were unable to recall this.

5.4 Key Results

One finding germane to all analyses was that feedback often looked similar at the macro level (e.g. the amount of feedback contributions and the patterning of the feedback landscape).

Therefore scrutiny of the feedback at category level did not always reveal differences. These tended to be clearer at subcategory and further subcategory levels. For example, levels of content-related feedback often looked similar, but when broken down into comments or symbols it became apparent that the bulk of content-related feedback for non-White names was simply tick-based. This suggests that future analyses of feedback will need to be thorough if they are serious about exploring these small, but potentially influential differences.

5.4.1 **Key Results: Gender**

Although findings related to marker variability reduced the capacity to examine expectancy effects with real breadth and clarity, some interesting results remained. The most substantial differences for gender were found when White British males (WBM) and White British females (WBF) were compared. There were several stylistic differences in in-text feedback. More fundamentally however, WBFs gained less positive feedback, more negative feedback, and less developmental feedback than WBMs, demonstrating what might be considered an unhappy triad of feedback. Particularly in terms of developmental feedback, WBF students were asked fewer *reflective questions* and received fewer *alternative* comments than WBMs. Given the developmental nature of feedforward (Hyatt, 2005), and research which highlights that positive feedback enhances motivation and esteem, and moderates the effect of negative comments (Hyland, 1998; Lizzio et al., 2003; Shields, 2015), expectancy effects may have favoured WBMs in this instance. These findings might also provide an indication of the expectancies markers held about the potential of the students, since historically high expectancy students have been provided with more positive feedback than low expectancy students (e.g., Brophy, 1983; Cooper, 1979; Jussim, 1986). Furthermore, they add weight to the perception by WBF students in Pitt and Norton's (2018) research that marking anonymously is fairer.

Nonetheless, these in-text findings contradicted those found in the summary feedback for White British students. Here markers provided more *Supportive Feedback* related to *Praise,* and more *Positive Descriptive* feedback to females. Moreover, the *Positive Descriptive* feedback provided to females was more extensive and related to higher-level academic skills (i.e. the ability to develop argument). Therefore, on this occasion WBFs have arguably gained an advantage over their male peers. Here the WBF would be considered the high expectancy student since she was provided with more feedback (e.g., Brophy, 1983; Cooper, 1979; Jussim, 1986) which was emotionally supportive (e.g. Jussim, 1986; Rubovitz & Maehr, 1973).

Finding in-text feedback themes which were disconnected to summary feedback themes was not an isolated occurrence. For example, in the female ethnicity analysis Asian females (AF) were provided with more ticks than both other female ethnicities. However, they gained the least amount of *Supportive Feedback* compared to their female peers in the summary text*,* and their content also failed to move beyond the generic. Such differences provide students with mixed messages about the quality of their work. On the one hand, the AF name received many ticks, but on the other, received little praise or encouragement for those efforts in the final comments. This feedback falls short in meeting one of the key principles of good feedback which is to, "… help clarify what good performance is" (Nicol & McFarlane-Dick, 2006, p.203). Students' understanding of the feedback they receive is often modest (Mutch, 2003; Carless, 2006), but of primary importance for their personal engagement and subsequent behaviours (Higgens et al., 2001; Pitt & Norton, 2016). Therefore the provision of contradictory feedback simply adds ambiguity to what is already a complex task for students.

Evidence of differences between feedback types has been found elsewhere. Kumar and Stracke (2007) found that in-text comments included more referential (i.e., presentational and content), and directive (i.e., suggestions and instructional) feedback, whereas summary comments were dominated by expressive feedback (i.e., praise, criticism, and opinion). These differences may be explained through research exploring analytic and holistic marking. Sadler (2009) explained that markers judgments were often contradictory such that holistically they may have judged a piece of work as outstanding but when they referred back to the analytic criteria it only looked mediocre. Sadler (2009) claimed that experienced markers attributed these discrepancies to their judgments being influenced by criteria not present on the list. In a similar way markers here appeared to be making a series of small qualitative judgments in-text in a more atomised or analytic way, but then writing holistic feedback which did not always correlate with the analytic judgements. Future research may wish to explore where students' attention falls when they receive feedback (i.e., do they pay more or less attention to in-text versus summary feedback), as this may have implications for how feedback is received, interpreted and used.

Further gender-based results only found small in-text differences for Asian names. For example, AFs gained fewer than half the comments on the content but received more symbol-related (i.e. ticks) feedback in this domain. Mirroring the results for WBFs, AFs also gained fewer *alternative* comments than their male peers, but did receive more feedback on *reflective questions* a trend which reversed the finding for WBFs. No summary feedback differences were found for Asian students

Limited in-text feedback differences were also identified for Chinese students. However, male students were provided with more *stylistic corrections* related to referencing and less structural feedback. There was only one difference in the summary feedback which demonstrated that Chinese females (CF) received more *Negative Descriptive* feedback than Chinese males (CM). *Negative Descriptive* feedback also pertained to a wider number of academic skills for females, whereas for males comments largely focused on the quality of academic writing. While negative feedback has sometimes been identified as useful and motivational by students (Winstone et al., 2016; Pitt & Norton, 2016), it might be more damaging to ethnic minority students who already lack trust in educational establishments and are prone to disengagement (Croft & Schmader, 2012).

The data represented fewer gender differences for Asian and Chinese students. However, rather than accurately reflecting reality this might be a limitation of the research. It is possible that participants were unable to identify student gender from the names provided on the Asian and Chinese essays. While there is research showing that certain names sound more masculine or feminine than others (Van Fleet & Atwater, 1997), and therefore participants might have accurately guessed student gender, it is also likely that they were less confident in doing so than with White British names. As such category activation on the basis of gender may not have occurred, thus the potential for expectancy effects and bias to materialise was attenuated. The results of the ethnicity-based analyses support this argument, since Asian and Chinese names did attract different feedback when they were compared to names from different ethnic groups. In fact, these names were most responsible for the differences reported here. This notwithstanding, the WBM versus WBF was the most reliable comparison of gender differences.

Although results were contradictory across feedback types, there were differences evident which generally disadvantaged female students. This suggests that feedback provision was prone to expectancy effects and bias on the basis of gender. The differences may not be large, or wholly due to the prime of the student name, but they do run counter to more recent research suggesting no evidence of gender bias on the basis of grades (Van Ewijk, 2011; Sprietsma, 2013; Krawcyzk, 2017; Pitt & Norton, 2018). Clearly further research is required to measure both their pervasiveness and additional causes of such bias.

5.4.2 **Key Results: Ethnicity**

Both ethnicity analyses revealed that non-White British names received different types of in-text feedback. For males it was the Chinese name that was the catalyst for such differences and for females it was the Asian name. Specifically, for the male Chinese name, differences consisted of fewer comments and corrections on their writing style and presentation; more corrective feedback on referencing-related issues and punctuation; more content-related feedback which was symbol-based (i.e. ticks). Importantly, the CM also received the lowest amount of developmental feedback related to *alternatives.* Nonetheless, these differences did not transfer to the summary feedback, where no between-groups differences were found related to CMs. There was however a within-group difference which demonstrated that markers who went on to mark the CM name in the experimental condition, provided more *Positive Descriptive* comments than for the WBM control name. Furthermore, a broader range of positive elements were identified.

For female ethnicities it was the Asian name that emerged as receiving different types and amounts of feedback. Many of the differences were similar to those found for the CM name, since AFs received more corrective feedback on punctuation, more comments on referencing-related issues, and more symbol-based feedback on content (i.e. ticks). The limiting nature of such feedback is explored in the overall findings which follow.

Nonetheless, the feedback landscape was not always negative for non-White British students. There were occasional examples of the summary feedback being advantageous for them. For example, AMs gained more *Supportive Feedback* than WBMs and the feedback was more detailed. Furthermore, within-group differences revealed that participants who went onto mark either a CM or an AF provided more *Positive Descriptive* comments than they did for the WBM in the control condition. Additionally, a broader range were identified for non-White British names. One explanation for these findings resides in social psychology literature which demonstrates that White markers consistently apply a positive feedback bias to minority groups (e.g. Devine, 1989) to satisfy self- presentational (Littleford, et al., 2005; Harber et al., 2010) or sympathy motives (Jones et al. 1995). Malouff and Thornsteinsson's (2016) recent meta-analyses of grade bias in assessment has also shown that a reverse bias is possible for students of non-dominant ethnicities. The applicability of this concept is strengthened for this thesis because ninety-five percent of participants identified as White British. Evidence of this bias might also help to explain the multiple occasions that Asian and Chinese names received more ticks on their work than White British names. Educational research might interpret these summary feedback findings as evidence of sympathetic marking at work - a concept which has been used to argue for anonymous marking (Shay, 2008). Expectancy-based explanations emphasise that Whites have lower expectations of work produced by non-White students and therefore the positive feedback is simply a reflection of the markers surprise when their expectations are exceeded (Biernat & Manis, 1994).

While it might be considered advantageous to evoke a positive feedback bias this is a contentious issue. Despite the fact that such feedback might protect self-esteem (Shields, 2015), stimulate positive motivational beliefs (Thorpe, 2000) and be preferred by students (Lipnevich & Smith, 2009; Li & DeLuca, 2014), the opposing view is that a feedback withholding bias (where negative feedback is withheld from ethnic minority groups), prevents learning opportunities and impedes academic development (Harber, 1998; Croft & Schmader, 2012).

Despite in-text feedback demonstrating evidence of some bias according to ethnicity, summary feedback differences were infrequent and often contradictory. However, once again it was the non-White British names which generated these differences. Specifically, the CF gained less *Instructional Autocratic* feedback and less *Developmental* feedback than Asian and White British names which gained comparable amounts of each. The AM name gained more *Supportive* and detailed feedback than both the WBM and the CM, but the AF name gained less *Supportive* and more generic feedback than the CF and the WBF (who gained the most *Supportive Feedback*). This topsy-turvy collection of results provided little consistent evidence of either a positive feedback bias or expectancy effects being present within the summary feedback.

5.4.3 Key Results: Gender and Ethnicity

Multiple differences were found when interactive expectancy effects were explored, although only the most significant are discussed here. Developmental feedback related to *alternatives* showed that CMs received fewer comments than WBFs. When compared with other analyses it was evidence that the CM also scored lower in this category than both other male ethnicities, but that no differences existed when they were compared to a Chinese or Asian females. This might indicate that gender became the defining trait when CM were compared to Chinese or Asian females, but that ethnicity moved to the foreground when comparisons were made with other male ethnicities and White British females. In other words, gender was strong enough to advantage the CM against other minority groups, but not against the WBF. However, developmental feedback related to *reflective questions* failed to reinforce the idea that WBFs ethnicity operated as a protective factor since she scored lower than any other group. Finally, for all between groups analyses for which differences were found, *informational comments* showed greater amounts of feedback being provided to White British students of both genders. Therefore Whiteness once more appeared to outweigh gender as an influential factor in expectancy effects.

In-text stylistic feedback was extensive and led to numerous differences. Most significantly more punctuation errors were noticed and corrected for Chinese names as opposed to Asian or White British names. This suggests that interactive expectancies operated here for both gender and ethnicity. The CM was also treated differently in terms of *syntax/word order/grammar* in comparison to the WBF and other male ethnicities but not to Chinese or Asian females. This infers that when the CM was compared to other male names, ethnicity was the determining factor and gender moved to the background. However, when compared to the Asian or Chinese females neither gender or ethnicity was sufficient to evoke a change. Once again the privileged ethnic status of the WBF may have been a protective factor for her gender.

Content-related feedback illustrated that Asian females gained less useful types of content-related feedback than WBM and females of other ethnicities. No comparisons were available for AM and CMs. This showed that the AFs gender may only have become a central trait when compared to the WBM, but that her ethnicity was sufficient to elicit changes too, when compared to other female names.

Though these results are far from conclusive, they do at least represent an initial step towards understanding the interactive effects of expectancies on feedback. It is true that this set of analyses was sometimes contradictory. For example, WBFs gained less developmental comments related to *reflective questions* than both AM and CM, but more comments on developmental comments related to *information*. This lack of consistency made it difficult to draw concrete conclusions from the findings and once more illustrated how capricious marking can be. Future research needs to be conducted on interactive effects to demonstrate a recognition that expectancies do not simply operate on singular characteristics but are activated by multiple interacting sources of information available to the perceiver.

5.5 Key Results: Overview

Overall in-text feedback differences were associated with stylistic elements of the work, or what Harber (1998) referred to as essay mechanics. Non-White British names attracted considerably more feedback in this area. Receiving feedback on elements of punctuation, grammar, referencing, and presentation have potential to be useful, but only if it is substantial enough for students to understand and correct errors in future work. Therefore it is even more troubling that much of the stylistic feedback provided for non-White British names was corrective or emphasis-based and not comment-based.

Another consistent difference concerned content-related feedback. An overarching view of the in-text feedback indicated that non-White British names often received more of this type of feedback. However, more detailed analysis revealed that this additional content was not comment-based. Instead non-White British names often attracted more ticks on their work. This was the case for AFs when compared to WBMs, AFs compared to Chinese and WBFs, and CMs compared to Asian and WBMs. When compared to a non-White British name, the White British name almost always gained the fewest ticks. Interestingly, with the exception of Asian names, there were no meaningful differences found in the amount of ticks provided on student work when participants who marked non-White British names were compared (i.e. AM versus CF; CM versus AF; CM versus CF), demonstrating that the differences in ticks only became evident when a non-White British name was compared with a White British name. While the provision of ticks may act as positive reinforcement and can therefore enhance student's confidence and motivation (an issue identified as particularly important among ethnic minority groups in HEIs) (Cohen et al., 1999), their utility has already been questioned throughout this thesis.

The above differences reiterate that not all feedback is equal in terms of its developmental, informational, or educational potential, and therefore raises interesting questions about the equity of feedback more broadly. Furthermore, the contrasting feedback found contradicts Jussim's (1991) work. He contends that rather than being prone to stereotypes, expectancies, and biases, people are actually very accurate when judging others. If this was true identical essays should have produced near-identical (or accurate) feedback. This is particularly so for the more mechanical, objective, and unambiguous elements of writing such as grammar, spelling and referencing, since these are either correct or incorrect, and therefore subjectivity is reduced (Harber, 1998; Sadler, 2009). However, this was not the case. Instead differences related to mechanical aspects were observed on multiple occasions in the experimental condition after having remained constant (i.e. accurate) in the control condition when the name remained the same. This suggested that the student name did culminate in expectancy effects which led markers to be more or less accurate in their feedback.

Another interesting finding was that more evidence of expectancy effects were found for in-text feedback. While it is difficult to speculate on the procedures participants employed while marking, the concepts of analytic versus holistic grading may once more help interpret these results. It is possible that when composing in-text feedback participants marked in a more analytic way, making separate judgements on the marking criteria to help shape their perception of the work. This practice aligns itself with data-driven processing since each piece of information is systematically processed in an individuated way. This is said to attenuate expectancy effects and subsequent bias. Data-driven processing places more strain on the processing system, but perhaps because markers were working in close context with the text when they provided this type of feedback, they were more able to make such a cognitive investment. Conversely, when writing summary comments, markers are one step removed from the context of the text, and need to assimilate multiple pieces of information and criteria simultaneously before conveying a set of overall judgements about the work. This holistic grading allows markers to build up a complex mental response to the work. Nonetheless, this also places strain on the information processing system, and it is possible that due to the distance from the work markers are less able to assimilate all the information accurately. Under such conditions markers might be more likely to resort to categorical thinking and expectancies to formulate a judgement. Consequently the student work is likely to be processed more subjectively, with more implicit, emergent criteria being applied, in part based upon the person schema that has been activated. This holistic approach therefore mirrors that of schema-driven processing.

Appealing though this explanation might be, when interpreted in line with the results from this thesis it makes little sense. Expectancy effects and bias are presumed to reveal themselves more explicitly when schema-driven processing operates. However, fewer differences were found in the summary comments when this type of processing was hypothesised to be operational. Instead they occurred within the in-text comments when data-driven processing was more likely to dominate. One explanation for these contradictory findings is that the reflective component of holistic grading allows markers time to employ an inhibitory mechanism such as stereotype suppression (Bodenhausen & Macrae, 1998), or controlled inhibition (Devine, 1989), whereby schemas about the student are activated but not applied. However, notwithstanding the equivocal nature of research pertaining to the efficacy of such strategies, (Macrae et al., 1994; Macrae & Bodenhausen, 2000), these explanations require an acceptance that category activation is controllable and not automatic as many believe. Therefore an alternative explanation for the reduced evidence of expectancy effects in the summary feedback might once more be explained by the existence of a positive feedback bias or sympathetic marking (Harber 1998; Shay, 2008). It is therefore possible that participants felt they had been biased in their in-text feedback and subsequently used the summary feedback to resolve their cognitive dissonance and self-image concerns (Harber et al., 2010).

Taken collectively, neither type of feedback consistently demonstrated a bias towards one gender or ethnicity. Nonetheless, within some analyses there were differences that pointed to expectancy effects and biased feedback practice. Despite this, there were also feedback differences in the control condition when student names remained the same. This points to there being a host of confounding variables that contributed to these differences in addition to the name (e.g. mood state, motivation, circadian rhythms) (Bodenhausen, 1990; Bodenhausen, et al., 2001; Forgas, 2007). Therefore, even in instances when the feedback in the control condition was the same, the confidence with which claims that differences in the experimental condition resulted from the name as opposed to the same confounding variables potentially present in the control condition (e.g. mood state, motivation, circadian rhythms) was impacted.

Nevertheless, there was evidence that a small proportion of markers had noticed the prime of the student name and admitted that it had influenced their perception of the work. Moreover, the intra-individual changes identified in feedback provision within groups and across conditions lends weight to the argument that the name played a role. This was best illustrated when participants who marked a White British name in both the control and experimental conditions remained consistent in their feedback, but participants who marked a White British followed by a non-White British name did not. The premise being that if results showed that confounding variables such as mood state and circadian rhythms failed to impact on one group of markers it makes it less likely that these had an impact on the remaining groups, and increases certainty that the name provoked the changes. Furthermore, there were numerous examples that non-White names gained different feedback to White names and some evidence of gender differences.

So what does this say about expectancy effects and anonymous marking? Are the NUS claims of bias valid, and should HEIs be moving towards marking anonymously where possible? Since White British females (WBF), Asian females (AF) and students with non-White British names received less usable and educative in-text feedback across a range of domains this points to a need to mark anonymously. Nonetheless, this move would not be wholly popular, since there remains an ongoing tension between the perceived benefits of anonymous marking and the pedagogical losses that such a move would incur. For the most part the pedagogical arguments have related to feedback. Specifically that it interrupts the feedback loop (Whitelegg, 2002), prevents both useful dialogue and personalised feedback (Nicol, 2010; Bols, 2013; Laryea, 2013), and attenuates learning potential (Pitt & Winstone, 2018). However, findings from this thesis suggest that in terms of in-text feedback, WBF students and non-White British students did not receive feedback rich in learning potential, and therefore the cost-benefit ratio of anonymous marking for these student's shifts. Given the evidence that in-text comments are used and valued by students this is not insignificant (Ferguson, 2011), and suggests that certain types of students may not gain the benefits that marking anonymously claims to engender.

Ironically despite many arguments for non-anonymous marking revolving around feedback, research has largely explored grade bias to justify its necessity. Although recent research has found little evidence of such bias (Owen et al., 2010; Hinton & Higson, 2017; Pitt & Winstone, 2018), this research has limited utility when exploring biased feedback. For example, two students could be marked non-anonymously and gain the same grade, but be provided with different feedback which then impacts upon their ability to learn and improve. Admittedly the variable feedback practices evident in the control condition demonstrate that this can also occur when the student name remains the same, but at least the maintenance of anonymity would eliminate the bias having occurred as a result of the gender and ethnicity of the student.

Nevertheless, it is important to recognise that anonymous marking is not a panacea. Expectancy effects and subsequent biases can reveal themselves in multiple ways which are broader than knowing the student name. It is therefore possible that if the name were not visible markers would use other criteria to categorise the student. Additionally, the name is not always needed for a tutor to recognise work, particularly if they have been working closely with the student. Moreover, many assessment types cannot be anonymised. Therefore even if anonymous marking were able to eradicate expectancy effects for written assignments, the practice could not be universally applied without substantially reducing the assessment diet.

5.6 Tackling Expectancies and Bias

The awareness that expectancies and bias can arise from multiple stimuli, combined with the knowledge that they operate implicitly has led to strategies to reduce their effects. However, researchers in social psychology have argued that implementing strategies to reduce the *effects* of implicit bias (i.e., anonymous marking) is insufficient. Instead researchers should be focused on trying to reduce the *causes*. This sounds appealing, since if they can be reduced tutors may be able to mark non-anonymously and provide constructive feedback without accusations of bias. Additionally, a host of other expectancies, biases, and subjectivities which pollute the marking process could be eradicated.

The most common measure of implicit bias is the Implicit Association Test [IAT] (Greenwald, McGhee, & Schwartz, 1998). There are a variety of tests available which measure a range of biases. The test claims to detect an individual’s automatic association between schematic representations stored in memory. An IAT test exploring racial bias for example, would test the associations people make between “good” and “bad”, and “black” and “white”. The test is conducted under timed conditions designed to reduce the role conscious processing can play. The contention is that when two words appear which match people’s biases (e.g., White and professional) they respond more quickly than to those that do not (e.g., Black and Professor), because they are more closely associated in our memory. Given that for an individual to notice their bias, they first need to be aware of its existence, IAT tests have been a useful educational tool to make people aware of their prejudices. This awareness is critical if the issue of expectancy effects and bias in social judgement is to be addressed (Macrae & Bodenhausen, 2000; Bodenhausen, 2005).

Nonetheless, in recent years meta-analyses of IAT measures have claimed that these tests are no more predictive of attitudes than explicit measures of bias (Oswald, Mitchell, Blanton, Jaccard & Tetlock, 2013). More fundamentally, some research has found that attitudinal changes in implicit bias do not transfer to behavioural ones (Oswald et al., 2013; Forscher, Lai, Axt, Ebersole, Herman, Devine & Nosek, 2017). Nonetheless, despite these criticisms many eminent researchers value IAT measures and believe in their ability to raise consciousness and alter biased judgements and subsequent behaviours (Greenwald & Banaji, 2015).

Alongside IAT tests a variety of regulatory processes have been considered to counteract bias. The simplest strategy is to make adjustments to judgements in the opposite direction to the perceived bias (Wegener & Petty 1997; Macrae & Bodenhausen, 2000). However, since it is difficult to know how much bias is present, individuals often overcompensate (Bodenhausen, 2005). In the context of marking student work it is easy to see such practices being criticised in relation to arguments over accuracy and sympathetic marking. A more difficult strategy is thought/stereotype suppression, whereby individuals try to prevent expectancies and stereotypic thoughts entering their minds at all (Wegner, 1994; Bodenhausen & Macrae, 1998). While monitoring for such thoughts is easy, replacing them is more difficult, and crucially can only happen when individuals are not under high cognitive load (Wegner, 1994). Given that marking is a cognitively intensive activity it seems unlikely that this would be successful.

