OXFORD UNIVERSITY CENTRE FOR EDUCATIONAL ASSESSMENT

Marker effects and examination reliability

A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling

Jo-Anne Baird, Malcolm Hayes, Rod Johnson, Sandra Johnson & Iasonas Lamprianou

January 2013

Ofqual/13/5261

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Content

1 Executive Summary

1.1 Research context, aims and objectives

This collaborative research project was a comparative study of the contributions that three different analysis methodologies could make to the exploration of rater effects on examination reliability. The analysis techniques in question were Generalizability Theory (G-theory), Item Response Theory (IRT) – in particular the Many-Facets Partial Credit Rasch Model (MFRM), and Multilevel Modelling (MLM). The examination datasets supplied for use in the project were AS component papers in two large-entry subjects – geography and psychology – for the three consecutive years 2009, 2010 and 2011.

Rater effects can be grouped into two broad types: inter-rater effects and intra-rater effects (also known as interaction effects). Arguably the best known, and certainly the most researched, inter-rater effect is that of between-rater difference in overall rating standards, i.e. differences in severity/leniency. Intra-rater effects, which have been less well covered in the literature, take many different forms, but are all examples of departures from the general pattern of inter-rater difference in severity/leniency. Rater-candidate interaction effects, rater-question interaction effects, and rater-occasion interaction effects ('marker-drift' if the pattern of differences is clearly a trend over time) are all evidence of intra-rater effects.

This project set out to explore both inter-rater effects and intra-rater effects. Given the findings from previous research, we initially set out to investigate

- the impact of inter-rater reliability
- the stability of rater severity effects
- question effects
- the effect of training group on marker behaviour
- the effect of training approach (face-to-face versus on-line standardisation).

In the event, in only one year (2009) for one subject (psychology) was marker standardisation conducted face-to-face, so that the last research objective in the list was intractable with the available data.

1.2 The datasets

Following current practice in all the major English awarding organisations at this time, the completed papers (the 'scripts') produced by the examination candidates were scanned and split into 'clips' for online marking. A clip is the response of an individual candidate to a question or part-question within the paper. In the operational process each clip was marked online by a single marker. As marking progressed, two types of

periodic quality check took place. In one type, 'backreading', marked clips were diverted by the automated delivery system from time to time to be re-marked by the marker's team leader (now acting as marking supervisor). The supervisor re-marked the clip, but with knowledge of the original marker's mark. Where there was a difference of opinion the supervisor's mark overrode the original mark as a contribution to the candidate's final composite score for the paper. In the other type of quality check a set of clips that had been pre-marked by the chief examiner at the time of marker standardisation were 'seeded' to markers in their clip allocations. The markers marked these clips without knowledge of their status or of their pre-assigned marks.

As a result, each annual subject dataset comprised three types of record, or three data subsets, reflecting the three different processes: the operational marking, the backread quality check marking, and the seeded clip quality check marking. The majority of records were 'operational', which means that they were the result of one marker marking one clip.

1.3 The three analysis techniques and their findings

G-theory, Rasch modelling and multilevel modelling each in principle have something to offer rater effects research. The first, G-theory, is essentially a sampling approach to reliability investigation, generalising sample-based findings to a given population or universe. It focuses on variance analysis, and makes no distributional assumptions. G-theory analysis is most effective when a given dataset allows the potential to explore not only main effects (e.g. in this context of markers, questions, teams) but also, critically, those interaction effects that contribute to measurement unreliability (marker-question interaction, marker-occasion interaction, etc.).

For Item Response Theory (IRT) the model of choice for data analysis was the Many-Facets Partial Credit Rasch Model (MFRM). Like all IRT modelling techniques, MFRM is in essence a scaling technique. The power of the technique is that, provided some quite strong assumptions are met, in particular that there are minimal interaction effects evident in the data, markers, questions and candidates can be located in terms of relative severity, difficulty and ability, respectively, on the same underlying logit scale.

Multilevel modelling is an explanatory approach, which, like G-theory, quantifies the relative contributions to variance of the factors identified in the analysis model. An essential assumption underpinning a multilevel analysis is that all variables are normally distributed. Multilevel models can be conceived of as an extension to regression models, in which the residuals are structured and modelled. Specifically, the hierarchical nesting of the residuals is taken into account, such as candidates' marks

being nested within examiners. Multilevel modelling has expanded our capacity to model complex data structures.

The unit of analysis in every case was a candidate mark on a question or part-question, i.e. a clip score. In each subject the three annual datasets were separately analysed using G-theory and MFRM to investigate rater effects and measurement reliability. Both techniques were used on the single-marked operational datasets; follow-on MFRM repeat analyses combined seeded clip data with single-marked operational data, whereas further G-theory analyses looked at seeded clip data alone. Multilevel modelling analyses, on the other hand, combined all backread and seeded clip data records over all three years across both subjects, and excluded the single-marked operational data.

It was not possible to carry out a 'team effect' analysis using MFRM, because of disjoined subsets (since individual markers were naturally assigned to one team only, team membership was mutually exclusive). G-study analyses were able to explore team effects within years, and found no evidence of these. Team effects were analysed in the multilevel model for geography and were found to be very small. Caveats have to be placed upon the multilevel findings regarding teams, however, as there were fewer teams (15) than necessary to safely conduct these analyses.

Stability of rater effects was investigated in different ways using the three methodologies. In MFRM, the rater effects were found to have non-significant correlations between years for the psychology data. Generalizability analyses investigated marker drift within each examination series and found no evidence that marker behaviour changed over the period of operational marking. The multilevel models were set up to investigate stability of rater effects across marker monitoring systems and found that rater effects correlated moderately across the backread and seeded systems. Across years, moderate correlations of rater effects were found in geography, but there were non-significant correlations in psychology (in keeping with the MFRM finding).

As far as marker contributions to clip score variance is concerned, both the G-theory analyses and the multilevel analyses found inter-rater effects to be modest in comparison with those from questions and other effects, in particular from between-candidate variance and from interaction effects involving candidates, which were confounded in the residual term (G-theory). The MFRM analyses, on the other hand, revealed significant inter-rater effects and a low reliability of measurement for the candidate scores. These findings were consistent across the three year groups, although

the effect was smaller for the 2011 psychology dataset. However, a particular feature to note that might create uncertainty about the validity of these findings is that there was an unusually high proportion of candidate misfit, both in psychology and in geography, meaning that the Rasch model did not fit the data well. This is probably explained by the presence of the kinds of interaction effects noted in the G-theory analyses, given that in the MFRM these are assumed in principle to be absent (the Rasch model can accommodate interaction effects to some extent, but only up to a point).

1.4 Data issues

The single-marked operational data were useful for investigating rater effects of different kinds. But they could never provide an estimate of the reliability with which examination candidates had actually been measured by the component paper investigated. The backread data could not do this either, given that the (only) two markers who marked each clip were not marking independently (the backreading supervisor was aware of the mark already assigned by the regular marker). The only data with any possibility of offering this option would have been the seeded clip data, given that most or all of the regular markers had independently marked each seeded clip without knowledge of the pre-assigned 'expert' mark.

The seeded clip exercise, though, had not been designed to serve this purpose on these occasions. The samples of seed clips were very small and uneven in size. Moreover, they had not been randomly selected to represent either the total entry for the subject concerned in the year in question, nor were they randomly selected to represent some identifiable subset – such as clips within five marks of a grade boundary. This meant that any analysis based on seeded clip data would probably not be well-based (estimation errors would be high) and generalisation would be limited in value. The fact that from one part-question to another the seed clips samples were from different candidates reduced even further the possibility of estimating measurement reliability at component level. This is an area that warrants consideration for the future, since the seed clip strategy could without too much additional cost and effort be re-designed and expanded to serve both the current operational quality check needs and the need to provide component-level reliability statistics. The project team has experience of analysis of these kinds of data from a range of examinations and electronic systems and the issues we raise here are general.

1.5 Conclusions

In keeping with previous findings within Ofqual's reliability programme, we did not generally find that there was a large impact of raters upon clip scores when compared with the effect of items. However, when estimated up to whole-paper level effects might have been more noticeable: the MFRM analyses, which located markers on a scale in

terms of their estimated overall severity/leniency, found large marker differences (though model fit was poor). Neither did the G-theory analyses reveal any evidence of training team effects. Measures of marker performance were found to be stable within an examination series in the generalizability analyses, but the MFRM and the multilevel analyses showed these to be moderately stable at best between years and between the two monitoring systems (backread and seeding). However, since rater effects were small, we would not necessarily expect high correlations.

We aimed to tackle a number of research questions using three advanced statistical methods, to compare them and to do all of this using operational datasets. In the timescale of this study, the project was ambitious and difficult. It is clear that much more could be done to compare the data using these techniques, as the methods offer the capacity to analyse the data in a range of other ways. As such, some of our findings are not relative limitations of the methods per se, but artefacts of the ways in which we have currently analysed the data for the purposes of this report. In a short project it is not possible to explore all possibilities, but this report nevertheless represents an important advance upon the previous literature and a basis upon which future research can build.