In addition to regulatory processes, a series of training programmes exist which offer training to tackle implicit biases. While employers in the fields of recruitment, law, and law enforcement have regularly integrated these programmes the field of education has lagged behind. Nonetheless, there is some research exploring the reduction of bias through the promotion of diversity education. Rudman, Ashmore and Gary (2001) gave White students an ethnicity-based IAT test alongside more explicit measures of racism. They were then part of a seminar about prejudices and bias taught by a Black Professor. Results demonstrated that anti-black biases reduced over the semester, although how long this reduction lasted for was not measured. At present no research has explored how implicit bias training might be used to reduce the expectancies and prejudices that operate in the marking process in HEIs.

Nonetheless, while the existence of programmes is welcomed, the efficacy of the cognitive strategies that underpin them and subsequently help to reduce bias has been questioned. Indeed, considerable evidence suggests that trying to suppress stereotypes can actually increase their activation and use (Wegner, 1994; Bodenhausen, 2005). These post-suppression rebound effects are more likely when the perceiver is cognitively busy or under time pressures (Macrae & Bodenhausen, 2000). Moreover, these effects have been found to be more prevalent when dealing with sensitive social groups (Sherman et al., 1997). Therefore raising awareness about bias through training programmes might ironically increase bias.

Encouragingly the sophistication of implicit bias training has increased and researchers have acknowledged that blanket programmes are unlikely to produce positive effects. Instead optimism surrounds longer term programmes which move beyond awareness-raising to consider skill development. For example, Lindsey, King, Membere and Cheung (2017) demonstrated that perspective-taking (which refers to being empathic), and goal-setting around issues related to diversity (e.g. setting goals to challenge sexist behaviour), led to long-term attitudinal changes (nine months post-test), but also importantly to behavioural changes (three months post-test). While more research is clearly needed, these findings do counter those of Oswald et al. (2013) and Forscher et al., (2017) who found no evidence for behavioural change. More recent research has also acknowledged the role personality characteristics play in the success of these programmes, claiming that people who score high in empathy and are motivated to address the issue will be more able to challenge their implicit thinking (Lindsey et al., 2017).

The general lack of confidence surrounding inhibitory processes and implicit bias training paints a negative picture, and the potential to reduce discrimination within marking and feedback looks bleak. Although contemporary research on training interventions has produced more positive results, assurances about their consistent ability to reduce implicit bias are remote. Until such assurances can be given, it may be incumbent upon HEIs to proactively remove one of the few controllable catalysts for expectancies; the student name. Although this thesis has shown that expectancies emanating from the name might be less influential as a single factor than others which can produce bias, it is at least controllable. This controllability is a luxury in the murky field of implicit process, and while anonymous marking would admittedly fall short in eliminating all causes of bias it would demonstrate a commitment to controlling for category-based expectancies derived from student names.

Advocates of marking as a social practice suggest students should be told about the subjective elements of marking. In this way the, "… collusion with the myth of objectivity" (Shay, 2005, p.676) can be stopped. They believe that students should be involved with creating marking criteria so that their awareness of the subjective element is enhanced. While such transparency is to be applauded, and the confidence in the student body to understand this dilemma may not be misplaced, a key issue has been overlooked. Students may be accepting of a social process which includes both subjective and objective elements, and therefore has to rely on the tacit knowledge of an experienced professional to make final decisions on feedback and grades. However, if this process is to be truly transparent, it also needs to be acknowledged that nested within it are markers who are prone to expectancy effects and bias. Furthermore, students would need to be informed these often operate at an implicit level which makes them difficult to control. Furnished with this knowledge it seems unlikely that students confidence in the marking process could be maintained, and since 'grades matter' (Sadler, 2009, p.808), who can blame students for calling for any measures which are perceived to reduce such bias.

Of course many assessment types would prevent a wholesale move towards anonymous marking. These more authentic assessments, which arguably better prepare students for employment, would remain prone to such effects even if anonymous marking were adopted. Consequently, despite questions surrounding the effectiveness of such interventions, universities might be prudent to incorporate anti-bias or implicit training programmes into their practice. In this way, even if perceptions of bias are low (Pitt & Norton, 2018) they can make assurances that they are proactively addressing this important issue.

6 LIMITATIONS

The first limitation concerned the names chosen to represent ethnic groups. An International student recruitment officer was asked whether the names selected for the essays were representative of the ethnicities and recognisably male or female. Nonetheless, some participants worked in HEIs with a less ethnically diverse student population than others. Resultantly, although participants might have been able to distinguish ethnicity for Asian and Chinese names, recognising gender might have been more problematic. This had implications for the gender analyses since confidence that gender was being compared was low for all but WBM versus WBF. Furthermore, the category Asian required greater clarity, since it is an umbrella term for numerous ethnic groups. Both Asian names in this thesis derived from Southern Asia and are of Indian descent.

A further limitation surrounded marker variability. Several participants failed to provide in-text feedback, and a large percentage failed to provide summary feedback. Subsequently some comparisons were either impossible, because there was no data, or the data was incomplete. Moreover, because low numbers of participants provided summary feedback it was difficult to conduct a hierarchical content analysis in the truest sense. The exploration of meaning through a more latent analysis, and the ability to indwell were curtailed by the quantity and depth of the data received. These shortcomings made it difficult to explore expectancy effects with breadth and clarity. Although it is conceivable that more prescriptive instructions might have alleviated these problems, asking participants to ‘mark in line with current teaching practice’, was considered to be more authentic and fit closely with the pragmatist worldview which underpinned the experimental design.

Substantial differences were evident in feedback practices in the control condition. Therefore it might have been prudent to establish marker variability over more than one essay in order to more confidently judge participants’ feedback practices. Additionally, more stringent measures could have been taken to establish similarities between the control and experimental essays. While they had the same title and occupied the same grade boundary, they could have been better aligned in terms of stylistic elements, structure, use of literature and the ability to construct argument. It is possible that the control essay was well written but lacked other aspects of academic quality, whereas the experimental essay was poorly written but well-researched and thought provoking. However, if a marker is a pedant for clarity of writing, their judgement of the experimental essay may be obscured by them activating categories of ‘lazy’ or ‘incompetent’, and this may shape their feedback rather than the student name. Therefore, ensuring greater symmetry between essays would provide more confidence that differences could be attributed to the student name.

In addition to addressing the above limitations, and extending research on student gender, ethnicity, and the interactive effects of these on feedback, further expectancy-based research might explore how expectancies derived from sources other than the student name can influence the feedback provided (i.e., motivation, mood state, order effects etc.) Knowing what these factors are, and understanding their potential influence would allow us to gauge the relative influence of the name more assuredly. Additionally, continuing the research conducted on marker gender and feedback would be interesting. Perhaps different genders are prone to different expectancies, or perhaps one gender is more prone to sympathetic marking. Future research might also extend work that has correlated feedback with grades to see whether groups that receive different types of feedback also receive different grades. It is possible that these might differ in the same way that in-text and summary feedback can.

7 CONCLUSION

This thesis sought to examine the need for anonymous marking within HEIs. Using a novel approach, it explored the impact of expectancy effects on feedback for non-anonymised undergraduate student essays. Using Olson et al.'s (1996) model of expectancy effects, and associated concepts from both social and cognitive psychology, it sought to examine whether differences existed in relation to gender, ethnicity, and gender and ethnicity of the student (as inferred from the name).

Findings demonstrated that in-text feedback showed more evidence of expectancy effects than summary feedback. However, expectancy effects were not consistently present across all analyses. Reasons for this have been addressed in the limitations. Nonetheless, in-text feedback pertaining to gender found White British female (WBF) names received less positive feedback, more negative feedback, and less developmental feedback when compared to White British males (WBM). This was reversed in the summary feedback where WBFs received more praise and descriptive feedback linked to higher-level academic skills. Additionally, Asian females (AF) gained fewer comments on content and how to develop than their male counterparts. They did receive more reflective questions than Asian males (AM), but there were no differences in the summary provision.

Non-White British names also received different in-text feedback to White British names. These differences were generated by the Chinese male (CM) and AF, and comprised of receiving more feedback on mechanical aspects of writing which were only corrective or emphasis-based. Furthermore, non-White British names gained little comment-based developmental feedback, instead being provided mainly with ticks. Once again there were some contradictions between in-text and summary feedback which showed isolated examples of non-White British names receiving more positive feedback when compared to WBMs.

Several differences were evident for gender and ethnicity. CMs gained less developmental feedback related to *alternatives* than any other group, while the WBF scored lowest for *reflective questions*. White British students of both genders gained more *informational* comments. Punctuation errors were cited more for Chinese students of both genders than either White British or Asian students. AF gained less useful content-related feedback than both WBM and females of other ethnicities. The relative effects of gender and ethnicity changed according to who the students was being judged alongside and these have been further teased out in the discussion.

Some unanticipated findings emerged which showed that feedback was both highly variable and littered with poor practice. Without exception, participants' comments failed to deliver the types of feedback that the literature recommends. Moreover, despite marking criteria being provided, multiple in-text feedback differences existed when the student name remained the same. This indicated that expectancy effects outside those primed by the student name were present, and might be even more influential than the name itself. Importantly, it also signified that the implementation of anonymous marking would only remove one of the factors responsible for expectancy effects within feedback. Nonetheless, since there was some evidence of expectancy effects for some groups the results would suggest a need for anonymous marking.

Importantly, the findings also nullified the feedback-based arguments often raised to justify marking non-anonymously. While generally the feedback was of poor quality, often the pockets of good practice that were included in the in-text feedback did not extend to WBF, AF, and non-White British students. Therefore when markers knew names, these groups received poorer feedback than some other groups. Ironically therefore a move towards anonymous marking might enhance rather than diminish feedback quality for these students.

This thesis therefore suggested that if HEIs are serious about trying to reduce expectancy effects and bias from marking and feedback processes they might be advised to enrol their staff in implicit bias training. Although the success of these training programmes has been equivocal, contemporary research suggests that more sophisticated designs can reap the benefits of both attitudinal and behavioural changes. Furthermore, engagement with such strategies would demonstrate to the student body that while marking remains a social practice, efforts to reduce expectancy effects and implicit bias have been addressed. At present no training programmes exist which specifically address bias in marking and feedback processes and this is therefore an attractive and important avenue for future research.

8 REFERENCES

Agius, N. M., & Wilkinson, A. (2014). Students' and teachers' views of written feedback at undergraduate level: A literature review. *Nurse Education Today, 34,* (4), 552–59 doi:10.1016/j.nedt.2013.07.005.

Ali, N., Rose, S., & Ahmed, L. (2015). Psychology students' perceptions of and engagement with feedback as a function of year of study. *Assessment & Evaluation in Higher Education, 40*, (4), 574-586, DOI: 10.1080/02602938.2014.936355

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Asch, S.E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258-290.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111,* (2), 256-274.

Archibald, M. M. (2016). Investigator Triangulation A Collaborative Strategy With Potential for Mixed Methods Research. *Journal of Mixed Methods Research*, *10,* (3), 228-250.

Armitage, A., & Campus, R. (2007, September). Mutual research designs: Redefining mixed methods research design. In *Paper presented at the British Educational Research Association Annual Conference* (5), p. 8.

Attride-Stirling, J. (2001). Thematic networks: an analytic tool for qualitative research. *Qualitative Research, 1,* 385 - 405.

Auwarter, A. E., & Aruguete, M. S. (2008). Effects of student gender and socioeconomic status on teacher perceptions. *The Journal of Educational Research*, *101,* (4), 242-246.

Babad, E. Y. (1980). Expectancy bias in scoring as a function of ability and ethnic labels. *Psychological Reports*, *46,* 2, 625-626.

Babad, E. Y., Inbar, J., & Rosenthal, R. (1982). Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology*, *74*, 459 – 474.

Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, *40*, (2), 191-202.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.). *Handbook of social cognition* (2nd ed.), (pp. 1 - 40).Hillsdale, NJ: Erlbaum.

Bargh, J. A. (1999). The cognitive monster. In S. Chaiken & Y. Trope (Eds.). *Dual Process Theories in Social Psychology,* (pp. 361 - 382). New York: Guilford Press.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, *43*, 437-449.

Baron, R. A. (1988). Negative effects of destructive criticism: impact on conflict, self-efficacy, and task performance. *Journal of Applied Psychology, 73*, (2), 199-207.

Baron, R., Tom, D., & Cooper, H. (1985). Social class, race and teacher expectations. In J. Dusek (Ed.), Teacher expectancies. Hillsdale, NJ: Erlbaum.

Bartram, B. (2016). Emotion as a resource in higher education. *British Journal of Educational Studies, 63,* (1), 67-84.

Baty, P. (2007). Trust eroded by blind marking. Retrieved August 18th, 2014, from http://www.timeshighereducation.co.uk/story.asp?storyCode=209075&sectioncode=26

Batten, J., Batey, J., Shafe, L., Gubby, L., & Birch, P. (2011). The influence of reputation information on the assessment of undergraduate student work. *Assessment and Evaluation in Higher Education*, *38*, (4), 417-435.

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The Effect of Professor Ethnicity and Gender on Student Evaluations: Judged Before Met. *Journal of Diversity in Higher Education*. Advance online publication. doi: 10.1037/a0020763

Bazeley, P. (2004). Issues in mixing qualitative and quantitative approaches to research. *Applying qualitative methods to marketing management research*, 141-156.

Beadle, P. (2012). The importance of marking. *Times Educational Supplement*. Available at http://newteachers.tes.co.uk/news/importance-marking/45966 Accessed 22/01/16.

Belsey, C. (1988). Marking by numbers. *AUT Women, 15* (pp. 1–2). London: Association of University Teachers.

Bertrand, M. & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94,* (4), 991-1013

Biddle, S. J., Markland, D., Gilbourne, D., Chatzisarantis, N. L., & Sparkes, A. C. (2001). Research methods in sport and exercise psychology: Quantitative and qualitative issues. *Journal of Sports Sciences*, *19,* (10), 777-809.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66,* (1), 5-20.

Biernat, M., Crandall, C. S., Young, L. V., Kobrynowicz, D., & Halpin, S. M. (1998). All that you can be: Stereotyping of self and others in a military context. *Journal of Personality and Social Psychology*, *75,* (2), 301.

Biesanz, J. C., Neuberg, S. L., Smith, D. M., Asher, T., & Judice, T. N. (2001). When Accuracy-Motivated Perceivers Fail: Limited Attentional Resources and the Re-emerging Self-Fulfilling Prophecy. *Personality and Social Psychology Bulletin*, *27*, (5), 621-629.

Blair, A., & McGinty, S. (2012). Feedback-dialogues: Exploring the student perspective. *Assessment & Evaluation in Higher Education, 38,* (5), 554–566.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*, (3), 242-261.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163.

Bloxham, S. (2009) Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education, 34*, (2), 209-220.

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36,* (6), 655-670.

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: testing process models of stereotype use. *Journal of Personality and Social Psychology*, *55,* (5), 726 -737

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science*, *1*, 319-322.

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotype-based discrimination: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.). *Affect, Cognition, and Stereotyping: Interactive processes in group perception* (pp. 13-35). San Diego, CA: Academic Press.

Bodenhausen GV, Macrae CN. 1998. Stereotype activation and inhibition. In R. S. Wyer (Ed.).*Stereotype Activation and Inhibition: Advances in Social Cognition,* 11, (pp.1–52). Hillsdale, NJ: Erlbaum.

Bodenhausen, G. V., Mussweiler, T., Gabriel, S., & Moreno, K. N. (2001). Affective influences on stereotyping and intergroup relations. In J. Forgas (Ed.). *Handbook of Affect and Social Cognition* (pp. 319-343). Mahwah, NJ: Lawrence Erlbaum.

Bodenhausen, G. V. (2005). The role of stereotypes in decision-making processes. *Medical Decision Making*, *25*, 112-118.

Bols, A. (2013). Student Views on Assessment. In L. Clouder, C. Broughan, S. Jewell, and G. Steventon (Eds.). *Improving Student Engagement and Development through Assessment: Theory and Practice in Higher Education* (pp. 4‑18). London: Routledge.

Boud, D. (1995). Assessment and Learning: Contradictory or Complementary? In P. Knight (Ed.). *Assessment for Learning in Higher Education,* pp. 35–48. London: Kogan Page.

Boud, D., & Malloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38,* (6), 698-712. http://dx.doi.org/10.1080/02602938.2012.691462

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology.* Cambridge: Polity Press.

Bradley, C. (1984). Sex bias in the evaluation of students. *British Journal of Social Psychology*, *23,* (2), 147-153.

Breda, T., & Ly. S. T. (2014). Professors in core science fields are biased in favour of women: Evidence from France." Working paper. http://www.parisschoolofeconomics.eu/docs/ly-son-thierry/gendergapulm.pdf.

Brennan, D. J. (2008). University student anonymity in the summative assessment of written work. *Higher Education Research and Development, 27*, (1), 43-54.

Brennan, J., & Shah, T. (2000). Quality assessment and institutional change: Experience from 14 countries. *Higher Education, 40*, 331 -349.

Broecke, S., & Nicholls, T. (2006) Ethnicity and degree attainment. *Department for Education and Skills.* Research Report RW92. Available at http://www.dfes.gov.uk/research

Brophy, J. (1983). Research on the self-fulfilling prophecy and teachers expectations. *Journal of Educational Psychology*, *75*, 631-661.

Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences.* New York: Holt.

Brown, E., Gibbs, G., & Glover, C. (2003). Evaluation tools for investigating the impact of assessment regimes on student learning. *Bioscience Education, 2*, (1), 1-7.

Brown, E., & Glover, C. (2006). Evaluating written feedback. In C. Bryan & K. Clegg (Eds.). *Innovative Assessment in Higher Education* (pp.81–91). Abingdon: Routledge.

Burgess, S., & Greaves, E. (2009). Test scores, subjective assessment and stereotyping of ethnic minorities. CMPO working paper, No.09/221. *The Centre for Market and Public Organisation.*

Buscombe, R., Greenlees, I., Holder, T., Thelwell, R., & Rimmer, M. (2006). Expectancy effects in tennis: The impact of opponents' pre-match non-verbal behaviour on male tennis players. *Journal of Sports Sciences*, 24, (12), 1265-1272.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119,* 197-253.

Campbell, T. (2013). Stereotyped at seven? Biases in teacher judgements of pupils' ability and attainment. *University of London: Institute of Education:* London, UK.

Canon (2015). The Lab: Decoy – A portrait session with a twist Available at: https://www.youtube.com/watch?v=F-TyPfYMDK8 [Accessed 10th July 2018]

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology*, *71,* (4), 672-678.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education*, *31,* (2), 219-233.

Carlsson, M., & Roothe, D. O. (2007). Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics*, 14, (4), 716-729.

Chaiken, A. L., Sigler, E., & Derlega, V. J. (1974). Nonverbal mediators of teacher expectancy effects. *Journal of Personality and Social Psychology*, *30*, (1), 144-149.

Chapman, L. J., & Chapman, J.P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, (3), 193-204.

Chen, M., & Bargh, J.A. (1997). Nonconscious behavioural confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology, 33,* 541–560.

Chory-Assad, R.M. (2002). Classroom justice: perceptions of fairness as a predictor of student motivation, learning and aggression. *Communication Quarterly, 50,* 58–77.

Christie, M., Grainger, P., Dahlgren, R., Call, K., Heck, D., & Simon, S. E. (2015). Improving the quality of assessment grading tools in master of education courses: A comparative case study in the scholarship of teaching and learning. *Journal of the Scholarship of Learning and Teaching, 15*, 22–35. doi:10.14434/josotl.v15i5.13783.

Clark, M. S., & Isen, A. M. (1982). Towards understanding the relationship between feeling states and social behavior. In A. H. Hastorf & A. M. Isen (Eds.). *Cognitive Social Psychology* (pp.73–108). New York: Elsevier.

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin*, *25*, (10), 1302-1318.

Cooper, H. M. (1979). Pygmalion grows up: A model for teacher expectation communication and performance influence. *Review of Educational Research*, *49*, (3), 389-410.

Cooper, H. M., Baron, R. M., & Lowe, C. A. (1975). The importance of race and social class information in the formation of expectancies about academic performance. *Journal of Educational Psychology, 30*, 846-855.

Cooper, H., *&* Baron, R. (1977). Academic expectations and attributed responsibility as predictors of professional teachers' reinforcement behavior. *Journal of Educational Psychology, 69,* 409-418.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using race to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83,* 1314–1329.

Cote, J., Salmela, J. H., Baria, A. & Russell, S. (1993). Organising and interpreting unstructured qualitative data. *The Sport Psychologist*, 7, 127-137.

Crammond, J. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication*, 15, (2), 230–268.

Crano, W. D., & Mellon, P. M. (1978). Causal influence of teachers' expectations on children's academic performance: A cross-lagged panel analysis. *Journal of Educational Psychology*, *70*, 1, 39-49, doi: 10.1037/0022-0663.70.1.39

Creswell, J. W., Shope, R., Plano Clark, V. L., & Green, D. O. (2006). How interpretive qualitative research extends mixed methods research. *Research in the Schools*, *13,* (1), 1-11.

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd edition). Los Angeles: Sage.

Crimmins, G., Nash, G., Oprescu, F., Liebergreen, M., Turley, J., Bond, R., & Dayton, J. (2016). A written, reflective and dialogic strategy for assessment feedback that can enhance student/teacher relationships. *Assessment & Evaluation in Higher Education, 41*, 141–153. doi:10.1080/02602938.2014.986644

Crook, C.K., Gross, H., & Dymott, R. (2006). Assessment relationships in higher education: The tension of process and practice. *British Educational Research Journal,* 32, (1), 95–114.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44*, 20–33.

Darley, J. M., & Fazio, R. H. (1980). Expectancy-confirmation processes arising in the social interaction sequence. *American Psychologist*, *35*, 867-881.

de Boer, H., Bosker, R. J., van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*, (1), 168-179.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86,* (1), 195-210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review*, *95,* (2), 158-165.

Dennis, I. (2007). Halo effects in grading student projects. *Journal of Applied Psychology*, *92*, (4), 1169 – 1176.

Dennis, I., & Newstead, S. E. (1994). The strange case of the disappearing sex bias. *Assessment & Evaluation in Higher Education*, *19,* (1), 49-56.

Dennis, I., Newstead, S. E., & Wright, D. E. (1996). A new approach to exploring biases in educational assessment. *British Journal of Psychology*, *87,* (4), 515-534.

Denzin, N. (1970). The research act: A theoretical introduction to sociological methods. Chicago, IL: Aldine.

Denzin, N. K. (1978). *Sociological methods: A sourcebook.* New York: McGraw-Hill Companies.

Denzin, N. K., & Lincoln, Y. S. (2005). Introduction: The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 1-32). Thousand Oaks, CA: Sage

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.

Diederich, P. B. (1974). Measuring growth in English. Urbana, IL: National Council of Teachers of English.

Dijker, A. M. (1987). Emotional reactions to ethnic minorities. *European Journal of Social Psychology, 17,* (3), 305-325.

Dochy, F.J., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279–298.

Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International, 13,* 313-321.

Duncan, N. (2007). 'Feedforward': Improving students' use of tutors' comments. *Assessment & Evaluation in Higher Education, 32,* (3), 271–83*.*

Dunning, D., & Sherman, D.A. (1997). Stereotypes and tacit inference. *Journal of Personality and Social Psychology, 73*, 459–471.

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*, (3), 327–346.

Eagly, A. H. & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: a meta-analysis of social influence studies, *Psychological Bulletin, 90,* (1), 1-20.

Ecclestone, K. (2001). "I know a 2:1 when I see it": Understanding degree standards in programmes franchised to colleges. *Journal of Further & Higher Education*, 25 (4) 301- 313.

Edwards J. A., & Weary, G. (1993). Depression and the impression-formation continuum: piecemeal processing despite the availability of category information. *Journal of Personality and Social Psychology, 64*, 636 – 645.

Elander, J., & Hardman, D. (2002). An application of judgment analysis to examination marking in psychology. *British Journal of Psychology, 93*, 303–328.

Elashoff, J. D., & Snow, R. E. (1971). *Pygmalion Reconsidered*. Worthington, Ohio: Charles A. Jones

Enzi, B. (2015). Gender differentials in test scores and teacher assessments: Evidence from Germany. Working paper. http:// www.edge-page.net/jamb2014/papers/Enzi%20-%20Draft.pdf.

Erwin, P. G., & Calev, A. (1984). The influence of Christian name stereotypes on the marking of children's essays. *British Journal of Educational Psychology, 54*, 223–227.

Esqueda, C. W., Espinoza, R. K., & Culhane, S. E. (2008). The effects of ethnicity, socio-economic status, and crime status on juror decision making: A cross-cultural examination of European American and Mexican American mock jurors. *Hispanic Journal of Behavioral Sciences*, *30*, 181-199.

Eysenck, M. W. (2009). *Fundamentals of Psychology*. New York: Psychology Press.

Fajardo, D. M. (1985). Author race, essay quality and reverse discrimination. *Journal of Applied Social Psychology, 15*, 255–268.

Fazio, R. H., & Zanna, M. P. (1981). Direct experience and attitude-behavior consistency. *Advances in Experimental Social Psychology*, *14*, 161-202.

Ferguson, P. (2011) Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education, 36,* (1), 51-62.

Fiske, S. T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. S. Uleman & J. A. Bargh (Eds.). *Unintended Thought*, 253-283. New York: Guilford Press.

Fiske, S. T., Lin, M., & Neuberg, S. (1999). The continuum model. In S. Chaiken & Y. Trope (Eds.). *Dual-process Theories in Social Psychology* (pp. 254-321). Guilford Press: London.

Fiske, S.T. (2002). What we know now about bias and intergroup conflict, the problem of the century. *Current Directions in Psychological Science*, *11,* (4), 123-128.

Fiske, S.T., & Taylor, S.E. (1984). Social Cognition. Reading, MA: Addison-Wesley.

Fiske, S.T. & Neuberg, S.L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M.P. Zanna (Ed.). *Advances in Experimental Social Psychology*, 23, 1-74. New York: Academic Press.

Fiske, S.T., & Taylor, S.E. (1991). *Social Cognition*. New York: McGraw-Hill

Fiske S.T., & Depret, E. (1996). Control, interdependence and power: understanding social cognition in its social context. *European Review of Social Psychology, 7*, 31-62.

Ford, T. E., & Thompson, E. P. (2000) Preconscious and postconscious processes underlying construct accessibility effects: An extended search model. *Personality and Social Psychology Review, 4*, 317-336.

Forgas J.P. (1995). Emotion in social judgments: review and a new affect infusion model (AIM). *Psychological Bulletin, 117*, 39 - 66.

Forgas, J. P. (2002a). Feeling and doing: Affective influences on interpersonal behaviour. *Psychological Inquiry, 13*, 1, 1-28.

Forgas, J. P. (2002b). Towards understanding the role of affect in social thinking and behaviour. *Psychological Inquiry, 13*, 1, 90-102.

Forgas, J. P. (2007). When sad is better than happy: Negative affect can improve the quality and effectiveness of persuasive messages and social influence strategies. *Journal of Experimental Social Psychology, 43*, 513–528.

Forgas, J. P., & Laham, S. (2009). Halo effects. In R. Baumeister, & K. D. Vohs (Eds.). *Encyclopedia of Social Psychology* (pp. 499–502). Thousand Oaks: Sage Publications.

Forgas, J. P. (2011). She just doesn't look like a philosopher….? Affective influences on the halo effect in impression formation. *European Journal of Social Psychology, 41*, 812-817.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2017). A meta-analyses of change in implicit bias. *Psychological Bulletin* (in review)

Fowell, S. L., Maudsley, G., Maguire, P., Leinster, S. J., & Bligh, J. (2000). Student assessment in undergraduate medical education in the United Kingdom, 1998. *Medical Education*, *34,* (s 1), 1-49.

Francis, B., Robson. J., & Read, B. (2001). An analysis of Undergraduate writing styles in the context of gender and achievement. *Studies in Higher Education, 26,* 313-326.

Francis, B, Read, B., & Melling, L. (2003). University lecturers' perceptions of gender and Undergraduate writing. *British Journal of Sociology in Education, 24,* (3), 357-373.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75-90.

Garratt, D. & Hodkinson, P. (1999). Can there be criteria for selecting research criteria? A hermeneutical analysis of an inescapable dilemma. *Qualitative Inquiry*, *4*, 515 -539.

Ghaye, T., Danai, K., Cuthbert, L., & Dennis, D. (1996). *Introduction to learning through critical reflective practice*. Newcastle-Upon-Tyne, UK: Pentaxion.

Gibbs, G. (2013). Feedback turnaround time: CADQ guide. Nottingham Trent University: Centre for Academic Development and Quality, p.1-9 www.ntu.ac.uk/cadq

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education* 1, (1), 1–31.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*, 509–517.

Gittoes, M., & Thompson, J. (2005). Higher education admissions: Assessment of bias (HEFCE 2005/47). *Bristol: HEFCE*.

Glascock, J., & Ruggiero, T. E. (2006). The relationship of ethnicity and sex to professor credibility at a culturally diverse university. *Communication Education, 55,* 197–207.

Glover, C., & Brown, E. (2006). Written feedback for students: too much, too detailed or too incomprehensible to be effective? *Bioscience Education*, *7,* (1), 1-16.

Goddard-Spear, M. (1984). The Biasing Influence of Pupil Sex in a Science Marking Exercise. *Research in Science & Technological Education 2*, (1), 55–60.

Greenlees, I. (2007). Person perception and sport performance. In S. Jowett & D. Lavallee (Eds.). *Social Psychology in Sport*, (pp.195-208). Champaign ILL.: Human Kinetics.

Greenlees, I., Buscombe, R., Thelwell, R., Holder, T., & Rimmer, M. (2005). Impact of Opponents' Clothing and Body Language on Impression Formation and Outcome Expectations. *Journal of Sport and Exercise Psychology*, *27*, 39-52.

Greifeneder, R., Zelt, S., Seele, T., Bottenberg, K., & Alt, A. (2012). Towards a better understanding of the legibility bias in performance assessments: The case of gender-based inferences. *British Journal of Educational Psychology*, *82,* (3), 361-374.

Hamilton, D. L. (1981). Illusory correlation as a basis for stereotyping. In D.L. Hamilton (Ed.). *Cognitive Processes in Stereotyping and Intergroup Behavior*, (pp.115-144). Hillsdale, NJ: Erlbaum.

Hamilton, D. L., Sherman, S. J., & Ruvolo, C. M. (1990). Stereotype-based expectancies: Effects on information processing and social behavior. *Journal of Social Issues*, *46,* (2), 35-60.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*, (3), 337-373.

Harari, H., & McDavid, H. W. (1973). Name stereotypes and teacher's expectations. *Journal of Educational Psychology, 65*, 222–225.

Harber, K. D. (1998). Feedback to minorities: Evidence of a positive bias. *Journal of Personality and Social Psychology, 74,* (3), 622-628.

Harber, K. D., Stafford, R., Kennedy, K. A. (2010). The positive feedback bias as a response to self-image threat. *British Journal of Social Psychology, 49,* 207-218.

Hardy, L., Jones, G. & Gould, D. (1996). Understanding psychological preparation for sport. Chichester: John Wiley.

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*, (3), 363-386.

Harrison, L. Jr. (2001). Understanding the influence of stereotypes: Implications for the African American in sport and physical activity. *Quest*, *53,* (1), 97-114.

Harrison, L., Harrison, K., & Moore, L. (2002). African American racial identity and sport. *Sport, Education and Society 7*, 121–33.

Hattie, J., Biggs, J., & Purdie, N. (1996) Effects of learning skills intervention on student learning: a meta-analysis. *International Journal of Educational Research, 11,* 187–212.

Hattie, J. & Jaeger, R. (1998) Assessment and classroom learning: a deductive approach. *Assessment in Education, 5*, (1), 111–122.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77,* (1), 81-112.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

HEFCE (2005). *Higher education admissions: assessment of bias* (HEFCE publication 2005/47).

HEFCE (2010). Student ethnicity: profile and progression of entrants to full-time, first degree study (HEFCE publication 2010/13).

Hepplestone, S., & Chikawa, G. (2014). Understanding how students process and use feedback to support their learning. *Practitioner Research in Higher Education 8,* 1,
41–53.

Heywood, J. (2000). *Assessment in Higher Education.* London: Jessica Kingsley.

Higher Education Statistics Agency (HESA) (2004) *Qualifications data* (Cheltenham, HESA).

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141-154.

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual review of psychology*, *38*, (1), 369-425.

Higgins, R., Hartley, P., & Skelton, A. (2002). The conscientious consumer: reconsidering the role of assessment feedback in student learning. *Studies in Higher Education, (27),* 1, 53-64

Hinnerich, B.T., Hoeglin, E., & Johannesson, M. (2011). *Ethnic discrimination in high school grading: Evidence from a field experiment.* SSE/EFI Working Paper 733, Stockholm: Stockholm School of Economics.

Hinton, D. P., & Higson, H. (2017). A large-scale examination of the effectiveness of anonymous marking in reducing group performance differences in higher education assessment. *PLoS ONE 12* (8), e0182711. https://doi.org/10.1371/journal.pone.0182711

Hodson, G., & Busseri, M.A. (2013).Bright minds and dark attitudes: Lower cognitive ability predicts greater prejudice through right-wing ideology and low intergroup contact. *Psychological Science, 24,* (7), 1-9*.*

Hounsell, D. (1987). Essay writing and the quality of feedback. In J.T.E. Richardson, M.W. Eysenck, and D. Warren-Piper. Student learning: Research in education and cognitive psychology, (Eds.) (pp.109–19). Milton Keynes: Open University Press and Society for Research into Higher Education.

Hounsell, D. (2003). Student feedback, learning and development. In M. Slowey and D. Watson (Eds.). *Higher education and the lifecourse* (pp.67–78). Maidenhead: Society for Research into Higher Education and Open University Press.

Hsieh, H., & Shannon, S. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*: 1277-1288.

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment & Evaluation in Higher Education, 36,* (1), 109-124.

Husserl, E. (1931). *Ideas: General Introduction to Pure Phenomenology*, trans. Gilson W. R. B.. New York: Humanities Press.

Hyatt, D. F. (2005). 'Yes, a very good point!': a critical genre analysis of a corpus of feedback commentaries on master of Education assignments. *Teaching in Higher Education, 10,* (3), 339-353.

Hyland, P. (2000). Learning from feedback on assessment. In: P. Hyland & A. Booth (Eds.).*The Practice of University History Teaching.* Manchester: Manchester University Press.

Hyland, F. (2001). Providing effective support: investigating feedback to distance language learners. *Open Learning 16,* (3), 233–247.

Isen, A. M. (1984). Towards understanding the role of affect in cognition. In R. S. Wyer & T. K. Srull (Eds.). *Handbook of Social Cognition* (Vol. 3, pp. 179–236). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Isen, A. M. (1987). Positive affect, cognitive processes and social behaviour. In L. Berkowitz (Ed.). *Advances in Experimental Social Psychology (20),* pp. 203–253). New York: Academic.

Isen, A. M., Shalker, T. E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory and behavior: A cognitive loop? *Journal of Personality and Social Psychology, 36,* 1–12.

Isen, A. M., Johnson, M. M., Mertz, E., & Robinson, G. (1985). The influence of positive affect on the unusualness of work associations. *Journal of Personality and Social Psychology, 48,* 1413–1426.

Isen, A. M., Niedenthal, P., & Cantor, N. (1992). An influence of positive affect on social categorization. *Motivation & Emotion, 16,* 65–78.

Jampol, L. (2014). *How gender-based feedback contributes to the 'glass ceiling'.* Unpublished thesis. https://ecommons.cornell.edu/handle/1813/47/browse?value=Jampol%2C+Lily&type=author

Joffe, H., & Yardley, L. (2004). Content and thematic analysis. In D. F. Marks, & L. Yardley. *Research methods for clinical and health psychology,* (Eds.) (pp. 56-68). London: Sage.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, *33,* 7, 14-26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, *1*, 2, 112-133.

Jonason, P. K. (2015). How "dark" personality traits and perceptions come together to predict racism in Australia. *Personality and Individual Differences*, *72*, 47-51.

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education 14*, 63–76.

Jones, E. E. (1986). Interpreting interpersonal behaviour: The effects of expectancies. *Science*, *234*, 41-46.

Jones, E. E. (1990). Interpersonal perception. New York: Freeman

Jones, E. E., & McGillis, I. (1976). Correspondent inferences and the attribution cube: A comparative reappraisal. In J. H. Harvey, W. J. Ickes, & R. K. Kidd (Eds.). *New directions in attribution research*. New York: Wiley.

Jones, E., Farina, A., Hastorf, A., Markus, H., Miller, D., & Scott, R. (1984). *Social stigma: The psychology of marked relationships.* New York: Freeman.

Jones, M. V., Paull, G. C., & Erskine, J. (2002). The impact of a team's aggressive reputation on the decisions of Association Football referees. *Journal of Sports Sciences, 20*, 991-1000.

Judice, T. N., & Neuberg, S. L. (1998). When interviewers desire to confirm negative expectations: Self-fulfilling prophecies and inflated applicant self-perceptions. *Basic and Applied Social Psychology*, *20,* 175-190.

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review*, *93,* (4), 429-445.

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology*, *57*, (3), 469-480.

Jussim, L. (1990). Social reality and social problems: The role of expectancies. *Journal of Social Issues*, *46*, (2), 9-34.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 98*, 54-73.

Jussim, L. (1993). Accuracy in interpersonal expectations: A reflection-construction analysis of current and classic research. *Journal of Personality, 61,* 637–668.

Jussim, L. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in experimental social psychology*, *37*, 1-93.

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. Oxford: Oxford University Press.

Jussim, L., Coleman, L. M., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology, 52,* 536–546.

Jussim, L., & Eccles, J. S. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of personality and social psychology*, *63*, (6), 947-961.

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in experimental social psychology*, *28*, 281-388.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*, (2), 131-155.

Jussim, L., Robustelli, S. & Cain, T. (2009). Teacher expectations and self-fulfilling prophecies. In A. Wigfield and K. Wentzel (Eds.), *Handbook of Motivation at School* (pp. 349-380). New Jersey: Erlbaum.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, (3), 582-591.

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin*, *24*, (4), 407-416.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence.* New York: Wiley.

Kember, D. (2003). To control or not to control: The question of whether experimental designs are appropriate for evaluating teaching innovations in higher education. *Assessment & Evaluation in Higher Education*, *28,* (1), 89-101.

Kierein, N. M., & Gold, M. A. (2000). Pygmalion in work organizations: a meta-analysis. *Journal of Organizational Behavior*, *21*, 913–928.

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach, *Education Economics, 21*, 5, 447-463, DOI: 10.1080/09645292.2011.585019

Klein, O., & Snyder, M. (2003). Stereotypes and behavioral confirmation: From interpersonal to intergroup perspectives. *Advances in experimental social psychology*, *35,* 153-234.

Knight, P. T. (2002). Summative assessment in higher education: practices in disarray. *Studies in Higher Education, 27*, (3), 275-286.

Knight, P. T., & Yorke, M. (2003). *Assessment, learning and employability.* Maidenhead: Open University Press.

Kobrynowicz, D., & Biernat, M. (1997). Decoding subjective evaluations: How stereotypes provide shifting standards. *Journal of Experimental Social Psychology*, *33,* (6), 579-601.

Krause, K., Hartley, R., James, R., & McInnis, C. (2009).The first year experience in Australian universities: Findings from a decade of national studies. http://www.cshe.unimelb.edu.au/research/experience/docs/FYE_Report_1994_to_2009.pdf

Krawczyk, M. (2018). Do gender and physical attractiveness affect college grades? *Assessment & Evaluation in Higher Education, 43,* (1), 151-161 https://doi.org/10.1080/02602938.2017.1307320

Krueger, R. F. (2002). Personality from a realist's perspective: Personality traits, criminal behaviors, and the externalizing spectrum. *Journal of Research in Personality*, *36*, 6, 564-572.

Kuklinski, M. R., & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development*, *72* (5), 1554-1578.

Kumar, V., & Stracke, E. (2007). An analysis of written feedback on a PhD thesis. *Teaching in Higher Education*, *12,* (4), 461-470.

Kvale, S. (2008). Qualitative inquiry between scientistic evidentialism, ethical subjectivism and the free market. International Review of Qualitative Research, 1,(1), 5-18.

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, *29,* (3), 299.

Lanzetta, J. T., & Hannah, T. E. (1969). Reinforcing behavior of "naive" trainers. *Journal of Personality and Social Psychology*, *11*, 3, 245-252.

Laryea, S. (2013). Feedback Provision and Use in Teaching and Learning: A Case Study. *Education and Training 7*, 665‐680.

Lavy, V. (2004). *Do Gender Stereotypes Reduce Girls' Human Capital Outcomes? Evidence from a Natural Experiment*: SSRN.

Lea, M., & Street, B. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23,* (2), 157-172.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72*, 275–287.

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: the role of implicit theories. *Journal of Personality and Social Psychology, 74*, 1421–1437.

Li, J., & De Luca, R. (2014). Review of assessment feedback. *Studies in Higher Education*, *39,* (2), 378-393.

Lindahl, E. (2007). Gender and ethnic interactions among teachers and students: evidence from Sweden (No. 2007: 25). *Working Paper, IFAU-Institute for Labour Market Policy Evaluation.*

Lindsey, A., King, E., Membere, A., & Cheung, H. K. (2017). Two types of diversity training that really work. *Harvard Business Review* (Available at: http://www.hbr.org)

Lipnevich, A. A., & Smith, J. K. (2009). I really need feedback to learn: Students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability 21,* (4), 347–367*.*

Littleford, L. N., Wright, M. O., & Sayoc-Parial, M. (2005). White students' intergroup anxiety during same-race and interracial interactions: A multimethod approach*. Basic and Applied Social Psychology, 27*, 85–94.

Lizzio, A., & Wilson, K. (2008). Feedback on assessment: students' perceptions of quality and effectiveness. Assessment and Evaluation in Higher Education, 33, (3), 263-275.

Lubker, J. R., Watson II, J. C., Visek, A. J. & Geer, J. R. (2005). Physical appearance and the perceived effectiveness of performance enhancement consultants. The Sport Psychologist, *19*, 446-458.

Maclellan, E. (2001). Assessment for learning: The differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, *26,* (4), 307-318.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: stereotypes on the rebound. *Journal of Personality and Social Psychology, 67,* 808–817.

Macrae, C. N., Bodenhausen, G. V., & Milne, A. B. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology*, *69*, 397–407.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., Thorn, T. M. J., & Castelli, L. (1997). On the activation of social stereotypes: the moderating role of processing objectives. *Journal of Experimental Social Psychology, 33*, 471–489.

Macrae, C. N., & Bodenhausen, G. V. (2000). Thinking categorically about others. *Annual Review of Psychology*, *51*, 93-120.

Macrae, C. N., & Bodenhausen, G. V. (2001). Social Cognition: Categorical Person Perception. *British Journal of Psychology, 92*, 239-255.

Madon, S., Willard, J., Guyll, M., & Scherr, K. C. (2011). Self-fulfilling prophecies: Mechanisms, power and links to social problems. *Social and Personality Psychology Compass, 5*, (8), 578-590.

Malouff, J. M., Emmerton, A. J., & Schutte, N. S. (2013). The risk of a halo bias as a reason to keep students anonymous during grading. *Teaching of Psychology*, 0, 0, 1-5. doi: 10.1177/0098628313487425.

Malouff, J. M. & Thornsteinsson, E. B. (2016) Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 0,* 0, 1-12

Mandhane, N., Ansari, S., Shaikh, T.P., & and Deolekar, S. (2015) Positive feedback: A tool for quality education in field of medicine. *International Journal of Research in Medical Sciences 3,* 8, 1868–1873. doi:10.18203/2320-6012.ijrms20150293.

Manley, A., Greenlees, I., Thelwell, R., & Smith, M. (2010). Athletes' use of reputation and gender information when forming initial expectancies of coaches. *International Journal of Sports Science and Coaching*, *5,* (4), 517-532.

Marsh, H. W. (1990). Causal ordering of academic self-concept on academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology,* 82, 646 – 656. doi:10.1037/0022-0663.82.4.646.

Martin, W. D. (1972). The Sex Factor in Grading Composition. *Research in the Teaching of English*, *6,* (1), 36–47.

Mayring, P. (2007). Mixing qualitative and quantitative methods. *Mixed methodology in psychological research*, 27-36.

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological review*, *90*, 3, 215-238, doi: 10.1037/0033-295X.90.3.215

McArthur, J. (2015). Assessment for social justice: the role of assessment in achieving social justice. *Assessment & Evaluation in Higher Education*, 1-15. http://dx.doi.org/10.1080/02602938.2015.1053429

McDavid, J. W., & Harari, H. H. (1966). Stereotyping of names and popularity in grade-school children. *Child Development*, *37,* 453-459.

McKown, C., & Weinstein, R. S. (2002). Modelling the role of child ethnicity and gender in children’s differential response to teacher expectations. *Journal of Applied Social Psychology, 32,* 159–184.

McLean, A. J., Bond, C H., & Nicholson, H. D. (2014). An anatomy of feedback: a phenomenographic investigation of undergraduate students' conceptions of feedback. *Studies in Higher Education, 40,* (5), 921-932, DOI: 10.1080/03075079.2013.855718

McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: a meta-analysis of the result. *Journal of Applied Psychology*, 85, (2), 314 – 322.

Meadows, M., & Billington, L. (2005). A review of the literature on marking reliability. *Unpublished AQA report produced for the National Assessment Agency*.

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review*, 193-210.

Miller, D. T., & Turnbull, W. (1986). Expectancies and interpersonal processes. *Annual Review of Psychology*, *37*, 233-256.

Miller, A., Imrie, B., & Cox, K. (1998). *Student assessment in higher education: a handbook for assessing performance*. London: Kogan Page.