1.6 Acknowledgements

This research was commissioned by Ofqual. We would also like to acknowledge Pearson for providing data for analysis. Rose Clesham and Jeremy Pritchard of Edexcel also assisted the research by clarifying our interpretations of the data. Jo Hazell administered the project and Yasmine El Masri assisted with the finalisation of the report.

Our methodological critical friends were invaluable. They were George Leckie, George Marcoulides and Jim Tognolini. Any remaining errors are those of the authors.

2 Introduction

2.1 Project overview

In the eyes of the public, rater reliability is paramount for English public examinations (Ipsos MORI, 2009, p.21). Public examinations need to offer a common yardstick, so inter-rater reliability assumes a great deal of significance in systems where external examinations are used for high stakes assessments. Kingsbury (1922) categorised the effects that examiners could have upon scores in terms of

- severity,
- halo and
- central tendency effects.

Severity effects arise when examiners differ markedly in their leniency/severity, halo effects involve a positive biasing effect of part of a candidate's performance upon the assessment of other parts, and central tendency effects relate to the extent to which examiners use the entire range of the mark scale. These are all systematic effects that can readily be observed in the resulting marking data. In addition to these are unsystematic effects that can be grouped under the term 'marker inconsistency'. Examples are particular markers scoring particular scripts more or less severely than normal, and markers scoring erratically over a period of time as opposed to showing consistent trends towards greater or lesser severity. These various aspects of marker behaviour are widely referred to as 'rater effects' in the literature and we use this term here.

Information about rater reliability for a given set of scores produced under particular administration conditions for specific assessments is useful for quality assurance and quality control. To improve the quality of rating, assessment organisations can address systemic aspects of their processes, such as marker training. Additionally, the performance of individual examiners is monitored and controlled. With the introduction of on-screen marking, the capacity for measuring and assuring inter-rater reliability has improved. Availability of data and new research avenues have also followed from the application of new technology to marker reliability issues. Operationally, assessment organisations put a lot of resource into rater quality assurance based upon sample checks of marking performance. These checks typically measure rater severity, but can in principle identify other effects such as central tendency and rater interactions.

This project set out principally to investigate rater severity effects and their stability across examination administrations. But we looked also at the issue of marker

inconsistency in its different forms. We analysed General Certificate of Education (GCE) AS-level examination data in two subjects, geography and psychology, using generalizability theory (G-theory), Rasch modelling and multilevel modelling (MLM).

2.2 Previous research on raters and their effects

There is a sizeable literature on rater severity effects. Within the English examinations system in particular, research into marker behaviour and performance has a long history, and has been extensive in scale and in scope. Meadows and Billington (2005) give a relatively recent and comprehensive review. Several different marking-related issues that might impact on assessment reliability have been investigated in recent studies. These include: comparisons of the results of paper-based and online marking (e.g. Johnson, Nadas and Bell, 2010); investigations into the possibility of training team effects and of training strategies (e.g. Shaw, 2002; Baird, Greatorex and Bell 2004; Greatorex and Bell, 2008); the effectiveness of different types of marker standardisation procedures on inter-marker reliability (e.g. Baird et al., 2004); studies into markers' thought processes as they mark (e.g. Crisp, 2010); the effect of different models of double marking (Vidal-Rodeiro, 2007) and the impact of marker characteristics (Royal-Dawson and Baird, 2009; Suto, Nadas and Bell, 2011).

Elsewhere, research has also been conducted on the relationship between rater characteristics and rater effects (e.g. Lunz and Stahl, 1990; Powers and Kubota, 1998; Shohamy, Gordon and Kraemer, 1992) and recently, including in the UK, there has been an active interest in whether rater effects are stable over time (Congdon and McQueen, 2000; Harik, Clauser, Grabovsky, Nungester, Swanson and Nandakumar, 2009; Hoskens and Wilson, 2001; Lamprianou, 2006; Myford and Wolfe, 2009; Leckie and Baird, 2011).

2.2.1 Rater severity

Several studies have investigated changes in rater severity effects over the examining period, with mixed findings. Lunz and Stahl (1990) conducted research on three assessments – essay, clinical and oral examinations. They did not find significant effects of time on severity of scoring for the oral examination. For the clinical and essay examinations, severe ratings were found on the second and third half day sessions (respectively), but thereafter ratings became more lenient on average (significantly so for the clinical examination). Pinot de Moira, Massey, Baird and Morrissy (2002) analysed students' English extended writing test scores, assigned by 78 raters, to investigate rater drift in severity over a scoring period which lasted several weeks. They found that raters were, on average, significantly more severe towards the end of their scoring period than at the beginning, although this was a small effect. Congdon and McQueen's (2000) study on the writing performances of 8,285 elementary school students, with 10 raters, was conducted over a one week marking period. Nine of the

raters became more severe over time and one became more lenient, which showed that different raters can exhibit different trends in their scoring severity over time. Indeed, Myford and Wolfe's (2009) study of 101 raters and 28 check essays on an Advanced Placement English Literature and Composition examination found significant positive and negative drift in rater severity over time for a small proportion of their raters. They also found that where rater effects were not stable, it tended to be for a single check, rather than a trend.

Interestingly, Lamprianou (2006) investigated rater effects across academic subjects as well as over time. He found that there was considerable instability of rater severity effects. All of these studies showed significant within-rater variability of rater effects over time, which relates to the debate regarding whether rater effects are stable traits or are unstable states, and indicates the presence of marker interactions of different kinds. Finally, in a large study of rater effects for national curriculum tests, Leckie and Baird (2011) found that although there was no directional drift in severity over time, there were large within-individual variations, again suggesting the presence of interaction effects. These recent findings in the literature raise questions about how assessment agencies should measure and interpret inter-rater reliability in operational procedures, so we were interested to pursue issues regarding stability of our measures of rater severity. After all, this is the main gauge of examiner performance.

2.2.2 Training group effects

Wilson and Case (2000) found significant supervisor effects in a high school mathematics examination, but did not find an association with the subsequent behaviour of raters within teams. Hoskens and Wilson (2001) also found supervisor effects in their analysis of rating quality checks for a high school economics examination, with one team of raters being significantly more severe than the others. In a recent analysis, Baird, Leckie and Meadows (*in submission*) showed that training group affected the measures of severity of individual raters, even when the 'true' scores for check scripts were set centrally (rather than by the Team Leader who conducted the training). This means that marking cultures were brought about through interactions within standardisation training groups. Due to these findings in the literature, we wanted to investigate whether there were training effects in our current data. Unlike the studies in the literature, a system of on-line training was used (with the exception of one of the psychology papers), so there was less scope for training group effects in the data as all examiners experienced the same training and there was no face-to-face interaction between supervisors and examiners. Feedback was given to examiners on their marking if they went awry, so this was an opportunity to test whether the feedback alone could cause training group effects.

2.2.3 Monitoring system effects

Two ways of monitoring raters are typically used in electronic monitoring systems, as 'live' marking is checked and standard scripts are also second-marked by supervisors (these are explained more fully in Chapter 3). Use of standard scripts across raters is sometimes called 'seeding' and enables a consistent check to be conducted. Baird, Leckie and Meadows (*in submission*) found significant differences in the rater effects produced under these two types of monitoring system when analysing monitoring data from A-level examinations. When supervisors could see the original marking of examiners, there was significantly less discrepancy between the check mark (the supervisor's mark) and the original mark. Previous research had shown the biasing effect of the original mark upon second-marking (McVey, 1975; Murphy, 1979).

Another finding from the Baird, Leckie and Meadows (*in submission*) study was that supervisors created bigger differences between examiners, in terms of marking accuracy, when the supervisors second-marked live scripts than when standard, seeded scripts were used. From a single study, we cannot tell what features of the monitoring system caused these effects. Helpfully, there are a number of differences between the systems used in the Baird, Leckie and Meadows study and in the current study, which might help to shed light on the causes. As can be seen from the comparison of the two studies across Table 1 and Table 2, in the current study raters were not able to select their own samples for checking, there were no differences in the presentation of marking from either system, training was conducted on-line and examiners did not mark an entire student's script. In the current study, examiners did not see whole candidates' scripts, but instead were presented on-screen with scanned parts of their written booklets, known as 'clips'.

Table 1 Monitoring systems used in the Baird, Leckie and Meadows study

	Pre-set true score system	**Evaluation true score system**
Examiner's original mark	Observable	Observable
Correct score assignment	Principal Examiner	Supervisor
Selection of check sample	Principal Examiner	Rater
Check sample	Same 5 for all raters for a test	Different 5 for each rater
Presentation of check sample	Photocopied student work from current test	Original student work from current test
Training	Face-to-face meetings on tables with the supervisor	
Marking	Entire script	

Table 2 Monitoring systems used in the current study

	Seeded clip system	**Backread system**
Examiner's original mark	Observable	Observable
Correct score assignment	Principal Examiner	Supervisor
Selection of check sample	Principal Examiner	E-marking system
Check sample	Selection from a pre-chosen standard set of clips	Individual students' work (clips)
Presentation of check sample	Scanned work presented on-screen	
Training	On-screen system, with the exception of face-to-face training on tables with supervisor for the 2009 psychology examination	
Marking	Item level	