Modood, T. (2006). Ethnicity, Muslims and higher education entry in Britain. *Teaching in Higher Education, 11*, (2), 247-250.

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A metaanalytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. doi:10.3102/0034654309337522

Morse, J. M., & Niehaus, L. (2009). *Mixed method design: Principles and procedures* (Vol. 4). Left Coast Pr.

Mutch, A. (2003). Exploring the practice of feedback to students. *Active Learning in Higher Education, 4*, (1), 24-38.

Nash, G., Crimmins, G., & Oprescu, F. (2015). If first-year students are afraid of public speaking assessments what can teachers do to alleviate such anxiety? Assessment & Evaluation in Higher Education. doi:10.1080/02602938.2015.1032212.

Nash, R. A., Winstone, N. E., Gregory, S. E. A., & Papps, E. (2018). A memory advantage for past-oriented over future-oriented performance feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/xlm0000549

Neuberg, S. L. (1994). Expectancy-confirmation processes in stereotype-tinged social encounters: The moderating role of social goals. In M. P. Zanna & J. M. Olson (Eds.). *The psychology of prejudice: The Ontario symposium* (pp. 103–130). Hillsdale, NJ: Erlbaum.

Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology*, *56*, 374-386.

Neuberg S. L., Fiske S. T., (1987). Motivational influences on impression formation: outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology, 53*, 431.41.

Newstead, S. E. (1996). The psychology of student assessment. *Bulletin of the British Psychological Society, 9,* 543–547.

Newstead, S. E., & Dennis, I. (1990). Blind marking and sex bias in student assessment. *Assessment & Evaluation in Higher Education*, *15,* (2), 132-139.

Newstead, S. E. (2002). Examining the examiners: Why are we so bad at assessing students? *Psychology Learning and Teaching, 2* (2) 70-75

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education, 35,* (5), 501-517.

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self- regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31,* (2), 199–218.

Nightingale, D., & Cromby, J. (1999). *Social constructionist psychology: A critical analysis of theory and practice.* Buckingham: Open University Press

Noden, P., Shiner, M., & Madood, T. (2014). University offer rates for candidates from different ethnic categories. *Oxford Review of Education, 40,* (3), 349-369, http://dx.doi.org/10.1080/03054985.2014.911724

NUS (National Union of Students) (2008). Mark my words, not my name. Available at: http://www.nus.org.uk/cy/news/mark-my-words-not-my-name/ (Accessed October 2015).

NUS (National Union of Students) (2016). Available at: http://nussl.ukmsl.net/campaigns/highereducation/archived/learning-and-teaching-hub/anonymous-marking/(Accessed February 2016).

Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In A. W. Kruglanski & E. T. Higgens (Eds.) *Social Psychology: Handbook of Basic Principles* (1st Ed), pp.211-238. New York: The Guildford Press.

O'Neill, G. (1985). Self, teacher and faculty assessments of student teaching performance: a second scenario, *The Alberta Journal of Educational Research, 31*, (2), 88-98.

Onwuegbuzie, A. J., & Leech, N. L. (2004). Enhancing the interpretation of significant findings: The role of mixed methods research. *The Qualitative Report*, *9,* (4), 770-792.

Orr, S. (2007). Assessment moderation: Constructing the marks and constructing the students. *Assessment & Evaluation in Higher Education, 32*, (6) 645–56.

Orrell, J. (2006) Feedback on learning achievement: Rhetoric and reality. *Teaching in Higher Education, 11,* (4), 441-456.

Orrell, J. (2008). Assessment beyond belief: the cognitive process of grading. *Balancing Dilemmas in Assessment and Learning in Contemporary Education*, 251-63.

Orsmond, P., Merry, S. & Reiling, K. (2000). The use of student derived marking criteria in peers and self-assessment. *Assessment & Evaluation in Higher Education*, *25*, 23–38.

Orsmond, P., Merry, S., & Reiling, K. (2005). Biology students' utilisation of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education 30*, 369–86.

Orsmond, P., & Merry, S. (2011). Feedback and alignment: Effective and ineffective links between tutors' and students' understanding of coursework feedback. *Assessment & Evaluation in Higher Education, 36*, (2), 125-136.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*. Advance online publication. doi: 10.1037/a0032734

Ouazad, A. (2009). Assessed by a teacher like me: Race, gender and subjective evaluation. *CentrePiece,* 12-13.

Owen, C., Stefaniak, J., & Corrigan, G. (2010). Marking identifiable scripts: Following up student concerns. *Assessment & Evaluation in Higher Education*, 35, (1), 37-44.

Owen, D., Green, A., Pitcher, J., & Maguire, M. (2000). *Minority Ethnic Participation and Achievements in Education, Training and the Labour Market* (Research Report No. 225). London: UK Department for Education and Skills. https://www.education.gov.uk/publications/eOrderingDownload/RR225.pdf.

Packman, T., Brown, G. S., Englert, P., Sisarich, H., & Bauer, F. (2005). Differences in personality traits across ethnic groups with New Zealand and across an International sample. *Differences in Personality Traits across Ethnicities and Countries,* 34, (2), 77-85.

Panadero, E., & Jonsson, A. (2013). Review: The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. doi:10.1016/j.edurev.2013.01.002.

Parks, F. R., & Kennedy, J. H. (2007). The impact of race, physical attractiveness, and gender on Education majors’ and teachers’ perceptions of student competence. *Journal of Black Studies, 37,* (6), 936-943.

Parsons, J. E., Kaczala, C. M., & Meece, J. L. (1982). Socialization of achievement attitudes and beliefs: Classroom influences. *Child Development, 53,* 322-339.

Patton, M. Q. (1980). *Qualitative evaluation and research methods.* Newbury Park, CA: Sage.

Payne, B. K. (2006). Weapon Bias: Split-second decisions and unintended stereotyping. *Current Directions in Psychological Science*, *15*, 287-291.

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38,* 384–396.

Perry-Langdon, N. (1990). Marking by numbers: Evaluation of the marking of final degree examinations in the Faculty of Humanities and Social Studies. *Retrieved from British Library Document Supply Centre- DSC: q94/21236*

Pheterson, G. I., Kiesler, S. B., & Goldberg, P. A. (1971). Evaluation of the performance of women as a function of their sex, achievement, and personal history. *Journal of Personality and Social Psychology, 19,* (1), 114-118.

Piche, G. L., Michellin, M., Rubin, D., & Sullivan, A. (1977). Effects of dialect-ethnicity, social class and quality of written compositions on teachers' subjective evaluations of children. *Communications Monographs, 44*, 60–72.

Pitt, E., & Norton, L. (2016): 'Now that's the feedback I want!' Students' reactions to feedback on graded work and what they do with it. *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2016.1142500

Pitt, E., & Winstone, N. (2018). The impact of anonymous marking on students' perceptions of fairness, feedback and relationships with lecturers, *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2018.1437594

Plessner, H. (2005). Positive and negative effects of prior knowledge on referee decisions in sport. In T. Betsch & S. Haberstroh (Ed.). *The Routines of Decision Making*, 311-324. Mahwah NJ.: Lawrence Erlbaum.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport and Exercise Psychology*, *21*, (2), 131-144.

Poulos, A., & Mahony, M. J. (2009) Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education, 33,* (2), 143-154.

Price, M. (2005). Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education*, *30,* (3), 215-230.

Price, M., & Rust, C. (1999). The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education*, *5,* (2), 133-144.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*, (3), 277-289.

Quality Assurance Agency (2013). UK quality code for higher education: Chapter B6: *Assessment of student and recognition of prior learning.* Gloucester (Available at: http://www.qaa.ac.uk/docs/qaa/quality-code/chapter-b6_-assessment-of-students-and-the-recognition-of-prior-learning.pdf?sfvrsn=9901f781_8

Race, P. (1995). What has assessment done for us – And to us? In P. Knight (Ed.). *Assessment for Learning in Higher Education.* (pp. 61–74). London: SEDA/Kogan Page.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76,* 85 – 97.

Read, B., Robson, J., & Francis, B. (2004). Re-viewing undergraduate writing: tutors' perceptions of essay qualities according to gender. *Research in Post-Compulsory Education*, *9,* (2), 217-238.

Read, B., Francis, B., & Robson, J (2005). Gender, 'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education, 30*, (3), 241-260.

Reay, D. (2000) ‘Dim dross’: Marginalized women both inside and outside the academy. *Women’s Studies International Forum, 23,* (1), 13-21.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*, 435–448. doi:10.1080/02602930902862859.

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*, (3), 137-152.

Richardson, J. T. E. (2008). The Attainment of Ethnic Minority Students in UK Higher Education. *Studies in Higher Education, 33,* (1): 33–48.

Richardson, J. T. E., Alden Rivers, B., Whitelock, D. (2014). The role of feedback in the under-attainment of ethnic minority students: evidence from distance education. *Assessment & Evaluation in Higher Education.* http://dx.doi.org/10.1080/02602938.2014.938317

Richardson, J. T. E. (2015).The under-attainment of ethnic minority students in UK higher education: what we know and what we don’t know. *Journal of Further and Higher Education, 39*, (2), 278-291, DOI: 10.1080/0309877X.2013.858680

Ridge, R. D., & Reber, J. S. (2002). "I Think She's Attracted to Me": The Effect of Men's Beliefs on Women's Behavior in a Job Interview Scenario. *Basic and Applied Social Psychology,* 24, 1, 1-14.

Rigsby, L. C. (1987). Changes in students’ writing and their significance. Paper presented at the meeting of the Conference on College Composition and Communication, Atlanta.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research, 62*, 413-426.

Rist, R. C. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, *40,* (3), 411-451.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research, 62*, 413-426.

Robinson, P., Harrison, M., Law, I., & Gardinier, J. (1992). Ethnic monitoring of university admission: some Leeds findings, *Social Policy and Sociology Working Paper, No. 7* (Leeds: University of Leeds).

Robson, J., Francis, B. & Read, B. (2002). Writes of Passage: Stylistic features of male and female undergraduate history essays. *Journal of Further and Higher Education, 26*, (4), 351-362, DOI: 10.1080/0309877022000021757

Roese, N. J. (1994). The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology*, 66*,* (5), 805 – 818.

Rosengren, K. E. (1981). Advances in Scandinavia content analysis: An introduction. In K. E. Rosengren (Ed.), *Advances in content analysis* (pp. 9-19). Beverly Hills, CA: Sage.

Rosenthal, R. (1973). The mediation of Pygmalion effects: A four factor "theory". *Papua New Guinea Journal of Education, 9*, 1-12.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, *3*, (6), 176-181.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist*, *57*, 839–849.

Rosenthal, R. (2003). Covert communication in laboratories, classrooms, and the truly real world. *Current Directions in Psychological Science*, *12*, (5), 151-154.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development*. Rinehart and Winston.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 1*, 377-386.

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development* (2nd Ed.) Carmarthen: Crown House Publishing.

Rothbart, M., Dalfen, S., & Barrett, R. (1971). Effects of teacher's expectancy on student-teacher interaction. *Journal of Educational Psychology*, *62*, (1), 49-54.

Rowe, A. D., & Wood, L. N. (2008). Student perceptions and preferences for feedback. *Asian Social Science, 4*, (3), 78–88.

Rubin, D. & Greene, K. (1992). Gender-typical style in written language. *Research in the Teaching of English*, *26*, 7–40.

Rubovitz, R., & Maehr, M. (1971). Pygmalion analyzed: Toward an explanation of the Rosenthal-Iacobson findings. *Journal of Personality and Social Psychology, 19,* 197-203.

Rubovits, P. C., & Maehr, M. L. (1973). Pygmalion black and white. *Journal of Personality and Social Psychology, 25*, (2), 210-218.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). “Unlearning” automatic biases: The malleability of implicit prejudice and stereotypes*. Journal of Personality and Social Psychology, 81,* (5), 856-868.

Rust, C. (2007). Towards a scholarship of assessment. *Assessment & Evaluation in Higher Education,* 32, 2, 229-237.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education*, *28,* 2, 147-164.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*, 2, 175-194.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 34*, 159–179.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment and Evaluation in Higher Education*, 35, (5), 535-550.

Sadowski, C. J., & Cogburn, H. E. (1997). Need for cognition in the Big-Five factor structure. *The Journal of Psychology*, *131*, 307-312.

Schaefer, E. (2008) Rater bias patterns in an EFL writing assessment. *Language Testing*, 5, (4), 465-493.

Scherr, K. C., Madon, S., Guyll, M., Willard, J., & Spoth, R. (2011). Self-verification as a mediator of mothers' self-fulfilling effects on adolescents' educational attainment. *Personality and Social Psychology Bulletin*, *37*, 587–600.

Schultz, P. W., & Oskamp, S. (2000). *Social Psychology: An Applied Perspective.* Upper saddle river, NJ: Prentice-Hall

Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genetic, Social, and General Psychology Monographs*. Retrieved August 06, 2013 from HighBeam Research: http://www.highbeam.com/doc/1G1-90681651.html

Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. In E. T. Higgins & R. Sorrentino (Eds.). *Handbook of Motivation and Cognition: Foundations of Social Behaviour* (Vol. 2, pp. 527–561). New York: Guilford.

Schwartz, N. (1998).Warmer and more social: Recent developments in cognitive social psychology. *Annual Review of Sociology*, *24*, 239-264.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513 -523.

Senko, C., Belmonte, K., & Yakhkind, A. (2012). How students' schievement goals shape their beliefs about effective teaching: A 'build-a-professor' study'. *British Journal of Educational Psychology, 82,* (3), 420–435. doi:10.1111/j.2044-8279.2011.02036.x.

Shay, S. B. (2004). The assessment of complex performance: A socially situated interpretive act. *Harvard Educational Review*, *74,* 3, 307-329.

Shay, S.B. (2005). The assessment of complex tasks: A double reading. *Studies in Higher Education. 30*, 663–79.

Shay, S. (2008). Researching assessment as social practice: Implications for research methodology. *International Journal of Educational Research, 47,* 159-164.

Shields, S. (2015). 'My work is bleeding': Exploring students' emotional responses to first-year assignment feedback. *Teaching in Higher Education, 20*, 6, 614-624, http://dx.doi.org/10.1080/13562517.2015.1052786

Shiner, M., & Modood, T. (2002). Help or hindrance? Higher education and the route to ethnic equality. *British Journal of Sociology of Education, 23*, 2, 209-232.

Shiner, M., & Modood, T. (2010). Help or hindrance? Higher education and the route to ethnic equality. *British Journal of Sociology of Education, 23,* 2, 209-232.

Sinclair, R. C. (1988). Mood, categorization breadth, and performance appraisal: The effects of order of information acquisition and affective state on halo, accuracy, information retrieval, and evaluations. *Organizational Behavior and Human Decision Processes*, *42*, 1, 22-46.

Sinclair, R.C., & Mark, M. M. (1992). The influence of mood state on judgement and action: Effects of persuasion, categorization, social justice, person perception and judgemental accuracy. In L. L. Martin & A. Tesser (Eds.). The construction of social judgement (pp. 165-193). Hillsdale, NJ: Erlbaum.

Skipper, Y., & Douglas, K. (2012). Is no praise good praise? Effects of positive feedback on children's and university students' responses to subsequent failures. *British Journal of Educational Psychology, 82*, 2, 327–339. doi:10.1111/j.2044-8279.2011.02028.x.

Smale, G. G. (1977). *Prophecy, behaviour and change: An examination of self-fulfilling prophecies in helping relationships*. London: Routledge.

Small, F., & Attree, K. (2015). Undergraduate student responses to feedback: expectations and experiences. *Studies in Higher Education*, http://dx.doi.org/10.1080/03075079.2015.1007944

Smith, A. E., Jussim, L., Eccles, J., VanNoy, M., Madon, S., & Palumbo, P. (1998). Self-fulfilling prophecies, perceptual biases, and accuracy at the individual and group levels. *Journal of Experimental Social Psychology, 34*, 530– 561.

Smith, E. R. (1998). Mental representation and memory. In D.Gilbert, S.T Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed.), (pp.391-445) Boston: McGraw-Hill.

Smith, E. R., & Mackie, D.M. (2007). Social Psychology (3rd Ed.). Hove: Psychology Press.

Smith, E., & Coombe, K. (2006). Quality and qualms in the marking of university assignments by sessional staff: An exploratory study. *Higher Education*, *51,* 1, 45-69.

Smith, F. (1982). *Understanding Reading* (3rd ed.). New York: Holt, Rinehart & Winston.

Smithers, R. (1999). NUS claims racial bias in exams. The Guardian. Available at: https://www.theguardian.com/uk/1999/mar/10/rebeccasmithers [Accessed 12th May 2015].

Snow, R. E. (1995). Pygmalion and intelligence? *Current Directions in Psychological Science*, *4*, (6), 169-171.

Snyder, M. (1992). Motivational foundations of behavioral confirmation. In M.P. Zanna (Ed.). *Advances in Experimental Social Psychology, 25*, 67–114. Orlando, FL: Academic Press.

Snyder, M. (1984). When belief creates reality. In L. Berkowitz (Ed.). *Advances in Experimental Social Psychology, 18*, 247-305. New York: Academic Press.

Snyder, M., & Swann, W. B., Jr. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology, 14*, 148–162.

Snyder, M., & Haugen, J. A. (1995). Why does behavioral confirmation occur? A functional perspective on the role of the target. *Personality and Social Psychology Bulletin*, *21*, 9, 963-974.

Snyder, M., & Stukas, Jr., A. A. (1998). Interpersonal processes: The interplay of cognitive, motivational, and behavioural activities in social interaction. *Annual Review of Psychology, 50,* 273-303.

Sparkes, A. C., & Smith, B. (2014). *Qualitative research methods in sport, exercise and health.* London: Routledge.

Spence, J. T., Deaux, K., & Helmreich, R. L. (1985). Sex roles in contemporary American society. In G. Lindzey & E. Aronson (Eds*.), Handbook of Social Psychology* (3rd ed.), (pp.149-178). New York: Random House.

Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C., & Dunn, M. (1998). Automatic activation of stereotypes: the role of self-image threat. *Personality and Social Psychology Bulletin, 24,* 1139–1152.

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. *Empirical Economics, 45* 1,523–538..

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons. Some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660-1672.

Srull, T. K. & Wyer, R. S. (1980). Category accessibility and social perception. Some implications for the study of person memory and interpersonal judgment. *Journal of Personality and Social Psychology, 38*, 841-856.

Stemmler, S. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation*, *7*, (7, pp.1-9.

Stock, P. L., & Robinson, J. L. (1987). Taking on testing: Teachers as tester researchers. *English Education, 19*, 93-121.

Stroessner, S. J. (1996). Social categorization by race or sex: effects of perceived non-normalcy on response times. *Social Cognition, 14*, 247–76.

Sudcamp, A., Kaiser, J., & Moller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*, 3, 743-762.

Swann, W. B., & Ely, R. J. (1984). A battle of wills: self-verification versus behavioral confirmation. *Journal of Personality and Social Psychology*, *46*, 1287 – 1302.

Swansea University (2018). Available at: http://www.swansea.ac.uk/academic-services/academic-guide/assessment-issues/marking/

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105*, 3, 409-429.

Tajfel, H. (1979). Individuals and groups in social psychology. *British Journal of Social and Clinical Psychology*, *18,* 2, 183-190.

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in the social and behavioral sciences.* Thousand Oaks, CA: Sage.

Taylor, S. E. (1998). The social being in social psychology. In D. Gilbert., S.T Fiske., & G. Lindzey. (Eds.) *Handbook of Social Psychology,* (4th ed.), (pp. 58-95). New York: McGraw- Hill.

Terborg, J.R. & Illgen, D.R. (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. *Organizational Behaviour and Human Performance, 13,* 352-376.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. Journal of Educational Psychology, *99*, 253–273.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. Bristol, PA: Falmer.

Thomas, V., & Miles, S. (1995). Psychology of Black women: Past, present, and future. In H. Ladrine (Ed.). *Bringing cultural diversity to feminist psychology.* Washington, DC: American Psychological Association.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, (1), 25-29.

Thorndike, E. L. (1968). Review of Robert Rosenthal and Lenore Jacobson, Pygmalion in the classroom. *American Educational Research Journal*, *5,* 708-711.

Thorpe, M. (2000). Encouraging students to reflect as part of the assignment process: Student responses and tutor feedback. *Active Learning in Higher Education, 1,* (1), 79-92.

Tight, M. (2004). Research into higher education: an a-theoretical community of practice? *Higher Education Research & Development, 23*, 4, 395-411, DOI:10.1080/0729436042000276431

Tight, M. (2013). Discipline and methodology in higher education research. *Higher Education Research & Development, 32,* 1, 136-151, DOI:10.1080/07294360.2012.750275

Tight, M. (2014). Discipline and theory in higher education research. *Research Papers in Education, 29,* 1, 93-110, DOI: 10.1080/02671522.2012.729080

Towler, A., & Dipboye, R. L. (2006). Effects of trainer reputation and trainees' need for cognition on training outcomes. *The Journal of Psychology*, *140*, 549-564.

Trautwein, U., Ludtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology, 90,* 334–349. doi:10.1037/0022-3514.90.2.334

Tsiplakides, I., & Keramida, A. (2011). The relationship between teacher expectation and student achievement in the teaching of English as a foreign language. *English Language Teaching*, www.ccsenet.org/elt

Turner, G., & Gibbs, G. (2010). Are assessment environments gendered? An analysis of the learning responses of male and female students to different assessment environments. *Assessment & Evaluation in Higher Education*, *35,* (6), 687-698.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Vail, K. (2005). What's in a name? Perhaps a student's grade. *American School Board Journal*, *192*, (8), 6-8.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The Implicit Prejudiced Attitudes of Teachers Relations to Teacher Expectations and the Ethnic Achievement Gap. *American Educational Research Journal*, *47,* (2), 497-527.

Van Dinther, M., Dochy, F., & Segers, M. (2011). Factors affecting students’ self-efficacy in higher education. *Educational Research Review*, 6, 95-108.

Van Fleet, D. D. & Atwater, L. (1997). Gender neutral names: Don’t be so sure! *Sex Roles, 37*, (1/2), 111-123.

Van Matre, J. C., Valentine, J. C., & Cooper, H. (2000). Effect of students’ after-school activities on teachers’ academic expectancies. *Contemporary Educational Psychology, 25*, 167–183.

Van Ryn, M., & Fu, S.S. (2003). Paved with good intentions: Do public health and human service providers contribute to racial/ethnic disparities in health? *American Journal of Public Health*, *93*, (2), 248-255.

Von Baeyer, C. L., Sherk, D. L., & Zanna, M. P. (1981). Impression Management in the Job Interview When the Female Applicant Meets the Male (Chauvinist) Interviewer. *Personality and Social Psychology Bulletin*, *7*, 45-51.

Vorauer, J., & Kumhyr, S. M. (2001). Is this about you or me? Self- versus other-directed judgment and feeling in response to intergroup interaction. *Personality and Social Psychology Bulletin, 27,* 706–719.

Walker, M. (2009). An investigation into written comments on assignments: Do students find them usable? *Assessment & Evaluation in Higher Education, 34,* (1), 67-78.