2.3 Marker-focused research within Ofqual's reliability programme

Little of the research into marker reliability has produced, or been aimed to produce, empirical quantifications of the effects of marker-related factors on the reliability of candidate measurement, even when multiple-marking has been a feature:

> … it is fair to say that much research involving multiple marking has not had the calculation of a reliability coefficient for a particular examination as its primary purpose. (Bramley and Dhawan, 2012, p.269)

It was the absence of empirical evidence about the level of reliability actually achieved for English examinations that prompted Ofqual to launch its three-year reliability programme (Opposs and He, 2012). Two of the commissioned research projects within the programme empirically addressed the issue of the potential contribution to component reliability of marker-related effects (Bramley and Dhawan, 2012; Johnson and Johnson, 2012a).

One of the project reports (Bramley and Dhawan, 2012) includes a particularly extensive, thoughtful and informative section on the general marking issue. An overview of current procedures for quality assuring marking is presented: marker standardisation, seeded script exercises, random checks of the marking of individual markers by team leaders, mark adjustment practices, and so on. Also offered is an equally informative description of how marking is now organised for many components in the new electronic age of randomised allocation of scripts to markers for online marking. This strategy is rapidly replacing the traditional method of organisation, in which batches of paper-based scripts were despatched by courier from centres to markers for marking, and then from markers to the awarding body with marks and comments added. Several useful chart formats for visually monitoring marker behaviour during the operation marking process are illustrated in the report with real-data examples. The report falls short, though, of providing any quantifications of the reliability actually achieved at component-level or qualification-level for any subject, whether through paper-based script marking or electronic marking. On the other hand, the authors do offer an interesting analysis of the relative contributions to mark variation that could be attributed to markers, to candidates, and to marker-candidate interaction (typically confounded in residual variance), using seeded script data.

A second report (Johnson and Johnson, 2012a) went a step further, by using similar variance contribution information to provide quantifications of likely reliability achieved for a traditional paper-based GCSE History component paper. The project carried out a generalizability study whose design had been fully described in an earlier report in the reliability programme (Johnson and Johnson 2012b). The intention was to illustrate with examination data how information about the relative contributions of candidates, markers, questions and their interactions can be used to produce reliability indicators, not only for candidate measurement but also for marker measurement.

Sadly, due to data access problems, only one such study could be reported. The dataset analysed comprised the marks awarded to each of three history essay questions from each of 40 markers to each of 30 candidates. While the average inter-marker correlation was very high (a traditional, but misleading, indicator of marker reliability), there were

clear differences in overall severity of marking amongst examiners, and sometimes large differences in the marks awarded to particular candidates by different markers. In other words, there was evidence of both inter-marker and intra-marker variation at play in the data, as Bramley and Dhawan (2012) also found in their analyses. Of the total mark variation at candidate-marker-question level in the dataset, just over 60% could be attributed to candidates, over 15% to the candidate-question interaction, just 3% to questions, 5% to markers and 2% to each interaction involving markers, *viz*. marker-candidate interaction and marker-question interaction.

The predicted generalizability coefficient for the case of single marking (the normal operational situation in the English system) was 0.86, while the 95% confidence interval around a total test score was around ± 9 marks, giving a band spanning 24% of the mark scale. Doubling the number of questions in the test to six, whilst retaining single marking, would be predicted to increase the reliability coefficient to 0.91, and to reduce the 95% confidence interval around at test score to roughly ± 14 marks, giving a band spanning just over 18% of the mark scale.

2.4 Research questions

Given the findings from previous research, we set out to investigate

- the impact of inter-rater reliability
- stability of rater severity effects
- question effects
- the effect of training group on marker behaviour
- the effect of training approach (face to face versus on-line standardization).

Our data were from two AS examination question papers (geography and psychology), across three years. The intention was to look at examiner performance across occasion, i.e. across years. This is a form of intra-rater reliability. The extent to which we could address the research questions was dependent upon the form of the available data and the requirements of the statistical techniques. For example, the fact that examiners were nested within teams without multiple membership made the data less than ideal for Rasch modelling; having fewer than 20 training teams made the data problematical for multilevel modelling; not having a complete, balanced design for analysis (all markers scoring all scripts) was less than ideal for generalizability theory. If the findings seem complex, it is because the data are complex and this is one of the challenges of conducting research in this area. Each monitoring system is different and there are idiosyncrasies to each examination, even in each year. In the chapters that follow, the same data are treated very differently in an attempt to use each method to address the same research questions. In the event, none of the researchers tackled a

comparison of on-line with face-to-face training because only one examination (psychology 2009) was conducted through face-to-face training, so that any comparison of training format would be confounded with any other features particular to that examination.

In Chapters 4, 5 and 6 we overview each of the three techniques, and describe the analyses carried out using them. We offer a comparative summary of findings in Chapter 7, and draw out implications for future practice, both for the awarding body and for reliability researchers. But before moving on to the analyses, findings and implications, Chapter 3 provides a comprehensive account of the datasets that were made available to us, and the procedures that were in place in the awarding organisation to quality assure marking.

3 The datasets

3.1 The component papers

Two Advanced Subsidiary GCE subject examination components were explored in this project: a geography unit and a psychology unit. In each case marked response data were supplied for the component papers used in each of three years: 2009, 2010 and 2011. The structures of the three papers within each subject were the same or similar over the years, but the questions within them naturally differed.

The geography paper was a two-section paper, with each section comprising two three-part open-ended questions. Candidates were required to answer one question from each section, so that in practice there were four different pathways through the one paper. Within every question two of the part-questions carried 10 marks each while the third carried 15 marks, for a paper total of 75 marks. The part-questions generally extended to no more than one or two lines of text and the space allocated for answers extended to around 40-50 lines, which indicated the examiner's expectation for responses. The time allowance was 60 minutes for the 2009 paper, increased to 75 minutes from 2010.

The 100-minute psychology paper was in three sections, with no question choice. The number of questions in each section, and the mark tariffs they carried, changed from one year to another. Section A comprised approximately ten multiple-choice questions in each year; while the majority were binary-scored, one or two were worth two marks. Section B comprised between five and seven questions, with a mixture of formats. Section C typically contained one or two extended-response questions. The maximum achievable mark for the paper was 80; Section A carried the lowest number of marks (12, 13 and 11, respectively, in 2009, 2010 and 2011); Section B carried over half the total paper marks (42, 49 and 44, respectively each year); Section C carried between a quarter and a third of the total marks (26, 18 and 25 marks, respectively).

Both subjects were large-entry. The total number of candidate entries across the three years for geography was 32,926. Of these, 1, 774 were re-sits leaving 31,152 unique candidates over the period. For psychology the total number of candidate entries across the three years was 20,276; some 2,259 candidates took re-sits in the period, leaving 18,017 unique candidates.

The following sections give a brief overview of the processes that were in place both to mark the scripts and to quality assure the marking of these and other component papers.

3.2 Marker standardisation

The *GCSE, GCE, Principal Learning and Project Code of Practice*, published jointly by Ofqual, Llywodraeth Cymru and CCEA, the examinations and qualifications regulators in England, Wales and Northern Ireland, respectively, outlines the quality assurance procedures which are followed by all awarding organisations for GCSE and GCE qualifications. A summary of principal features of the procedures follows.

3.2.1 Division of responsibilities

For each subject a chief examiner is responsible for the assessment specification as a whole and a principal examiner is responsible for the professional judgements underpinning the process of standardising marking for the component paper. The number of markers, or examiners, is intentionally kept to a minimum to reduce the scope for variability in marking standards. In both subjects studied in this project the entry was such that the principal examiner could not complete all of the marking and therefore team leaders and additional examiners were appointed: examiners were formed into marking teams, each team led and monitored by a particular team leader throughout the operational marking process. The following sections give a brief overview of the marking quality assurance processes that were in place. In practice, much more detailed process maps are in operation.

3.2.2 Marker training

The principal examiner, any assistant principal examiners and all team leaders attended a pre-standardisation meeting that included:

- Administrative briefings on procedures, timelines, documents and contracts
- A principal examiner briefing on the nature and significance of standardisation and any issues from current and previous examinations
- Discussion of the mark scheme
- Marking and discussion of sample responses
- The handling of unexpected, yet acceptable, responses
- Confirmation of the marks pre-awarded by principal examiners to a sample of candidates' question responses for use in marker performance monitoring throughout the operational marking period and any annotations for the sample responses; the pre-allocated marks are known within UK awarding bodies as 'true scores' – not to be confused with the 'true scores' of Classical True Score theory
- Confirmation of the final mark scheme

At the end of the pre-standardisation meetings the mark schemes were finalised and made available to all markers.

All examiners then attended a standardisation meeting that included:

- An administrative briefing on procedures, timelines, documents and contacts
- A principal examiner briefing on the nature and significance of standardisation and any issues from current and previous examinations
- Discussion of issues that emerged during familiarisation marking
- Discussion of the mark scheme, including any criteria for the assessment of quality of written communication (this is assessed in every subject paper)
- Marking and discussion of sample responses chosen to illustrate the range of performance and possible types of response
- The handling of unexpected, yet acceptable, responses.