Wallston, B. S., & O'Leary, V. E. (1981). Sex makes a difference: Differential perceptions of women and men. *Review of Personality and Social Psychology*, *2*, 9-41.

Warr, P. B., & Knapper, C. (1968). The perception of people and events. Oxford: John Wiley and Sons.

Weaver, M. R. (2006) Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education, 31,* (3), pp.379-394.

Weber, R. P. (1990). *Basic content analysis* (2nd Ed.). Beverly Hills, CA: Sage.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, *101*, 34-52.

Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.). *Handbook of Social Psychology* (4th Ed. Vol. 1) (pp. 446 - 496). New York: McGraw-Hill.

Weiss, R. (2000). Emotion and learning. *Training and Development, 54,* (11), 44–48.

White, K. J., Jones, K., & Sherman, M. D. (1998). Reputation information and teacher feedback: Their influences on children's perceptions of behavior problem peers. *Journal of Social and Clinical Psychology*, *17*, 11-37.

Wiers, R. W., Van De Luitgaarden, J., Van Den Wildenberg, E., & Smulders, F. T. Y. (2005). Challenging implicit and explicit alcohol-related cognitions in young heavy drinkers. *Addiction*, *100*, 806–819.

Wilke, A., & Mata, R. (2012). Cognitive Bias. In: V.S. Ramachandran (Ed.). The *Encyclopedia of Human Behavior, 1,* pp. 531-535. Academic Press.

Williams, J. E., & Best, D. L. (1982). *Measuring sex stereotypes: A thirty-nation study.* Beverly Hills, CA: Sage Publications.

Willig, C. (2008). *Introducing qualitative research in psychology* (2nd Ed.). New York: McGraw Hill.

Winstone, N. E., Nash, R. A., Rowntree, J., & Menezes, R. (2016). What do students want most from written feedback information? Distinguishing necessities from luxuries using a budgeting methodology. *Assessment & Evaluation in Higher Education, 41*, 1237–1253. http://dx.doi.org/10.1080/02602938.2015.1075956

Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': Barriers to university students' feedback seeking and recipience. *Studies in Higher Education, 42,* 2026–2041. http://dx.doi.org/10.1080/03075079.2015.1130032

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72*, (2), 262-274.

Woolf, A. (1995). *Competence-based assessment.* Buckingham: Open University Press.

Woolf, H. (2004) Assessment criteria: reflections on current practices. *Assessment & Evaluation in Higher Education, (29),* 4, 479-493, DOI: 10.1080/02602930310001689046

Wyer, N. A., Sherman, J. W., & Stroessner, S. J. (1998). The spontaneous suppression of racial stereotypes. *Social Cognition, 16*, 340–52.

Yang, M., & Carless, D. (2013). The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education, 18,* (3), 285–297.

Yorke, M., Bridges, P., & Woolf, H. (2000). Mark distributions and marking practices in UK higher education. *Active Learning in Higher Education 1*, (1) 7–27.

Zadney, J., & Gerard, H. B. (1974). Attributed intentions and informational selectivity. *Journal of Experimental Social Psychology*, *10*, 34-52.

Zanna, M. P., & Pack, S. J. (1975). On the self-fulfilling nature of apparent sex differences in behavior. *Journal of Experimental Social Psychology*, *11*, (6), 583-591.

9 APPENDICES

9.1 APPENDIX I: Brown, Gibbs, & Glover's (2003) Amended Feedback System

Feedback categories	Examples
Identifying errors	Underlined or circled words; 'X'; '?'; 'No'
Giving praise	Ticks; 'Good'; 'Excellent'
Correcting errors	Corrected grammar, spellings, dates or individual numerical data
Explaining misunderstandings	'This data is out of date. Recent data shows … '; 'Don't forget … which suggests … '; 'Using … Shows … '
Demonstrating correct practice	Underlined or crossed-out sentences or phrases together with a replacement version as a marginal comment; crossed-out whole paragraphs, tables or diagrams with a suggested alternative structure for these as a marginal comment.
Engaging students in thinking	'Why?'; 'Is this logical?'; 'Does this follow?'; 'Is this relevant?'; 'Meaning?'; 'Is there an alternative?'
Suggesting further study	'See … for more information'; 'Information on … is absent'
Justifying marks	'This assignment was given a Grade point 4 because….'; 'I could not award a higher mark because….'
Suggesting approaches to future assignments	'In future essays you should….'; 'Next time…..'

9.2 APPENDIX II: Kumar and Stracke's (2007) Feedback classification system

Examples of Categorisation		
Referential	Editorial	Please get rid of spaces.
	Organisation	This does not belong in the literature review.
	Content	Are you sure you can make such a claim?
Directive	Suggestion	Maybe this is not necessary.
	Question	Whose term is this?
	Instruction	Please clarify
Expressive	Praise	Good, nice example.
	Criticism	This table… does not add to the text
	Opinion	I would be interested to explore what triggered this.

9.3 APPENDIX III: Hyatt's (2005) Feedback classification system

Phatic comments	- Comment - Encouragement	Comments whose purpose is the establishment and maintenance of a good academic and social relationship between the tutor and the student. They are divided into two types. **Comment:** The tutor writes generally on the content indicating interest, surprise and so on at what has been written: 'This is a well-presented and well-written assignment'; 'John you write in a very eloquent and engaging style'. **Encouragement**: The tutor offers comments that are intended to encourage the student in future work: 'I hope you find these comments useful and best of luck with the rest of the course!'
Developmental comments	- Alternatives - Future - Reflective questions - Informational comments	These comments are made by the tutor with the intention of aiding the student with subsequent work in relation to the current assignment. **Alternatives:** The tutor offers alternatives, suggestions and recommendations in place of what the students has written or points out omissions in the student's work: 'It would have been helpful if you had indicated briefly what these counterarguments were'. **Future:** These are comments on how the student needs to address a point directly in subsequent work: 'This is a point to think about for the future'. **Reflective questions:** Here the tutor poses a question, as opposed to making a direct point, for the student to consider reflectively: 'It's important to consider limitations - were these the only ones?'. **Informational comment**. Here the tutor offers a direct comment on a related and complementary topic, with the intention of offering the student additional academic insight into the topic under discussion.
Structural Comments	- Discourse level - Sentence level	These comments refer to the structural organisation of the assignment, either as a whole or in sections. **Discourse level:** These comments consider the organisation of the assignment as a whole in terms of the constituent sections - introduction, literature review, conclusion. These comments may consider how each of these constituent sections may be put together, in terms of rhetorical moves, or how they themselves may fit together to give a

		structure to the overall assignment (coherence). "This introduction covers the main structural elements of aims, scope and sequence. It would have been improved by…" **Sentence level**: These comments look at the organisation of individual sentences, in terms of length, relation to other sentences (cohesion) and so on: 'This sentence is a good signpost to the next section and thus a good structural point'
Stylistic Comments	• Punctuation • Lexis • Syntax/word order/ grammar • Referencing/citation/ quotation/bibliography • Presentation • Register	These comments consider the use and presentation of academic language within the assignment. Areas under consideration include the following. **Punctuation:** 'Be careful with commas. They can make a big difference to readability!' **Lexis:** 'This is an example of what X would term a ''dangling'' particle, i.e. it appears as if the ''taking care'' is the action of the paper rather than part of the case you make.' **Syntax/word order/grammar:** 'This needs to be worded more clearly'; 'This is not a complete sentence'. **Proofreading/spelling:** 'Always proofread to check spellings, particularly of authors that can slip through a spellchecker.' **Referencing/citation/quotation/bibliography:** 'This needs a reference - you always need a reference when you offer a theorist's major claims.' **Presentation:** Comments cover page numbering, subtitling, figures, tables, captions, footnotes, endnotes, contents pages, word length, acronyms and so on: 'This table could have been presented much more clearly. Each column really requires its own heading.' **Register:** These comments relate to appropriate language within a particular context (what, who and how of a text) - this would include such aspects as voice, audience and purpose of the text: 'A relatively informal style can be fine but I feel that sometimes you have slipped a little too far into a ''casual'' style…'
Content-Related Comments	• Positive evaluation • Negative evaluation • Non-evaluative summary	This section includes comments on the content of the assignment in terms of their appropriateness/ accuracy or their inappropriateness/inaccuracy. These divide into three categories: **Positive evaluation.** Here, comments on the strengths of the assignment are noted and tend to include features such as: synthesis of literature, theory and practice; appropriate synthesis of personal experience; clear argumentation; and reflection.

		Negative evaluation. Comments here are on weaknesses in the assignment, which may include a deficit in the above features as well as problems relating to the provision of evidence, lack of clarity or the need for clarification, or a lack of criticality in the work: 'Generally there is a need to substantiate claims based on more solid evidence than simply one's feelings about what is going on.' **Non-evaluative summary.** Comments here non-evaluatively offer a summary of aspects of the assignment: 'This project aims to measure the degree of success of a specific teaching intervention by statistical analysis of the results of pre- and post-tests'
Methodological Comments	• Approach • Procedures • Process	This section only applies to feedback on research-based assignments, where the presence or absence of appropriate discussion on aspects of the research design and analysis are discussed. **Approach:** Here, comments may be made on the philosophical and epistemological positions of the research, and how these relate to the research paradigm through which the enquiry is approached, and the researcher's consideration of positionality. **Procedures:** Here, comments are made on practical aspects of the research design/the collection and analysis of the data, the sample, recording and so on, including the researcher's criticality in these discussions. **Process:** Here, comments are made on the process, timeframe and practicality of the conduct of the research and might include issues such as piloting, distribution, nonresponse and problems encountered, including the researcher's criticality in these discussions: 'The question that comes to me is how has your data ''provided'' you with the insights you discuss.'
Administrative Comments		These comments relate to the administrative procedures of the course: 'Please submit two copies of the assignment in future.'

9.4 APPENDIX IV: Assessment Pack

Variations in Marking Practice across the Assessment Process

Demographic Questionnaire

Please fill in your details in the spaces provided.

Age__________ (Years)

Sex	Male	□
(Please Tick)	Female	□

Ethnicity	Asian/Asian British – Bangladeshi	□
(Please Tick)	Asian/Asian British – Pakistani	□
	Asian/Asian British – Indian	□
	Chinese	□
	Other Asian Background	□
	Mixed – White & Black Caribbean	□
	Mixed – White & Black African	□
	Mixed – White & Asian	□
	Black/Black British – African	□
	Black/Black British – Caribbean	□
	Other Black Background	□
	White - British	□
	White - Irish	□
	Other White Background	□

Other Mixed Background □

Other Ethnic Background □

Job Title ____

Number of Years' Experience in Higher Education	Less than Six Months	□
(Please Tick)	Six Months – One Year	□
	One Year – Two Years	□
	Two Years – Five Years	□
	Five Years – Ten Years	□
	More than Ten Years	□

Approximate Annual Marking Workloads	Less than Fifty Essays	□
(Please Tick)	Fifty – One Hundred Essays	□
	One Hundred – Two Hundred Essays	□
	Two Hundred – Five Hundred Essays	□
	Five Hundred – One Thousand Essays	□
	More than One Thousand Essays	□

ESSAY 1

SAMUEL JONES

Directions

Please use the Standardised Assessment Criteria Profile (ACP) tick sheet to help you award a final mark for the subsequent essay. Please also give feedback (in line with current teaching practice) on the essay as if the work was to be ***returned to the student***. The following essay was submitted for assessment on a first year Introduction to Research Methods module.

Drugs in Sport should be legalised

So why do athletes take drugs in sport? It has been claimed by Weinberg and Gould (2007) that top athletes take such drugs as anabolic steroids to enhance their performance whilst competing, to help cope with pain during working with an injury and to control their weight. Mottram (2003) explains drugs are something that interacts with the body and changes how the body naturally acts. Even though many drugs within sport are illegal, athletes still take them, but for what, the fame, the medals, the sponsorship? Still the underlying remains, should drugs in sport be legalised?

Links between drug taking and health problems have been found, Mottram (1996) reports that in the 1960's athletes died from amphetamine abuse due to respiratory/cardiac arrest. Although through the years, the progressions of performance enhancing drugs have improved dramatically yet still have damaging side effects. It is stated by Weinberg and Gould (2007), that stimulants have such side effects like anxiety, insomnia and in severe cases, death. Additionally anabolic steroids have just as serious side effects like liver disease and heart disease. If the legalisation of drugs in sport is allowed then the sporting hero's of today will not have a fulfilled life after their career. The opportunities for them to go on and help the foundation levels in sport participation are then gone due to the athletes not being well enough to make such visits to publically run organisations.

In contrast, if drugs were legalised, would this make it easier? Doctors would be able to monitor the amount consumed by the athletes this was suggested by Toohey and Veal (2007), they also stated that from doing this, it could then go on to reduce deaths from drug taking in sports. If this was the case then doctors would then be able to advise athletes what drugs do actually help with performance and direct them away from the potentially really harmful drugs. Also the money saved from doing the expensive drug testing could be then put back into the development of performance drugs in order for doctors to have a greater knowledge of the drugs out there and how to go on and progress the drugs further.

Many argue that drugs in sports should remain prohibited due to the message it sends out to young audience that are influenced so much by sporting role models within the media. As Weinberg and Gould (2007), go on to explain that athletes are seen not just through the television but other sources such as the internet. Due to ongoing development of technology nowadays, it is much easier for teenagers to access such articles in the press, therefore if drug taking by athletes was in the media constantly, this would then send out the wrong message to the impressionable young children that look up to their 'hero's' and allow them to think drug taking was a good thing and not think about the negativities. This could then lead onto much harsher more damaging drugs such as

Cocaine because they think it is okay because the athletes they see across the media are also doing drugs.

Legalisation of drugs in sport would bring added excitement within competitions such as the Olympic Games and Commonwealth Games. This would be due to the ongoing progression of performance enhancing drugs leading to performances getting better and adding more excitement to events. From this, more spectators would want to watch such events through being at the venues or watching it through television/internet. Additionally from more followers of each sporting event, the sport participation levels locally in communities would increase due to the inspiration they are receiving from the athletes the watch doing so well. Therefore, a healthier nation from doing sport would reduce money spent on the NHS dealing with the growing problem that is obesity in the United Kingdom.

As an overview, taking drugs in sport could be seen as cheating in many people's opinions and many athletes that have been found out to be cheating or have been caught have been stripped of medals and in worst case scenarios been banned from competing in any upcoming events such as Dwaine Chambers. Drug taking is seen as cheating due to the athletes not using their own ability and using the drugs to enhance their performances. Consequently, through drug use in sport it is seen to be ethically wrong and the whole achievement of winning a gold medal would be based upon what drugs make your body do, not what your body is naturally able to do.

However, if the legalisation of drugs were to happen then all the competitors would be at a level playing field, meaning they would all stand a chance instead of a clear favourite such as Usian Bolt and Beijing Olympics, 2008. All athletes would have an equal opportunity to be the gold medallist they aspire to be. Toohey and Veal (2007) explain that drug taking doesn't have to be done by all athletes, but they would all have a choice if the ban was lifted. Furthermore, if all athletes stood in with a chance of winning, the more likely it would be for world records to be broken on a regular basis, also creating more excitement.

In conclusion, drugs should be legalised is seen to be a very controversial topic and many people express different opinions upon the subject. Looking at the topics in the different lights does tend to make you think if it would be easier if drugs were legalised? However, the topic that never goes unnoticed is the health issues that go hand in hand with drug taking in sport. "Tell them that by taking drugs, what they would be doing would literally be DYING TO WIN." (Waddington, Smith, 2009:17). How could the legalisation of drugs in sport be justified when faced with that?! Therefore for that main reason, drugs should not be legalised in sport. The risks upon athletes' health are not worth the ongoing development of performance enhancing drugs neither the excitement it may bring; it is just something to think about.

References

Waddington, I., & Murphy, P. (1992). Drugs, Sport and Ideologies. In E. Dunning & C. Rojek (eds.). *Sport and Leisure in the Civilising Process: Critique and Counter-Critique.* Basingstoke: Macmillan.

Mottram, D.R (1996) Side effects of drugs. Drugs in Sport (2nd ed.) London : Routledge

Sleap, M. (1998). Drug Abuse in Sport. In *Social Issues in Sport.* Basingstoke: Macmillan.

Cashmore, E. (2000). Champs or Cheats? Drug Use and Attempts to Contain It. In *Making Sense of Sports* (3rd ed.). London: Routledge.

Coakley, J. J. (2001). Deviance in Sports. In *Sport in Society: Issues and Controversies* (7th ed.). Boston: Irwin McGraw-Hill.

Mottram, D.R (2003) Side effects of drugs. Drugs in Sport (3rd ed.) London : Routledge

Morgan, W.J (2007) Ethics in Sport (2nd ed.) Champaign IL.: Human Kinetics.

Toohey, K and Veal, AJ (2007). The Olympic Games: A Social Science Perspective (2nd ed.). Trowbridge : Cromwell Press

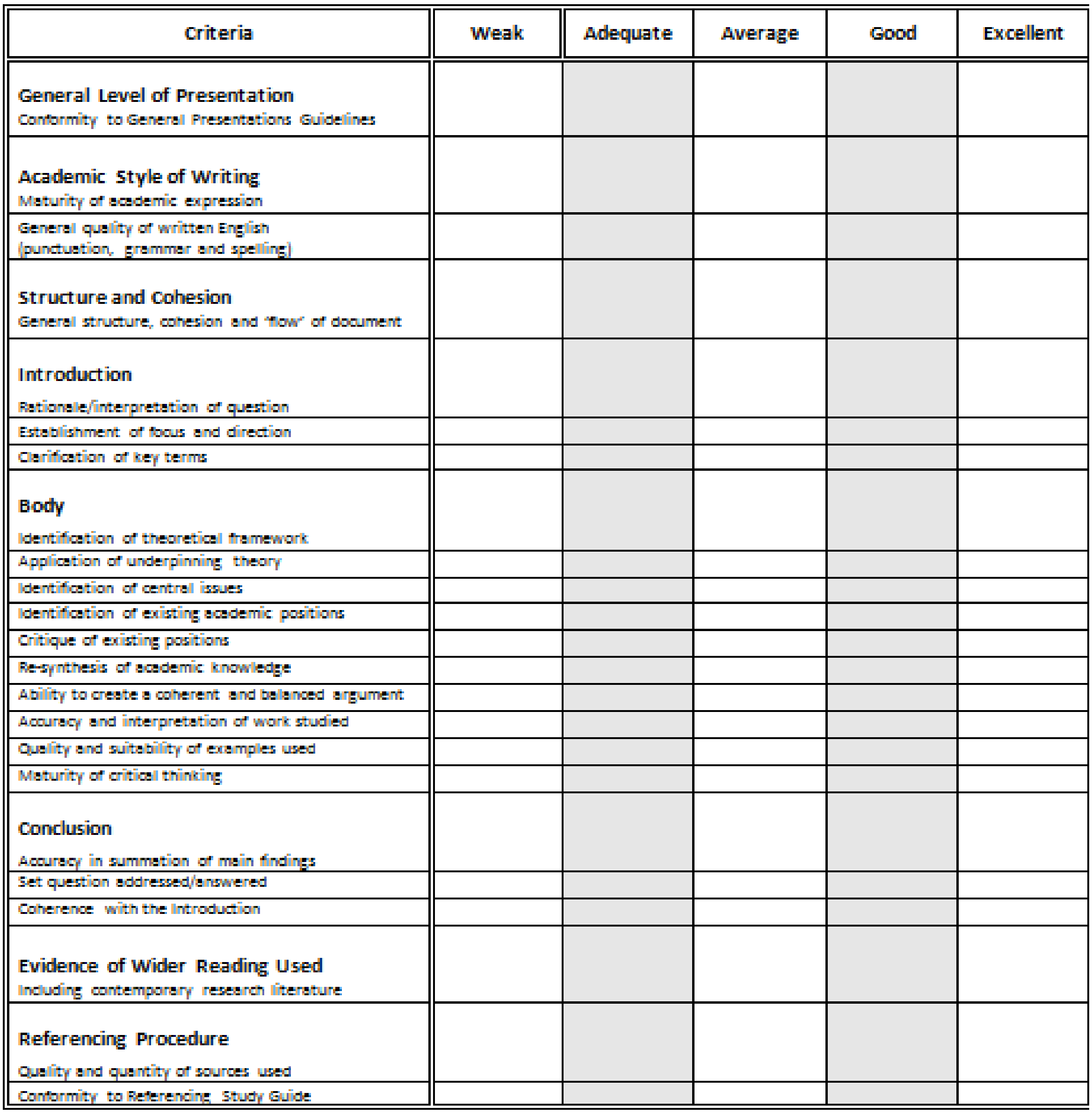
Weinberg, R. & Gould, D. (2007) *Foundations of Sport and Exercise Psychology.* (4th Ed). Champaign IL.: Human Kinetics.

Kristensen, J.K and Petersen, T.S (2009) Should Athletes Be Allowed to Use

All Kinds of Performance-Enhancing Drugs?—A Critical Note on Claudio M. Tamburrini [online] Available:http://web.ebscohost.com/ehost/pdf?vid=5& hid=7&sid=dc1a327d-b232-4923-b3d7-5ec591e0d5ed%40sessionmgr14 [1st Nov. 2009]

Waddington, I and Smith, A (2009) An Introduction to Drugs in Sport. Addicted to Winning? Oxon : Routledge.

Assessment Criteria Profile



Criteria	Weak	Adequate	Average	Good	Excellent
General Level of Presentation Conformity to General Presentations Guidelines					
Academic Style of Writing Maturity of academic expression					
General quality of written English (punctuation, grammar and spelling)					
Structure and Cohesion General structure, cohesion and "flow" of document					
Introduction Rationale/interpretation of question					
Establishment of focus and direction					
Clarification of key terms					
Body Identification of theoretical framework					
Application of underpinning theory					
Identification of central issues					
Identification of existing academic positions					
Critique of existing positions					
Re-synthesis of academic knowledge					
Ability to create a coherent and balanced argument					
Accuracy and interpretation of work studied					
Quality and suitability of examples used					
Maturity of critical thinking					
Conclusion Accuracy in summation of main findings					
Set question addressed/answered					
Coherence with the Introduction					
Evidence of Wider Reading Used Including contemporary research literature					
Referencing Procedure Quality and quantity of sources used					
Conformity to Referencing Study Guide					

Classification Awarded   Fail   3rd Class   2:2   2:1   1st Class

Final Mark Awarded ________ %

What factors influenced your perceptions of the essay?

ESSAY 2

JAMES SMITH

Directions

Please use the Standardised Assessment Criteria Profile (ACP) tick sheet to help you award a final mark for the subsequent essay. Please also give feedback (in line with current teaching practice) on the essay as if the work was to be ***returned to the student***. The following essay was submitted for assessment on a first year Introduction to Research Methods module

Drugs in sport should be legalised

There are numerous arguments both for and against the legalisation of drugs in sport. In the course of this essay I will explore the most controversial of these arguments before concluding whether or not they should be legalised. Mottram, (2003) explains that a drug is a chemical substance that when taken by humans can lead to a change in the way the body functions.