For each subject paper, the principal examiners were responsible for establishing and setting the standard for marking using their professional judgement about how to interpret and apply the mark scheme. The principal examiner's judgement on these issues is always final.

In 2009 the psychology paper was standardised in the traditional way, in a face-to-face meeting involving markers and senior examiners. The 2010 and 2011 psychology papers were standardised online, as were the geography papers in all three years. The main difference between the two standardisation formats is that whereas in a face-to-face meeting samples of candidates' work are marked during the meeting itself, in online standardisation markers work through their work samples online at home and at their own pace. Feedback prepared by the principal examiner was provided after the initial standardisation, and further support was given where required by team leaders.

Within 48 hours of standardisation, examiners were required to complete practice sets of candidate work using the final version of the mark scheme. They were free to contact their team leader to discuss any issues or queries raised during this marking. Marker performance was checked by team leaders at this point, and markers who demonstrated the required conformity in standards were allocated their workload for operational marking.

3.3 The operational marking process

In all three years operational marking was carried out online. Candidate scripts, or those sections that were to be human marked, were first divided into electronic 'clips', a clip consisting of a single candidate's response to a question or part-question. Clips were randomly allocated to markers, each of whom typically saw and marked a succession of clips relating to the same question or part-question, rather than the traditional case of seeing all responses from single candidates (i.e. entire scripts).

For geography the markers were drawn from a pool of 52, with 36, 34 and 37 being used in each of the years 2009, 2010 and 2011, respectively. Each examiner marked approximately 2,100 clips in the year while team leaders marked around 300 clips in addition to the work they did monitoring their team of markers. The principal examiner marked a small number of clips first hand to gain direct evidence of the candidates' interpretation of questions and the application of the mark scheme. All examiners marked clips for every question or part-question, approximately in proportion to the number of candidates attempting them.

For the psychology paper, Section A was computer-marked, while Sections B and C were marked by the subject experts or by general markers, depending on the level of need for subject expertise from part-question to part-question. Clips were categorised accordingly. Examiners marked across all clips in their category. The markers were drawn from a pool of 49 with 21, 26 and 29 marking, respectively, in each of the three years. Each examiner marked approximately 4,400 clips in a year, while team leaders marked around 2,800 clips, as usual in addition to monitoring their team of markers. The principal examiner marked a small number of clips first hand to gain direct evidence of the candidates' interpretation of questions and the application of the mark scheme.

Two strategies were adopted to monitor the performance of individual markers throughout the operational marking period: re-marking of sample clips by team leaders, known as 'backreading', and clip seeding, which is the distribution of pre-marked clips to markers for blind marking, the clips having been pre-marked by the principal examiner.

Seeded clips were presented to examiners at random at a pre-determined rate during their marking. Markers did not know when they were marking a seeded clip, and they did not see the mark already assigned by the principal examiner. Differences between the two marks were recorded, with 'significant differences' prompting intervention by the team leader (a 'significant difference' is when the regular marker's mark falls outside some pre-agreed threshold when compared with the expert's mark).

For both subjects, seed clips were constructed for every part-question that required human marking. The number of clips created for this purpose varied across the years, but was in every case extremely small (e.g. fewer than 0.2% of all clips in 2011). The number of examiner marks collected for each clip varied according to the number of candidates attempting the part-question and the speed with which individual markers

worked (see Table 3). In general, different candidates provided the clips for the different part-questions in the set of seed clips for geography.

Table 3 Clip statistics for geography

Year	*Number of part-questions*	*Number of clips*	*Number of candidates*	*Average marked responses per clip*
2009	12	138	114	20.4
2010	12	69	64	35.1
2011	12	130	109	21.7

For psychology, the numbers of clips also varied across the years and were also extremely small (like geography, fewer than 0.2% of all clips in 2011), but this time the number of marked clips was similar across part-questions within a year (since all candidates were required to attempt all the questions in the paper). See Table 4 for details.