Sport is played by billions of people around the world as a way to facilitate relaxation and enjoyment as well as promote an extensive, healthy lifestyle. However, others lose sight of this purpose by risking everything and taking drugs, such as sports men and women who practice doping using illegal methods to gain an advantage. Medical research has proven that the thousands of professional athletes, who take these prohibited substances may be physically and mentally damaged by drugs (Dimeo, 2007) what's more Gifford, (2004) states that as a result of drug use athletes can suffer life threatening health problems, an example of this is Birgit Dressel, who finished fourth in the heptathlon in April 1986. She suffered serious health problems as a result of taking steroids throughout her athletic career which resulted in her death in 1987.

A strong argument in favour of the legalisation of drugs is that they can be used recreationally. Sports minister Richard Caborn (BBC, [online] 12th December 2006) has commented that he believes sports men and women should not be banned for using "social" drugs. Social drugs such as Cocaine and Amphetamines are illegal and it could be said that these drugs should be left for the authorities to deal with as in most cases they are just used as a way to relax or provide a 'buzz' (Gifford 2004). Australian Rugby player Wendell Sailor said that he did not cheat but had succumbed to the temptation of a "so-called party drug"(BBC, [online] 12th December 2006). As well as making you feel depressed and run down, taking the drug may give you an air of overconfidence and result in the taking of unnecessary risks, in these circumstances it can't be said that it would affect your sporting performance in any positive manner. Becket (1988) comments that: "If we started looking at the social aspect of drug taking then we would not be doing our job (cited in an introduction to drugs in sport, 2009, p.36) which backs up the theory that if these drugs are not performance enhancing then athletes, at their own risk, should be legally permitted to take these recreational drugs.

However, it can also be argued that the short term affects of this drug can be positive upon a performer. Dr Gary Wadler, (2007) stated that although the acute affects would probably impair rather than enhance performance, within a two hour window the drug could overcome fatigue, reaction time and therefore could be performance enhancing (cited in Olympics, 2007[online]). In the late 1970's an American footballer, Hollywood Henderson carried a nasal inhaler filled with cocaine and water during games. It was said that he used this to help adopt a sense of euphoria and self confidence that scientists believe this drug can aid (Adler, 2007[online]). This is neither natural, nor fair that one performer gets a competitive advantage over another. Dimeo, (2007) comments that: "sport should be about fair play, all drugs are a form of cheating" (p.3). Therefore there are circumstances when it can be argued that a drug will positively assist a performer and as a consequence should not be legalised in sport.

On the other hand, another very influential argument especially in a modern society, against the legalisation of drugs in sport, is the concept of role models. Younger generations look up to the elite performers such as Mutu, as universal role models which they aspire to be like. By taking drugs, sports people are essentially condoning drug use; the younger generation imitate the actions of their idols and by taking drugs, they are setting a poor example to these teenagers who may then follow suit. Moreover, Gifford, (2004) states that sport is about reaching personal goals, but getting there by hard work and tenacity as opposed to taking drugs that upset the concept of fair play. Serena Williams, quoted by Gifford (2004, p.49), says "When I was a kid I dreamed of becoming the best. Drugs kill dreams. It's just not worth it." This reinforces the fact that the younger generation aspire to follow in the footsteps of professionals, and for them to witness their idols taking these drugs leaves them feeling disheartened and angry toward both the sport itself and the individual. This provides a strong argument against the legalisation of drugs in sport.

Alternatively it is not fair that the performer is labelled as a drugs cheat, if taking drugs for legitimate therapeutic use, for example to cure a common cold or in the case of Paddy Kenny to fight a chest infection. Paddy was given a nine month ban simply because he was negligent in failing to consult his club doctor or reading the accompanying leaflet given when purchasing the cough medicine over the counter (The Daily Mail[online] 7th September 2009). What's more, (Mottram 2005) suggests that when an athlete has a more serious medical condition such as epilepsy or diabetes it would be implausible for them to partake in sport without regular treatment with drugs. This harsh reality probes for a more relaxed attitude towards the legalisation of certain drugs in sports.

By legalising drugs in sport, this would help to combat various underlying problems relating to drug testing. One problem in particular is the excessive costs that face the anti-doping authorities, due to the extensive scale and cost of testing such a vast amount of people for such a large range of different substances. Furthermore, a positive test result can only be evidence that the athlete has been exposed to the substance found, yet it cannot be evidence to confirm it was taken intentionally in order to enhance performance (Mottram 2005). Moreover, McBay (1987) highlights the fact that a test result is not always completely reliable (cited in Mottram, 2005, p.336). What's more, reducing the list of banned substances and cutting down the cost spent on drugs testing may allow authorities more money and time directing young athletes away from drug use (Gifford, 2005). Therefore dedicating all this time and money to drugs testing is not ideal when the result may not always be accurate or reliable. This serves to promote the argument in favour of the legalisation of drugs in sport.

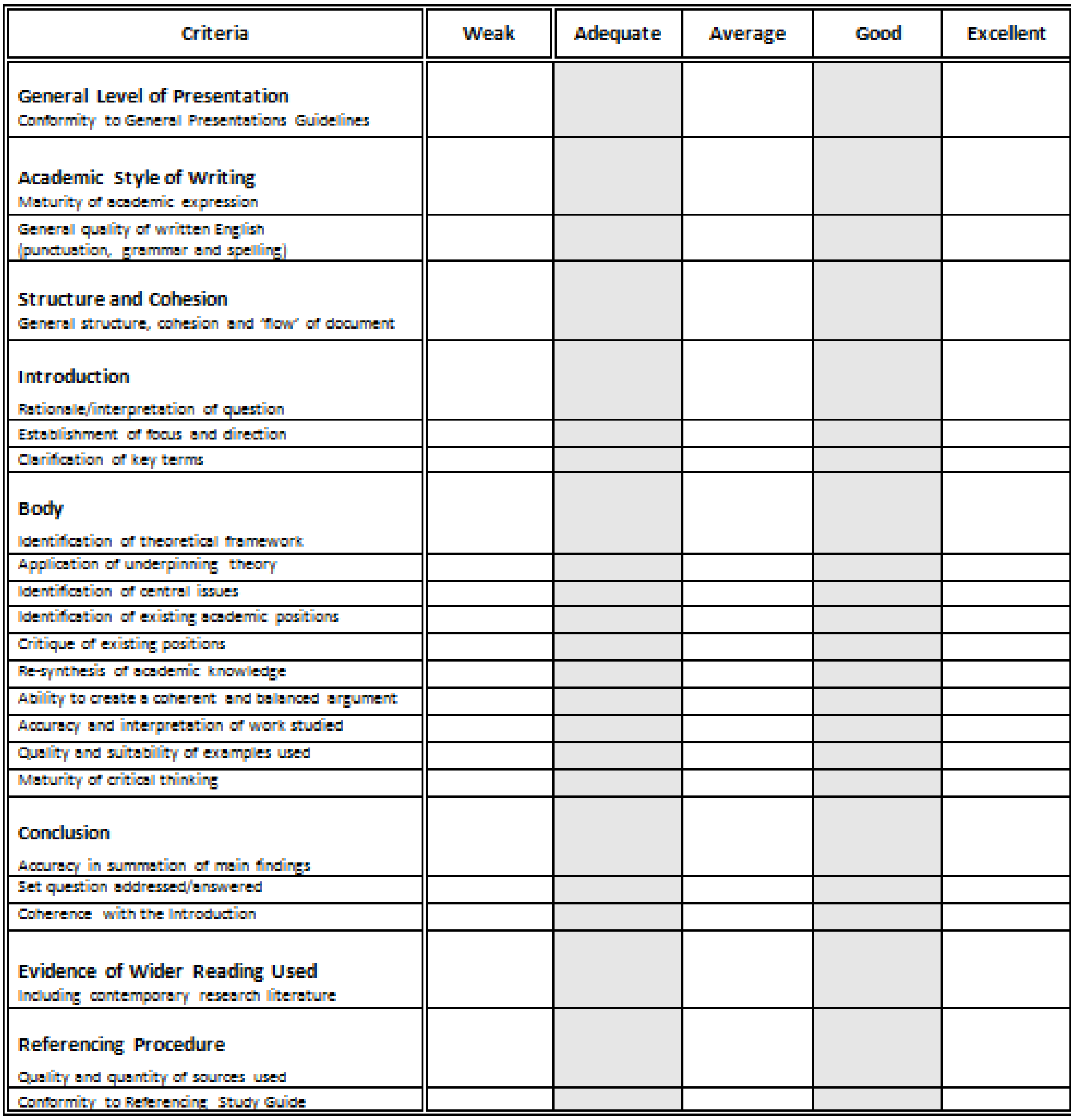
The idea of legalising drugs in sport is at the centre of an extremely controversial debate. While many would argue that the legalisation of drugs in a sporting environment have positive short term effects on a performer, others would argue that the long term health problems fashioned as a result of drug taking undermine any positive effects. Moreover, the expense of drug testing while accounting for the fact these results are not always accurate, coupled with the damaging effect it has on a young up and coming generation, prompts many to fight against legalising these drugs in sport. On balance, the solution may lie in the form of a compromise; while it is hard to justify legalising drugs with such damaging effects in relation to health, it is much easier to justify the legalisation of drugs used for legitimate therapeutic use as they have a more indirect affect on the actual sporting performance.

References

- Alper, J. (2007) Olympics, Is Cocaine a Performance Enhancing Drug?, [online]. Available: http://olympics.fanhouse.com/2007/11/02/is-cocaine-a-performance-enhancing-drug/ (28th October 2009)
- BBC (2006) Sport 'social drugs' ban queried. BBC [online]. Available: http://news.bbc.co.uk/1/hi/uk_politics/6171777.stm (28th October 2009)
- Dimeo, P. (2007) A history of drug use in sport 1876 – 1976 (1st edition). Oxon: Routledge.
- Gifford C. (2004) Drugs and sport. Oxford: Heinemann.
- Mottram, D.R (2003) Drugs in sport (3rd edition). London: Routledge.
- Mottram, D.R. (2005) Drugs in Sport (4th edition). Oxon: Routledge.
- Waddington, I and Smith, A. (2009) an introduction to drugs in sport (1st edition). Oxon: Routledge.

- Young, C. (2009) Paddy Kenny banned from football for nine months after failed drugs test, Mailonline [online]. Available://www.dailymail.co.uk/sport/footbal/article-1211781/Paddy-Kenny-football-months-failed-drugs-test.html (28th October 2009)

Assessment Criteria Profile



Criteria	Weak	Adequate	Average	Good	Excellent
General Level of Presentation Conformity to General Presentations Guidelines					
Academic Style of Writing Maturity of academic expression					
General quality of written English (punctuation, grammar and spelling)					
Structure and Cohesion General structure, cohesion and 'flow' of document					
Introduction Rationale/interpretation of question					
Establishment of focus and direction					
Clarification of key terms					
Body Identification of theoretical framework					
Application of underpinning theory					
Identification of central issues					
Identification of existing academic positions					
Critique of existing positions					
Re-synthesis of academic knowledge					
Ability to create a coherent and balanced argument					
Accuracy and interpretation of work studied					
Quality and suitability of examples used					
Maturity of critical thinking					
Conclusion Accuracy in summation of main findings					
Set question addressed/answered					
Coherence with the Introduction					
Evidence of Wider Reading Used Including contemporary research literature					
Referencing Procedure Quality and quantity of sources used					
Conformity to Referencing Study Guide					

Classification Awarded Fail 3rd Class 2:2 2:1 1st Class

Final Mark Awarded _________ %

What factors influenced your perceptions of the essay?

9.5 APPENDIX V: Initial coding

HIGHER ORDER THEMES

Type of comment	Details
Critical feedback	Includes comments that identify errors in the essay but fail to provide further information as to how the student may avoid making the same mistake in the future, or without an attempt to correct the error. All 'error' comments to be grouped here. For example, the marker circled or underlined a word/sentence/paragraph and identified the type of error that has been made, but failed to provide student with constructive feedback or edit the word/sentence/paragraph accordingly.
Ambiguous feedback	Includes comments that were either illegible or the researchers were unable to interpret.
Constructive feedback	Includes comments that identify an area for improvement, and provide the student with advice/ course of action for future essays or edits the word/sentence/paragraph accordingly.

LOWER ORDER THEMES

Type of comment	Details
Punctuation error	Indicates punctuation error (circled/underlined) but does not edit or provide information as to how to correct error.
Grammatical error	Indicates grammatical error but does not edit or provide information as to how to correct error.
Abbreviation error	Identifies abbreviated word (circled/underlined), but does not edit or provide information as to how to correct error.
Spelling error	Identifies spelling mistake but does not edit/correct
Informal language error	Identifies use of informal language (e.g. colloquial expression) or poor academic style, but does not edit or provide information as to how to correct error.
Citation error	Identifies citation error, such as a formatting issue or incorrect author, but does not edit or provide information as to how to correct error.
Format error	
Negative	Enters a cross or discounts the statement(s).
Positive	Enters a tick next to a statement, sentence, citation, or paragraph.
Punctuation correction	Indicates punctuation error (circled/underlined), and edits punctuation or provides information as to how to correct error.
Grammatical correction	Indicates grammatical error (circled/underlined), and edits grammar or provides information as to how to correct error.
Abbreviation correction	Identifies abbreviated word, and edits or provides information as to how to correct error.

Spelling correction	Identifies spelling mistake, and edits/corrects accordingly.
Informal language correction	Identifies use of informal language (e.g. colloquial expression) or poor academic style, and edits or provides information as to how to correct error.
Citation correction	Identifies citation error, such as a formatting issue or incorrect author, edits or provides information as to how to correct error.
Format correction	
Offers alternative wording	Provides student with alternative words that could be used instead of existing words, or incorporated into the sentence.
Offers information	Provides student with examples of incidents, investigations, references, and/or theory that they could have acknowledged in an attempt to bolster the arguments put forward.
Request for evidence	Requests the student for a supporting reference or to indicate a source that has informed the statement(s).
Request for wider reading	Over-reliance upon a limited number of sources.
Quality of source	Identifies poor quality of source, such as a secondary citation or a non-peer reviewed publication.
Request for theory	Requests the student to identify and explain theory that may support the statement(s).
Request for information	Requests the student for additional information, such as definitions and examples.
Request for explanation	Requests the student to further explain their statement.
Request for elaboration	Requests the student to continue with their statement in order to complete the issue/point they are trying to raise.
Request for clarity	Suggests to the student that their statement has not been clear, and requires further attention in order to clarify what they mean.
Request for rephrasing	Indicates to student that a sentence or paragraph requires further attention, rephrasing and/or restructuring.
Challenges statement	Disagree with student, and/or provides questions that may enable the student to contemplate issues surrounding their statement.
Praise	Provides the student with a comment praising a sentence, paragraph or aspect of the essay.

9.6 APPENDIX VI: Hyatt's (2005) Amended Feedback classification system

Hyatt's Amended Feedback Classification System (amendments in red italics)		
Feedback Category	**Feedback Subcategories**	**Further Subcategories & Explanation**
Phatic Comments	• Comment • Encouragement	Comments whose purpose is the establishment and maintenance of a good academic and social relationship between the tutor and the student. They are divided into two types. **Comment:** The tutor writes generally on the content indicating interest, surprise and so on at what has been written: 'This is a well-presented and well-written assignment'; 'John you write in a very eloquent and engaging style'. **Encouragement**: The tutor offers comments that are intended to encourage the student in future work: 'I hope you find these comments useful and best of luck with the rest of the course!'
Developmental comments	• Alternatives • Future • Reflective questions • Informational comments	These comments are made by the tutor with the intention of aiding the student with subsequent work in relation to the current assignment. **Alternatives:** The tutor offers alternatives, suggestions and recommendations in place of what the students has written or points out omissions in the student's work: 'It would have been helpful if you had indicated briefly what these counterarguments were'. **Future:** These are comments on how the student needs to address a point directly in subsequent work: 'This is a point to think about for the future'. **Reflective questions:** Here the tutor poses a question, as opposed to making a direct point, for the student to consider reflectively: 'It's important to consider limitations - were these the only ones?'. **Informational comment**. Here the tutor offers a direct comment on a related and complementary topic, with the intention of offering the student additional academic insight into the topic under discussion.
Structural Comments/ ***Corrections***	• Discourse level • Sentence level	These comments refer to the structural organisation of the assignment, either as a whole or in sections. **Discourse level:** These comments consider the organisation of the assignment as a whole in terms of the constituent sections - introduction, literature review, conclusion. These comments may consider how each of

		these constituent sections may be put together, in terms of rhetorical moves, or how they themselves may fit together to give a structure to the overall assignment (coherence). “This introduction covers the main structural elements of aims, scope and sequence. It would have been improved by…” **Sentence level**: These comments look at the organisation of individual sentences, in terms of length, relation to other sentences (cohesion) and so on: ‘This sentence is a good signpost to the next section and thus a good structural point’
Stylistic Feedback	Stylistic feedback containing **stylistic comments**, ***stylistic corrections*** or ***stylistic emphasis*** related to • Punctuation • Lexis • Syntax/word order /grammar • Referencing/citation/ quotation/bibliography • Presentation • Register	These comments, corrections and emphases consider the use and presentation of academic language within the assignment. Areas under consideration include the following. **Punctuation:** ‘Be careful with commas. They can make a big difference to readability!’ **Lexis:** ‘This is an example of what X would term a ‘‘dangling’’ particle, i.e. it appears as if the ‘‘taking care’’ is the action of the paper rather than part of the case you make.’ **Syntax/word order/grammar:** ‘This needs to be worded more clearly’; ‘This is not a complete sentence’. **Proofreading/spelling:** ‘Always proofread to check spellings, particularly of authors that can slip through a spellchecker.’ **Referencing/citation/quotation/bibliography:** ‘This needs a reference - you always need a reference when you offer a theorist’s major claims.’ **Presentation:** Comments cover page numbering, subtitling, figures, tables, captions, footnotes, endnotes, contents pages, word length, acronyms and so on: ‘This table could have been presented much more clearly. Each column really requires its own heading.’ **Register:** These comments relate to appropriate language within a particular context (what, who and how of a text) - this would include such aspects as voice, audience and purpose of the text: ‘A relatively informal style can be fine but I feel that sometimes you have slipped a little too far into a ‘‘casual’’ style…’
Content-related Feedback	Content-related feedback containing **Content-related Comments**, ***Content-related Symbols*** or ***Content-related Emphases*** pertaining to: • Positive evaluation • Negative evaluation	This section includes comments, corrections or emphases on the content of the assignment in terms of their appropriateness/ accuracy or their inappropriateness/ inaccuracy. These divide into three categories: **Positive evaluation.** Here, comments on the strengths of the assignment are noted and

	- Non-evaluative summary	tend to include features such as: synthesis of literature, theory and practice; appropriate synthesis of personal experience; clear argumentation; and reflection. **Negative evaluation.** Comments here are on weaknesses in the assignment, which may include a deficit in the above features as well as problems relating to the provision of evidence, lack of clarity or the need for clarification, or a lack of criticality in the work: 'Generally there is a need to substantiate claims based on more solid evidence than simply one's feelings about what is going on.' **Non-evaluative summary.** Comments here non-evaluatively offer a summary of aspects of the assignment: 'This project aims to measure the degree of success of a specific teaching intervention by statistical analysis of the results of pre- and post-tests'
Methodological Comments		***This section was removed on the basis that Hyatt stated,*** "This section only applies to feedback on research-based assignments, where the presence or absence of appropriate discussion on aspects of the research design and analysis are discussed".
Administrative Comments		These comments relate to the administrative procedures of the course: 'Please submit two copies of the assignment in future.'

9.7 APPENDIX VII: Steps for a hierarchical content analysis

Steps for a hierarchical content analysis (Sparkes & Smith, 2014)

Procedural Steps	Explanation
1) Immersion	This involves getting a sense of the database and becoming intimately familiar with it. This can be done by reading the interview transcripts on numerous occasions or listening to the recorded interview several times from an empathic view point. The combination of immersion and adopting an empathic position is what Maykut and Morehouse (1994) described as the posture of indwelling.
2) Search for, identify and label themes in each case	Search for and identify raw data themes characterising each participants' responses. To help with this the raw data is first tagged to obtain a set of concepts representative of the information collected. Idiographic profiles of each individual can also be developed.
3) Connecting and ordering themes	Independently cluster the raw data themes into meaningful categories that seem to connect and fit together. This analysis results in a cluster of raw data themes within categories of greater generality (sub-themes). These themes are then classified into larger, more inclusively meaningful clusters (higher-order themes and general dimensions), with each given a title that represents the themes contained within each category.
4) Cross-checking	The raw data themes and clusters are thoroughly examined again. The investigator or investigators who were present during data collection return to the original transcribed data and verify that all themes and categories were represented.
5) Confirmation	An investigator who was not present during data collection, but has experience in qualitative research, reviews the analysis.
6) Produce a table	The results are ordered in the form of a table or figure. The table or figure should be designed to display the hierarchical nature of the themes generated.

9.8 APPENDIX VIII: In-text Feedback Results Tables

White British Male Control (Group 1a)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	18	38.3%	5.5%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	19	40.4%	5.8%
Developmental Comment: Informational	10	21.3%	3.0%
Total Developmental Comments	47		14.3%
Structural Comment: Discourse Level	0	0.0%	0.0%
Structural Comment: Sentence Level	10	100.0%	3.0%
Total Structural Comments	10		3.0%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural Feedback	10		3.0%
Stylistic Comment: Punctuation	3	2.3%	0.9%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	30	23.4%	9.1%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	44	34.4%	13.4%
Stylistic Comment: Presentation	22	17.2%	6.7%
Stylistic Comment: Register	23	18.0%	7.0%
Stylistic Comment: Proof Reading	6	4.7%	1.8%
Total Stylistic Comments	128		39.0%
Stylistic Correction: Punctuation	27	45.0%	8.2%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	24	40.0%	7.3%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	7	11.7%	2.1%
Stylistic Correction: Presentation	0	0.0%	0.0%
Stylistic Correction: Register	0	0.0%	0.0%
Stylistic Correction: Proof Reading	2	3.3%	0.6%
Total Stylistic Corrections	60		18.3%
Stylistic Emphasis: Punctuation	6	16.2%	1.8%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	13	35.1%	4.0%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	7	18.9%	2.1%
Stylistic Emphasis: Presentation	7	18.9%	2.1%
Stylistic Emphasis: Register	1	2.7%	0.3%
Stylistic Emphasis: Proof Reading	3	8.1%	0.9%
Total Stylistic Emphasis Comments	37		11.3%
Overall Stylistic Feedback	225		68.6%
Content-related Comment: Positive Evaluation	2	6.3%	0.6%
Content-related Comment: Negative Evaluation	28	87.5%	8.5%
Content-related Comment: Non-Evaluative Summary	2	6.3%	0.6%
Total Content-related Comments	32		9.8%
Content-related Symbol: Positive	10	90.9%	3.0%

Content-related Symbol: Negative	1	9.1%	0.3%
Total Content-related Symbol Comments	11		3.4%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%
Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Feedback	43		13.1%
Administrative Comment	1		0.3%
Total Number of feedback Contributions	343		100.0%

White British Male Experimental (Group 1a)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	1	100.0%	0.29%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	1		0.29%
Developmental Comment: Alternative	32	42.1%	9.3%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	34	44.7%	9.9%
Developmental Comment: Informational	10	13.2%	2.9%
Total Developmental Comments	76		22.2%
Structural Comment: Discourse Level	2	13%	0.6%
Structural Comment: Sentence Level	14	88%	4.1%
Total Structural Comments	16		4.7%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	16		4.7%
Stylistic Comment: Punctuation	1	0.9%	0.3%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	7	0.0%	2.0%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	62	57.9%	18.1%
Stylistic Comment: Presentation	12	11.2%	3.5%
Stylistic Comment: Register	24	22.4%	7.0%
Stylistic Comment: Proof Reading	1	0.9%	0.3%
Total Stylistic Comments	107		31.2%
Stylistic Correction: Punctuation	28	45.9%	8.2%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	16	26.2%	4.7%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	4	6.6%	1.2%
Stylistic Correction: Presentation	4	6.6%	1.2%
Stylistic Correction: Register	5	8.2%	1.5%
Stylistic Correction: Proof Reading	4	6.6%	1.2%
Total Stylistic Corrections	61		17.8%
Stylistic Emphasis: Punctuation	10	71.0%	2.9%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	1	7.0%	0.3%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	3	21.0%	0.9%
Stylistic Emphasis: Presentation	0	0.0%	0.0%
Stylistic Emphasis: Register	0	0.0%	0.0%
Stylistic Emphasis: Proof Reading	0	0.0%	0.0%
Total Stylistic Emphasis Comments	14		4.1%
Overall Stylistics	182		53.1%
Content-related Comment: Positive Evaluation	27	56.0%	7.9%
Content-related Comment: Negative Evaluation	19	40.0%	5.5%
Content-related Comment: Non-Evaluative Summary	2	4.0%	0.6%
Total Content-related Comments	48		14.0%
Content-related Symbol: Positive	20	100.0%	5.8%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	20		5.8%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	68		20.0%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	343		100.0%

White British Female Control (Group 1b)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	2	100.0%	0.8%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	2		0.8%
Developmental Comment: Alternative	15	40.5%	6.0%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	21	56.8%	8.4%
Developmental Comment: Informational	1	2.7%	0.4%
Total Developmental Comments	37		14.9%
Structural Comment: Discourse Level	2	25%	0.8%
Structural Comment: Sentence Level	6	75%	2.4%
Total Structural Comments	8		3.2%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	8		3.2%
Stylistic Comment: Punctuation	4	5.5%	1.6%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	12	16.4%	4.8%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	33	45.2%	13.3%
Stylistic Comment: Presentation	14	19.2%	5.6%
Stylistic Comment: Register	10	13.7%	4.0%
Stylistic Comment: Proof Reading	0	0.0%	0.0%
Total Stylistic Comments	73		29.3%
Stylistic Correction: Punctuation	13	35.1%	5.2%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	15	40.5%	6.0%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	3	8.1%	1.2%
Stylistic Correction: Presentation	0	0.0%	0.0%
Stylistic Correction: Register	5	13.5%	2.0%
Stylistic Correction: Proof Reading	1	2.7%	0.4%
Total Stylistic Corrections	37		14.9%
Stylistic Emphasis: Punctuation	8	26.0%	3.2%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	12	39.0%	4.8%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	0	0.0%	0.0%
Stylistic Emphasis: Presentation	2	6.0%	0.8%
Stylistic Emphasis: Register	1	3.0%	0.4%
Stylistic Emphasis: Proof Reading	8	26.0%	3.2%
Total Stylistic Emphasis Comments	31		12.4%
Overall Stylistics	141		56.6%
Content-related Comment: Positive Evaluation	4	9.0%	1.6%
Content-related Comment: Negative Evaluation	42	91.0%	16.9%
Content-related Comment: Non-Evaluative Summary	0	0.0%	0.0%
Total Content-related Comments	46		18.5%
Content-related Symbol: Positive	14	100.0%	5.6%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	14		5.6%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	1	100.0%	0.4%
Overall Content-related Total	61		24.5%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	249		100.0%

White British Female Experimental (Group 1b)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.05	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	21	58.3%	10.0%
Developmental Comment: Future	1	2.8%	0.5%
Developmental Comment: Reflective question	7	19.4%	3.0%
Developmental Comment: Informational	7	19.4%	3.0%
Total Developmental Comments	36		16.0%
Structural Comment: Discourse Level	2	33.0%	1.0%
Structural Comment: Sentence Level	4	67.0%	2.0%
Total Structural Comments	6		3.0%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	6		3.0%
Stylistic Comment: Punctuation	1	2.1%	0.5%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	8	17.0%	3.7%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	14	29.8%	6.4%
Stylistic Comment: Presentation	10	21.3%	4.6%
Stylistic Comment: Register	13	27.7%	5.9%
Stylistic Comment: Proof Reading	1	2.1%	0.5%
Total Stylistic Comments	47		21.0%
Stylistic Correction: Punctuation	27	54.0%	12.0%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	15	30.0%	7.0%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	3	6.0%	1.0%
Stylistic Correction: Presentation	2	4.0%	1.0%
Stylistic Correction: Register	0	0.0%	0.0%
Stylistic Correction: Proof Reading	3	6.0%	1.0%
Total Stylistic Corrections	50		23.0%
Stylistic Emphasis: Punctuation	1	7.0%	0.5%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	5	36.0%	2.3%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	3	21.0%	1.0%
Stylistic Emphasis: Presentation	4	29.0%	2.0%
Stylistic Emphasis: Register	1	7.0%	0.5%
Stylistic Emphasis: Proof Reading	0	0.0%	0.0%
Total Stylistic Emphasis Comments	14		6.0%
Overall Stylistics	111		51%
Content-related Comment: Positive Evaluation	15	43.0%	7.0%
Content-related Comment: Negative Evaluation	20	57.0%	9.0%
Content-related Comment: Non-Evaluative Summary	0	0.0%	0.0%
Total Content-related Comments	35		16.0%
Content-related Symbol: Positive	31	100.0%	14.0%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	31		14.0%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	66		30.0%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	219		

Asian Male Control (Group 2a)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	23	50.0%	6.8%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	21	45.7%	6.2%
Developmental Comment: Informational	2	4.3%	0.6%
Total Developmental Comments	46		13.6%
Structural Comment: Discourse Level	6	43.0%	1.8%
Structural Comment: Sentence Level	8	57.0%	2.4%
Total Structural Comments	14		4.2%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	14		4.2%
Stylistic Comment: Punctuation	0	0.0%	0.0%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	31	26.7%	9.2%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	44	37.9%	13.1%
Stylistic Comment: Presentation	10	8.6%	3.0%
Stylistic Comment: Register	31	26.7%	9.2%
Stylistic Comment: Proof Reading	0	0.0%	0.0%
Total Stylistic Comments	116		34.4%
Stylistic Correction: Punctuation	21	35.6	6.2%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	29	49.2%	8.6%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	1	1.7%	0.3%
Stylistic Correction: Presentation	0	0.0%	0.0%
Stylistic Correction: Register	3	5.1%	0.9%
Stylistic Correction: Proof Reading	5	8.5%	1.5%
Total Stylistic Corrections	59		17.5%
Stylistic Emphasis: Punctuation	9	30.0%	2.7%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	5	17.0%	1.5%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	4	13.0%	1.2%
Stylistic Emphasis: Presentation	5	17.0%	1.5%
Stylistic Emphasis: Register	2	7.0%	0.6%
Stylistic Emphasis: Proof Reading	5	17.0%	1.5%
Total Stylistic Emphasis Comments	30		8.9%
Overall Stylistics	205		60.8%
Content-related Comment: Positive Evaluation	15	25.0%	4.5%
Content-related Comment: Negative Evaluation	44	75.0%	13.1%
Content-related Comment: Non-Evaluative Summary	0	0.0%	0.0%
Total Content-related Comments	59		17.5%
Content-related Symbol: Positive	7	78.0%	2.1%
Content-related Symbol: Negative	2	22.0%	0.6%
Total Content-related Symbol Comments	9		2.7%
Content-related Criticism: Negative Evaluation	3		0.9%

Content-related Emphasis: Negative Evaluation	0		0.0%
Overall Content-related Total	71		21.1%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	337		100.0%

Asian Male Experimental (Group 2a)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	29	61.7%	9.0%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	14	29.8%	4.3%
Developmental Comment: Informational	4	8.5%	1.2%
Total Developmental Comments	47		14.6%
Structural Comment: Discourse Level	2	20.0%	0.6%
Structural Comment: Sentence Level	8	80.0%	2.5%
Total Structural Comments	10		3.1%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	10		3.1%
Stylistic Comment: Punctuation	5	5.6%	1.6%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	10	11.1%	3.1%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	33	36.7%	10.2%
Stylistic Comment: Presentation	15	16.7%	4.7%
Stylistic Comment: Register	26	28.9%	8.1%
Stylistic Comment: Proof Reading	1	1.1%	0.3%
Total Stylistic Comments	90		28.0%
Stylistic Correction: Punctuation	40	50.0%	12.4%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	17	21.3%	5.3%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	1	1.3%	0.3%
Stylistic Correction: Presentation	12	15.0%	3.7%
Stylistic Correction: Register	6	7.5%	1.9%
Stylistic Correction: Proof Reading	4	5.0%	1.2%
Total Stylistic Corrections	80		24.8%
Stylistic Emphasis: Punctuation	2	18.0%	0.6%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	0	0.0%	0.0%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	4	36.0%	1.2%
Stylistic Emphasis: Presentation	3	27.0%	0.9%
Stylistic Emphasis: Register	2	18.0%	0.6%
Stylistic Emphasis: Proof Reading	0	0.0%	0.0%
Total Stylistic Emphasis Comments	11		3.4%
Overall Stylistics	181		56.2%
Content-related Comment: Positive Evaluation	35	58.0%	10.9%
Content-related Comment: Negative Evaluation	22	37.0%	6.8%
Content-related Comment: Non-Evaluative Summary	3	5.0%	0.9%
Total Content-related Comments	60		18.6%
Content-related Symbol: Positive	23	96.0%	7.1%
Content-related Symbol: Negative	1	4.0%	0.3%
Total Content-related Symbol Comments	24		7.5%
Content-related Criticism: Negative Evaluation	0		0.0%
Content-related Emphasis: Negative Evaluation	0		0.0%

Overall Content-related Total	84		26.1%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	322		100.0%

Asian Female Control (Group 2b)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	30	48.4%	12.2%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	23	37.1%	9.3%
Developmental Comment: Informational	9	14.5%	3.7%
Total Developmental Comments	62		25.2%
Structural Comment: Discourse Level	2	25.0%	0.8%
Structural Comment: Sentence Level	6	75.0%	2.4%
Total Structural Comments	8		3.3%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	8		3.3%
Stylistic Comment: Punctuation	2	3.7%	0.8%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	21	38.9%	8.5%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	26	48.1%	10.6%
Stylistic Comment: Presentation	1	1.9%	0.4%
Stylistic Comment: Register	3	5.6%	1.2%
Stylistic Comment: Proof Reading	1	1.9%	0.4%
Total Stylistic Comments	54		22.0%
Stylistic Correction: Punctuation	11	21.2%	4.5%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	14	26.9%	5.7%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	23	44.2%	9.3%
Stylistic Correction: Presentation	0	0.0%	0.0%
Stylistic Correction: Register	0	0.0%	0.0%
Stylistic Correction: Proof Reading	4	7.7%	1.6%
Total Stylistic Corrections	52		21.1%
Stylistic Emphasis: Punctuation	2	25.0%	0.8%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	1	13.0%	0.4%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	1	13.0%	0.4%
Stylistic Emphasis: Presentation	0	0.0%	0.0%
Stylistic Emphasis: Register	3	38.0%	1.2%
Stylistic Emphasis: Proof Reading	1	13.0%	0.4%
Total Stylistic Emphasis Comments	8		3.3%
Overall Stylistics	114		46.3%
Content-related Comment: Positive Evaluation	3	8.0%	1.2%
Content-related Comment: Negative Evaluation	32	86.0%	13.0%
Content-related Comment: Non-Evaluative Summary	2	5.0%	0.8%
Total Content-related Comments	37		15.0%
Content-related Symbol: Positive	18	95.0%	7.3%
Content-related Symbol: Negative	1	5.0%	0.4%
Total Content-related Symbol Comments	19		7.7%
Content-related Criticism: Negative Evaluation	0		0.0%

Content-related Emphasis: Negative Evaluation	0		0.0%
Overall Content-related Total	56		22.8%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	246		100.0%

Asian Female Experimental (Group 2b)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	2	100.0%	0.8%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	2		0.8%
Developmental Comment: Alternative	26	49.1%	10.7%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	23	43.4%	9.5%
Developmental Comment: Informational	4	7.5%	1.7%
Total Developmental Comments	53		21.9%
Structural Comment: Discourse Level	2	29.0%	0.8%
Structural Comment: Sentence Level	5	71.0%	2.1%
Total Structural Comments	7		2.9%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	7		2.9%
Stylistic Comment: Punctuation	1	3.0%	0.4%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	4	12.1%	1.7%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	21	63.6%	8.7%
Stylistic Comment: Presentation	0	0.0%	0.0%
Stylistic Comment: Register	6	18.2%	2.5%
Stylistic Comment: Proof Reading	1	3.0%	0.4%
Total Stylistic Comments	33		13.6%
Stylistic Correction: Punctuation	34	72.3%	14.0%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	8	17.0%	3.3%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	2	4.3%	0.8%
Stylistic Correction: Presentation	1	2.1%	0.4%
Stylistic Correction: Register	1	2.1%	0.4%
Stylistic Correction: Proof Reading	1	2.1%	0.4%
Total Stylistic Corrections	47		19.4%
Stylistic Emphasis: Punctuation	0	0.0%	0.0%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	1	17.0%	0.4%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	4	67.0%	1.7%
Stylistic Emphasis: Presentation	0	0.0%	0.0%
Stylistic Emphasis: Register	1	17.0%	0.4%
Stylistic Emphasis: Proof Reading	0	0.0%	0.0%
Total Stylistic Emphasis Comments	6		2.5%
Overall Stylistics	86		35.5%
Content-related Comment: Positive Evaluation	13	43.0%	5.4%
Content-related Comment: Negative Evaluation	17	57.0%	7.0%
Content-related Comment: Non-Evaluative Summary	0	0.0%	0.0%
Total Content-related Comments	30		12.4%
Content-related Symbol: Positive	64	100.0%	26.4%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	64		26.4%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	94		38.8%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	242		100.0%

Chinese Male Control (Group 3a)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	8	36.4%	2.8%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	13	59.1%	4.6%
Developmental Comment: Informational	1	4.5%	0.4%
Total Developmental Comments	22		7.8%
Structural Comment: Discourse Level	2	22.0%	0.7%
Structural Comment: Sentence Level	7	78.0%	2.5%
Total Structural Comments	9		3.2%
Structural Corrections: Sentence Level	1	10.0%	0.4%
Overall Structural	10		3.5%
Stylistic Comment: Punctuation	1	1.0%	0.4%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	18	17.3%	6.4%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	58	55.8%	20.5%
Stylistic Comment: Presentation	13	12.5%	4.6%
Stylistic Comment: Register	13	12.5%	4.6%
Stylistic Comment: Proof Reading	1	1.0%	0.4%
Total Stylistic Comments	104		36.7%
Stylistic Correction: Punctuation	26	42.6%	9.2%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	28	45.9%	9.9%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	3	4.9%	1.1%
Stylistic Correction: Presentation	1	1.6%	0.4%
Stylistic Correction: Register	2	3.3%	0.7%
Stylistic Correction: Proof Reading	1	1.6%	0.4%
Total Stylistic Corrections	61		21.6%
Stylistic Emphasis: Punctuation	5	22.0%	1.8%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	2	9.0%	0.7%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	8	35.0%	2.8%
Stylistic Emphasis: Presentation	4	17.0%	1.4%
Stylistic Emphasis: Register	0	0.0%	0.0%
Stylistic Emphasis: Proof Reading	4	17.0%	1.4%
Total Stylistic Emphasis Comments	23		8.1%
Overall Stylistics	188		66.4%
Content-related Comment: Positive Evaluation	12	30.0%	4.2%
Content-related Comment: Negative Evaluation	26	65.0%	9.2%
Content-related Comment: Non-Evaluative Summary	2	5.0%	0.7%
Total Content-related Comments	40		14.1%
Content-related Symbol: Positive	22	96.0%	7.8%
Content-related Symbol: Negative	1	4.0%	0.4%
Total Content-related Symbol Comments	23		8.1%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	63		22.3%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	283		100.0%

Chinese Male Experimental (Group 3a)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	13	56.5%	5.6%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	8	34.8%	3.4%
Developmental Comment: Informational	2	8.7%	0.9%
Total Developmental Comments	23		9.9%
Structural Comment: Discourse Level	0	0.0%	0.0%
Structural Comment: Sentence Level	3	100.0%	1.3%
Total Structural Comments	3		1.3%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	3		1.3%
Stylistic Comment: Punctuation	5	8.5%	2.1%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	0	0.0%	0.0%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	35	59.3%	15.0%
Stylistic Comment: Presentation	4	6.8%	1.7%
Stylistic Comment: Register	13	22.0%	5.6%
Stylistic Comment: Proof Reading	2	3.4%	0.9%
Total Stylistic Comments	59		25.3%
Stylistic Correction: Punctuation	29	60.4%	12.4%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	3	6.3%	1.3%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	16	33.3%	6.9%
Stylistic Correction: Presentation	0	0.0%	0.0%
Stylistic Correction: Register	0	0.0%	0.0%
Stylistic Correction: Proof Reading	0	0.0%	0.0%
Total Stylistic Corrections	48		20.6%
Stylistic Emphasis: Punctuation	6	40.0%	2.6%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	0	0.0%	0.0%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	3	20.0%	1.3%
Stylistic Emphasis: Presentation	6	40.0%	2.6%
Stylistic Emphasis: Register	0	0.0%	0.0%
Stylistic Emphasis: Proof Reading	0	0.0%	0.0%
Total Stylistic Emphasis Comments	15		6.4%
Overall Stylistics	122		52.4%
Content-related Comment: Positive Evaluation	25	68.0%	10.7%
Content-related Comment: Negative Evaluation	12	32.0%	5.2%
Content-related Comment: Non-Evaluative Summary	0	0.0%	0.0%
Total Content-related Comments	37		15.9%
Content-related Symbol: Positive	48	100.0%	20.6%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	48		20.6%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	85		36.5%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	233		100.0%

Chinese Female Control (Group 3b)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	17	33.3%	4.5%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	25	49.0%	6.7%
Developmental Comment: Informational	9	17.6%	2.4%
Total Developmental Comments	51		13.6%
Structural Comment: Discourse Level	4	31.0%	1.1%
Structural Comment: Sentence Level	9	69.0%	2.4%
Total Structural Comments	13		3.5%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	13		3.5%
Stylistic Comment: Punctuation	4	3.4%	1.1%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	17	14.4%	4.5%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	75	63.6%	20.1%
Stylistic Comment: Presentation	6	5.1%	1.6%
Stylistic Comment: Register	10	8.5%	2.7%
Stylistic Comment: Proof Reading	6	5.1%	1.6%
Total Stylistic Comments	118		31.6%
Stylistic Correction: Punctuation	26	36.1%	7.0%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	28	38.9%	7.5%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	8	11.1%	2.1%
Stylistic Correction: Presentation	3	4.2%	0.8%
Stylistic Correction: Register	3	4.2%	0.8%
Stylistic Correction: Proof Reading	4	5.6%	1.1%
Total Stylistic Corrections	72		19.3%
Stylistic Emphasis: Punctuation	15	27.0%	4.0%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	10	18.0%	2.7%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	18	33.0%	4.8%
Stylistic Emphasis: Presentation	0	0.0%	0.0%
Stylistic Emphasis: Register	9	16.0%	2.4%
Stylistic Emphasis: Proof Reading	3	5.0%	0.8%
Total Stylistic Emphasis Comments	55		14.7%
Overall Stylistics	245		65.5%
Content-related Comment: Positive Evaluation	6	16.0%	1.6%
Content-related Comment: Negative Evaluation	31	82.0%	8.3%
Content-related Comment: Non-Evaluative Summary	1	3.0%	0.3%
Total Content-related Comments	38		10.2%
Content-related Symbol: Positive	26	100.0%	7.0%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	26		7.0%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	64		17.1%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	374		100.0%

Chinese Female Experimental (Group 3b)

	Number of Comments	% of sub group	% of total
Phatic Comment: Comment	0	0.0%	0.0%
Phatic Comment: Encouragement	0	0.0%	0.0%
Total Phatic Comments	0		0.0%
Developmental Comment: Alternative	27	51.9%	8.3%
Developmental Comment: Future	0	0.0%	0.0%
Developmental Comment: Reflective question	23	44.2%	7.1%
Developmental Comment: Informational	2	3.8%	0.6%
Total Developmental Comments	52		16.0%
Structural Comment: Discourse Level	6	55.0%	1.8%
Structural Comment: Sentence Level	5	45.0%	1.5%
Total Structural Comments	11		3.4%
Structural Corrections: Sentence Level	0	0.0%	0.0%
Overall Structural	11		3.4%
Stylistic Comment: Punctuation	0	0.0%	0.0%
Stylistic Comment: Lexis	0	0.0%	0.0%
Stylistic Comment: Syntax/word order/Grammar	8	11.3%	2.5%
Stylistic Comment: Referencing/Citation/Quotation/Bibliography	39	54.9%	12.0%
Stylistic Comment: Presentation	6	8.5%	1.8%
Stylistic Comment: Register	18	25.4%	5.5%
Stylistic Comment: Proof Reading	0	0.0%	0.0%
Total Stylistic Comments	71		21.8%
Stylistic Correction: Punctuation	38	63.3%	11.7%
Stylistic Correction: Lexis	0	0.0%	0.0%
Stylistic Correction: Syntax/word order/Grammar	6	10.0%	1.8%
Stylistic Correction: Referencing/Citation/Quotation/Bibliography	10	16.7%	3.1%
Stylistic Correction: Presentation	1	1.7%	0.3%
Stylistic Correction: Register	1	1.7%	0.3%
Stylistic Correction: Proof Reading	4	6.7%	1.2%
Total Stylistic Corrections	60		18.4%
Stylistic Emphasis: Punctuation	10	31.0%	3.1%
Stylistic Emphasis: Lexis	0	0.0%	0.0%
Stylistic Emphasis: Syntax/word order/Grammar	2	6.0%	0.6%
Stylistic Emphasis: Referencing/Citation/Quotation/Bibliography	10	31.0%	3.1%
Stylistic Emphasis: Presentation	0	0.0%	0.0%
Stylistic Emphasis: Register	7	22.0%	2.1%
Stylistic Emphasis: Proof Reading	3	9.0%	0.9%
Total Stylistic Emphasis Comments	32		9.8%
Overall Stylistics	163		50.0%
Content-related Comment: Positive Evaluation	31	61.0%	9.5%
Content-related Comment: Negative Evaluation	20	39.0%	6.1%
Content-related Comment: Non-Evaluative Summary	0	0.0%	0.0%
Total Content-related Comments	51		15.6%
Content-related Symbol: Positive	49	100.0%	15.0%
Content-related Symbol: Negative	0	0.0%	0.0%
Total Content-related Symbol Comments	49		15.0%
Content-related Criticism: Negative Evaluation	0	0.0%	0.0%

Content-related Emphasis: Negative Evaluation	0	0.0%	0.0%
Overall Content-related Total	100		30.7%
Administrative Comment	0		0.0%
Total Number of feedback Contributions	326		100.0%

9.9 APPENDIX IX: Control Group Analyses for In-Text Feedback

In-Text Feedback Analysis 1: Perceived Student Gender

Analysis 1.1: White British Male vs. White British Female

Stage one of the analysis entailed comparing participants' in-text feedback for the control essay.

Analysis 1.1 Gender: White British Male Vs. White British Female (Control)		
	WBM (Group 1a)	WBF (Group 1b)
Phatic Feedback	0%	1%
Developmental Feedback	14%	15%
Structural Feedback	3%	3%
Stylistic Feedback	69%	57%
Content-related Feedback	13%	24%
Administrative Feedback	0%	0%
Ambiguous Feedback	1%	0%

Analysis 1.1: In-text feedback classifications for WBM versus WBF (Control)

The total number of in-text feedback contributions for the control essay differed across groups; Group 1a (n=328) and Group 1b (n=248). However, there were some patterns in the data. Results demonstrated that feedback related to the stylistic elements of the assignment were the dominant form of feedback for both groups, scoring 69% and 57% percent respectively. Specifically, *stylistic comments* dominated, followed by *stylistic corrections* and *stylistic emphasis* for both groups.

In terms of developmental feedback, the percentage weighting was almost identical (Group 1a = 14%; Group 1b = 15%). Exploration of the subcategories of developmental feedback revealed that *developmental comments* related to *reflective questions* was the most recorded feedback type, followed by *developmental comments* related to *alternative*s and finally *developmental comments* related to *information*. There were no *developmental comments* related to *future* for either group.

Content-related feedback comprised 13% of comments for Group 1a and 24% of comments for Group 1b demonstrating that the markers who comprised Group 1b appeared more prone to providing such feedback. This category was also dominated by *content-related comments* over *content-related symbols* for both groups. Disappointingly 88% percent of the *content-related comments* made by Group 1a fell into the *negative evaluation* category and this figure rose to 91% percent for Group 1b. On a more positive note, comparisons of *content-related symbols* which indicated either positive feedback (such as ticks) or negative feedback (such as crosses) showed that positive symbols accounted for 91% for Group 1a and 100% for Group 1b.

Finally, *structural comments* attracted a score of 3% across groups. *Sentence level* structural *comments* dominated over *discourse level structural comments* for both groups.

Analysis 1.2: Asian Male vs. Asian Female

Stage one of the analysis entailed comparing in-text feedback for the control essay.

Analysis 1.2 Gender: Asian Male Vs. Asian Female (Control)		
	AM (Group 2a)	**AF (Group 2b)**
Phatic Feedback	0%	0%
Developmental Feedback	14%	25%
Structural Feedback	4%	3%
Stylistic Feedback	61%	46%
Content-related Feedback	21%	23%
Administrative Feedback	0.0%	0%
Ambiguous Feedback	0.3%	2%

Analysis 1.2: In-text feedback classifications for AM versus AF (Control)

The total number of in-text feedback contributions differed across groups; Group 2a (n=337) and Group 2b (n=246). Nonetheless, there were still some patterns evident in the data. Once again feedback related to stylistic components of the work attracted the most feedback, scoring 61% for Group 2a and 46% for Group 2b. Specifically, *stylistic comments* dominated over *stylistic corrections* and *stylistic emphasis* for both groups.

Developmental comments yielded different amounts of feedback across groups, with Group 2a awarding 14% of total feedback to this category and Group 2b awarding 25%. However, feedback awarded to subcategories of developmental feedback followed the same patterns with *developmental comments* related to *alternatives* dominating, followed by *developmental comments* related to *reflective questions* and *developmental comments* offering *information*. There were no *developmental comments* related to *future* for either group.

Scores for content-related feedback were almost identical across groups comprising 21% for Group 2a and 23% for Group 2b. This category was dominated by *content-related comments* over *content-related symbols* for both groups. *Content-related comments* pertaining to *negative evaluation* dominated over *positive evaluation*, scoring 75% for Group 2a and 86% for Group 2b. This pattern was reversed when observing positive versus negative *content-related symbols*, where positive symbols accounted for 78% in Group 2a and 95% in Group 2b. These findings echo those in the control essay for Analysis 1.1 (WBM versus WBF).

Structural comments attracted similar percentages across groups (Group 2a = 4% and Group 2b = 3%) and *sentence level* comments dominated over *discourse level* comments for both groups.

Analysis 1.3: Chinese Male vs. Chinese Female

Stage one of the analysis entailed comparing in-text feedback for the control essay.

Analysis 1.3 Gender: Chinese Male Vs Chinese Female (Control)		
	CM (Group 3a)	**CF (Group 3b)**
Phatic Feedback	0%	0%
Developmental Feedback	8%	14%
Structural Feedback	4%	4%
Stylistic Feedback	66%	66%
Content-related Feedback	22%	17%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Analysis 1.3: In-text feedback classifications for CM versus CF (Control)

The total number of in-text feedback contributions differed across groups within the control condition; Group 3a (n=283) and Group 3b (n=374). When these contributions were analysed there were some noteworthy patterns in the data. As with the control condition for Analyses 1.1 and 1.2, stylistic elements of the work attained the most feedback, with both groups scoring 66%. The subcategories of stylistic feedback also replicated the earlier analyses with *stylistic comments* dominating over *stylistic corrections* and *stylistic emphasis* once more.

Content-related feedback attained the next highest number of feedback contributions for both groups, scoring 22% for Group 3a and 17% for Group 3b. Exploration of subcategories demonstrated that *content-related comments* outscored *content-related symbols* for both groups. Disappointingly comments about content were largely negative once more with *negative evaluation* percentages scoring 65% for Group 3a and 82% for Group 3b. The tendency to provide negative feedback once more disappeared when *content-related symbols* were analysed, with Group 3a providing 96% positive feedback and Group 3b providing 100%. These results all mirror the pattern of the earlier control group analyses for feedback related to content.

Developmental feedback was the third most popular category. There was a 6% difference across groups, with Group 3a providing 8% of their feedback in this category and Group 3b providing

14%. The subcategories revealed that *developmental comments* related to *reflective questions* attracted most feedback followed by *developmental comments* related to *alternatives* and *developmental comments* related to *information* for both groups. There were no *developmental comments* related to *future* for either group.

Finally, structural feedback scored 4% and was dominated by *sentence level* comments over *discourse level* comments for both groups. These results replicated the earlier control group analyses.

In-Text Feedback Analysis 2: Perceived Student Ethnicity

Analysis 2.1: Male Ethnicities

Stage one of the analysis involved comparing participants in-text feedback for the control essay.

Analysis 2.1: Ethnicity White British Male Vs. Asian Male Vs. Chinese Male Control			
	WBM (Group 1a)	**AM (Group 2a)**	**CM (Group 3a)**
Phatic Feedback	0%	0%	0%
Developmental Feedback	14%	14%	8%
Structural Feedback	3%	4%	4%
Stylistic Feedback	69%	61%	66%
Content-related Feedback	13%	21%	22%
Administrative Feedback	0%	0%	0%
Ambiguous Feedback	1%	0%	0%

Analysis 2.1: In-text feedback classifications for male ethnicities (Control)

The total number of in-text feedback contributions was different across groups; Group 1a (n=328), Group 2a (n= 337) and Group 3a (n=283). Feedback related to stylistic components of the assignment attracted the highest amount of contributions across groups; Group 1a = 69%, Group 2a = 61% and Group 3a = 66%. Specifically, *stylistic comments* dominated over *stylistic corrections* which in turn dominated over *stylistic emphasis* for all three groups.

Content-related feedback was the second highest scoring category for Group 2a and Group 3a comprising 21% and 22% of the overall feedback respectively. However this category was ranked third highest for Group 1a where it only reflected 13% of the overall feedback. *Content-related comments* prevailed over *content-related symbols* for each group.

Feedback pertaining to the developmental aspects of the work attracted identical percentages for Group 1a and Group 2a at 14%, but only 8% of the total score for Group 3a. Observation of the subcategories for developmental feedback revealed that these were ranked differently for each group. *Developmental comments* related to *reflective questions* was the highest scoring subcategory for Group 1a and 3a and gained 40% and 59% of the overall developmental feedback percentage, but was the second highest subcategory for Group 2a, scoring 46%. *Developmental comments* related to *alternatives* scored 38% for Group 1a, 50% for Group 2a (the highest ranked subcategory) and 36% for Group 3a. *Informational comments* scored 21% for Group 1a, but only 4% and 5% respectively for Group 2a and Group 3a. There were no *developmental comments* related to *future* for either group.

Structural feedback attracted between three and four percent of the overall feedback contributions for all groups and subcategories were unanimously dominated by *sentence level* comments over *discourse level* comments.

Analysis 2.2: Female Ethnicities

Stage one of this analysis involved comparing participants' in-text feedback for the control essay.

Analysis 2.2: Ethnicity White British Female Vs. Asian Female Vs. Chinese Female Control			
	WBF (Group 1b)	**AF (Group 2b)**	**CF (Group 3b)**
Phatic Feedback	1%	0%	0%
Developmental Feedback	15%	25%	14%
Structural Feedback	3%	3%	4%
Stylistic Feedback	57%	46%	66%
Content-related Feedback	25%	23%	17%
Administrative Feedback	0%	0%	0%
Ambiguous Feedback	0%	2%	0%

Analysis 2.2: In-text feedback classifications for female ethnicities (Control)

The total number of feedback contributions across groups was as follows: Group 1b (n= 249), Group 2b (n=246) and Group 3b (n=374). Stylistic feedback was the highest scoring category comprising 57% of overall feedback for Group 1b, 46% for Group 2b, and 66% for Group 3b. Once again *stylistic comments* dominated over *stylistic corrections* and *stylistic emphasis* for all groups.

Content-related feedback gained comparable scores across groups; Group 1b scored 25%, Group 2b scored 23% and Group 3b scored 17%. However, this category was ranked second highest for Groups 1b and 3b and third for Group 2b. *Content-related comments* were more prevalent than *content-related symbols* for each group.

Developmental feedback attracted almost identical scores for Group 1b and Group 3b who scored 15% and 14% respectively, but was much higher for Group 2b who provided 25% of their feedback in this way. Exploration of the subcategories of developmental feedback demonstrated that these were ranked differently for each group. *Developmental comments* related to *reflective questions* was the highest scoring subcategory for Groups 1b and 3b with scores of 57% and 50% respectively. However, this was the second highest subcategory of developmental feedback for Group 2b with a score of 37%. *Developmental comments* related to *alternatives* scored 41% for Group 1b, 48% for Group 2b (the highest ranked subcategory) and 33% for Group 3b. *Informational comments* only scored 3% for Group 1a, but increased significantly for Groups 2b and 3b scoring 15% and 18% respectively. There were no *developmental comments* related to *future* for either group.

Structural feedback was comparable across groups with scores of between 3-4% for each. Once again comments for all groups pertained to structural issues at *sentence level* as opposed to *discourse level*.

In-Text feedback Analysis 3: Perceived Student Gender and Ethnicity

Analysis 3.1: White British male vs. Asian Female

Stage one involved comparing the participants' feedback for the control essay.

Analysis 3.1 Perceived Student Gender & Ethnicity White British Male Vs. Asian Female (Control)		
	WBM (Group 1a)	**AF (Group 2b)**
Phatic Feedback	0%	0%
Developmental Feedback	14%	25%
Structural Feedback	3%	3%
Stylistic Feedback	69%	46%
Content-related Feedback	13%	23%
Administrative Feedback	0%	0%
Ambiguous Feedback	1%	2%

Analysis 3.1: In-text feedback classifications for WBM and AF (Control)

The total number of feedback contributions for the control essay varied across groups; Group 1a (n=328) and Group 2b (n=246). Examination of feedback categories demonstrated that stylistic feedback was the most prominent type of feedback provided for both groups although Group 1a did gain more than 20% more feedback in this domain. *Stylistic comments* prevailed *over stylistic corrections* for both groups although the difference was minimal for Group 2b. *Stylistic emphasis* was the lowest scoring type of stylistic feedback for both groups.

In terms of developmental feedback Group 1a received far fewer comments than Group 2b (14% versus 25%). Subcategories of developmental feedback revealed that comments related to *alternatives* was the highest scoring for Group 2b and equally highest for Group 1a, followed by *developmental comments* related to *reflective questions* and then *informational comments* for both groups. Neither group was provided with any developmental feedback related to their *future* work.

Content-related feedback amounted to 13% of comments for Group 1a and 23% of feedback for Group 2b. The subcategories revealed that *content-related comments* dominated over *content-related symbols* for both groups. Disappointingly negatively oriented comments surrounding content scored 88% for Group 1a and 86% for Group 2b.

Structural feedback made up 3% of the total for each group and was dominated by *sentence level* as opposed to *discourse level* comments.

Analysis 3.2: White British Male vs. Chinese Female

Stage one involved comparing the participants' feedback on the control essay.

Analysis 3.2 Perceived Student Gender & Ethnicity White British Male Vs. Chinese Female (Control)		
	WBM (Group 1a)	**CF (Group 3b)**
Phatic Feedback	0%	0%
Developmental Feedback	14%	14%
Structural Feedback	3%	4%
Stylistic Feedback	69%	66%
Content-related Feedback	13%	17%
Administrative Feedback	0%	0%
Ambiguous Feedback	1%	0%

Analysis 3.2: In-text feedback classifications for WBM and CF (Control)

The total number of in-text feedback contributions varied across groups; Group 1a (n=328) and Group 3b (n=374). Nonetheless, there were still some patterns visible in the data which illustrated where these contributions lay. Stylistic components of the work attracted the most feedback, scoring 69% for Group 1a, and 66% for Group 3b. Specifically, *stylistic comments* dominated over *stylistic corrections* and *stylistic emphasis* for both groups.

Feedback on the developmental aspects of the work attracted identical scores with both groups gaining 14% of their overall feedback in this domain. Patterns concerning the subcategories of developmental feedback demonstrated that *reflective questions* dominated, followed by *alternative* and *informational comments*. No *developmental comments* were made for either group concerning *future* work.

Scores for content-related feedback were 13% for Group 1a and 17% for Group 3b. *Content-related comments* dominated over *content-related symbols* for both groups. Comments related to *negative evaluation* dominated over those related to *positive evaluation* (Group 1a = 88%, Group 3b =82%). This pattern was reversed when observing *positive* versus *negative content-related symbols*. Positive symbols accounted for 91% for Group 1a and 100% for Group 3b.

Structural comments attracted between 3-4% of total feedback contributions for each group with *sentence level* comments outscoring *discourse level* comments for both groups.

Analysis 3.3: Asian Male vs. Chinese Female

Stage one of the analysis consisted of comparing feedback for the control essay.

Analysis 3.3 Perceived Student Gender & Ethnicity Asian Male Vs. Chinese Female (Control)		
	AM (Group 2a)	**CF (Group 3b)**
Phatic Feedback	0%	0%
Developmental Feedback	14%	14%
Structural Feedback	4%	4%
Stylistic Feedback	61%	66%
Content-related Feedback	21%	17%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Analysis 3.3: In-text feedback classifications for AM and CF (Control)

Once again the total number of feedback contributions was different across groups; Group 2a (n=337) and Group 3b (n= 374). Observation of feedback categories revealed that stylistic feedback was provided most frequently with percentages being comparable across groups; Group 2a = 61%, Group 3b = 66%. Examination of the subcategories of stylistic feedback demonstrated that *stylistic comments* outscored *stylistic corrections*, with *stylistic emphasis* being the lowest scoring type of stylistic feedback for both groups.

Content-related feedback was the next highest scoring category for both groups attracting 21% of the total feedback for Group 2a and 17% for Group 3b. The subcategories revealed that *content-related comments* dominated over *content-related symbols* for both groups, although there was a more even split between *comments* and *symbols* for Group 3b. It was disheartening to see that negatively oriented comments surrounding content scored 75% for Group 2a and 82% for Group 3b.

Observation of the developmental feedback category showed that both groups attracted the same percentage scores of 14%. However this percentage was distributed differently when the subcategories were scrutinised. Specifically, while comments related to *alternatives* scored highest for Group 2a this was only the second highest score for Group 3b. *Developmental comments* related to *reflective questions* was the next highest scoring subcategory and attracted comparable scores across groups. *Informational comments* attracted the next highest score although there was a 14% difference between groups (Group 2a = 4%, Group 3b = 18%). Neither group was provided with any *developmental feedback* related to *future* work.

Structural feedback made up 4% of the total for each group and was dominated by *sentence level* as opposed to *discourse level* comments.

Analysis 3.4: Asian Male vs. White British Female

Stage one of the analysis involved comparing participants in-text feedback for the control essay.

Analysis 3.4 Perceived Student Gender & Ethnicity Asian Male Vs. White British Female (Control)		
	AM (Group 2a)	**WBF (Group 1b)**
Phatic Feedback	0%	1%
Developmental Feedback	14%	15%
Structural Feedback	4%	3%
Stylistic Feedback	61%	57%
Content-related Feedback	21%	24%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Analysis 3.4: In-text feedback classifications for AM and WBF (Control)

The total number of in-text feedback contributions was different across groups; Group 2a (n=337), Group 1b (n= 249). The most prolific type of feedback for both groups was stylistic feedback attracting 61% of overall feedback contributions for Group 2a and 57% for Group 1b. In terms of subcategories *stylistic comments* dominated over *stylistic corrections* and *stylistic emphasis* for both groups.

Content-related feedback was the second highest scoring category for both Group 2a and Group 1b comprising 21% and 25% of the overall feedback respectively. Once again *content-related comments* dominated over *content-related symbols* for each group. Comments were dominated by *negative evaluation* for both groups (Group 2a = 75%, Group 1b = 91%) although this trend was reversed for *symbol-based* feedback which was predominantly positive in nature (Group 2a = 78%, Group 1b = 100%).

Similar percentages were also evident when observing the developmental feedback category with Group 2a scoring 14% and Group 1b scoring 15%. However differences at subcategory level were observed. Whereas *developmental comments* related to *alternatives* dominated for Group 2a, this was ranked in second place for Group 1b who provided most feedback for *reflective questions* instead. Both groups were provided with limited *informational comments* and no comments related to the *future* development of their work.

Structural feedback amounted to 4% of overall contributions for group 2a and 3% for Group 1b. *Sentence level* comments once more prevailed over *discourse level* comments.

Analysis 3.5: Chinese Male vs. White British Female

Stage one of the analysis involved comparing participants in-text feedback for the control essay.

Analysis 3.5 Perceived Student Gender & Ethnicity Chinese Male Vs. White British Female (Control)		
	CM (Group 3a)	**WBF (Group 1b)**
Phatic Feedback	0%	1%
Developmental Feedback	8%	15%
Structural Feedback	4%	3%
Stylistic Feedback	66%	57%
Content-related Feedback	22%	24%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	0%

Analysis 3.5: In-text feedback classifications for CM and WBF (Control)

The total number of in-text feedback contributions was different across groups; Group 3a (n=283), Group 1b (n= 249). The most common type of feedback provided for both groups was stylistic feedback which attracted 66% of total feedback contributions for Group 3a and 57% for Group 1b. Subcategories for Stylistic feedback were ranked in the same order for both groups too; *stylistic comments, stylistic corrections* and *stylistic emphasis*.

Content-related feedback attracted the next highest scores for both groups comprising 22% of total feedback for Group 3a and 25% for Group 1b. *Content-related comments* once more dominated over *content-related symbols* for both groups. The predisposition to provide comments conveying a *negative evaluation* of the assignment dominated for both groups although there were substantial differences (Group 3a = 65%, Group 1b = 91%). However this tendency was once more reversed when the feedback was symbol-based with positive symbols dominating (Group 3a = 96%, Group 1b = 100%).

Developmental feedback attracted 8% of total feedback contributions for Group 3a and 15% for Group 1b. Given this difference it was interesting to observe that percentages across subcategory level remained fairly constant in terms of developmental comments related to *alternatives, reflective questions* and *informational comments*. There were no comments related to the *future* work for either group.

Structural feedback was comparable across groups, gaining a score of 4% for Group 3a and 3% for Group 1b. *Sentence level* comments outweighed *discourse level* comments across the board.

Analysis 3.6: Chinese Male vs. Asian Female

Stage one of the analysis involved comparing the in-text feedback for the control essay.

Analysis 3.6 Perceived Student Gender & Ethnicity Chinese Male Vs. Asian Female (Control)		
	CM (Group 3a)	AF (Group 2b)
Phatic Feedback	0%	0%
Developmental Feedback	8%	25%
Structural Feedback	4%	3%
Stylistic Feedback	66%	46%
Content-related Feedback	22%	23%
Administrative Feedback	0%	0%
Ambiguous Feedback	0%	2%

Analysis 3.6: In-text feedback classifications for CM and AF (Control)

The total number of in-text feedback contributions was different across groups; Group 3a (n=283), Group 2b (n= 246). The most prolific type of feedback for both groups was stylistic feedback attracting 66% of overall feedback contributions for Group 3a and 46% for Group 2b. In terms of subcategories *stylistic comments* dominated over *stylistic corrections* and *stylistic emphasis* for both groups although *stylistic comments* and *stylistic corrections* attracted near identical scores for Group 2b (AF).

Content-related feedback attracted similar scores across groups comprising 22% of total feedback contributions for Group 3a and 23% for Group 2b. However, although this type of feedback was ranked in second place for Group 3a it was only ranked in third place for Group 2b whose score for Developmental feedback marginally surpassed that of Content. As has been common in previous analyses, *content-related comments* dominated over *content-related symbols* for both groups. Comments related to *negative evaluation* were most common (Group 3a = 65%, Group 2b = 86%) although this trend was reversed for *content-related symbol-based* feedback which was predominantly positive in nature (Group 3a = 96%, Group 2b = 95%).

There was a large discrepancy in feedback pertaining to Developmental issues. Specifically Group 3a gained 8% of their feedback in this area whereas Group 2b gained 25% representing huge differences in marking practice in operation across groups. There were also differences evident at subcategory level with Group 3a gaining the majority of their developmental feedback in the form of *reflective questions* (59%) and Group 2b gaining the majority of theirs in *alternatives* (48%). *Informational comment* feedback was ranked in third place for both groups and neither group received any comments related to the *future* development of their work.

Structural feedback amounted to 4% of overall contributions for Group 3a and 3% for Group 2b. *Sentence level* comments once more prevailed over *discourse level* comments.