Table 4 Clip statistics for psychology

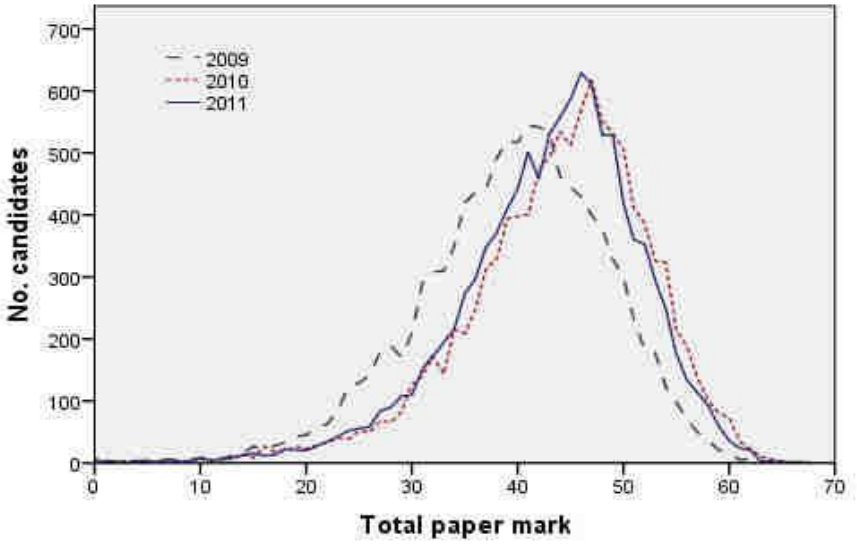
Year	*Number of part-questions*	*Number of clips*	*Number of candidates*	*Average marked responses per clip*
2009	15	313	159	11.5
2010	11	210	89	15.5
2011	14	223	54	20.0

In backreading, team leaders, now acting as marking supervisors, re-mark a proportion of clips (roughly 10% in geography and 6% in psychology in 2011, for example), the clips being selected and delivered to markers at random by the IT clip distribution system throughout the operational marking period. The supervisor either agrees to the original marker's mark or indicates that it should be replaced. The re-marking is not independent, since the team leader has sight of the original mark. Corrective action can include removing the examiner from the marking process altogether and re-marking the clips already marked by that individual. Whenever clips are re-marked, the mark awarded by the senior marker replaces the original mark.

3.4 The resulting mark distributions

Figure 1 shows the total mark distributions for all three years for geography unit papers, while Figure 2 shows the corresponding distributions for the psychology papers in those years.

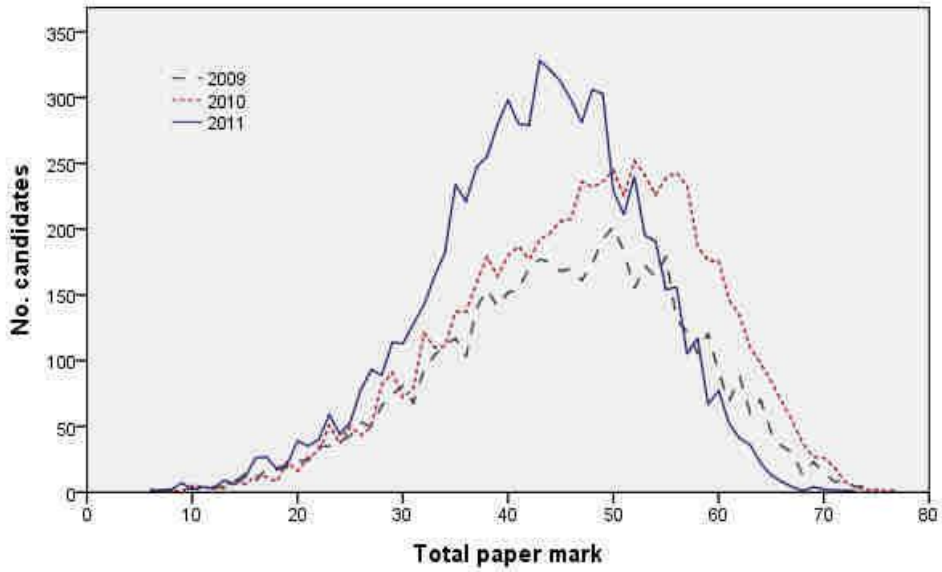
Figure 1 The mark distributions for the geography unit papers 2009-2011



Note in Figure 1 the similarity in the mark distributions for 2010 and 2011 compared with 2009. This is undoubtedly explained by that fact that, as noted earlier in this chapter, the 60-minute time allowance for the paper in 2009 was increased to 75 minutes in 2010. In 2010 and 2011, therefore, candidates had 25% more time in which to respond to the same number and type of extended response questions. This would explain the shift up the mark scale of the 2010 and 2011 mark distributions. The annual standard setting process will have addressed this issue and adjusted grade boundaries accordingly.

In psychology the picture is different, in that it is the 2011 distribution that differs from those of 2009 and 2010 in general shape. The mark distribution for 2011 is more symmetric than those for 2010 and 2009, both of which are left-skewed. Again, these differences will have been taken into account when grade boundary decisions were made during the annual standard setting processes.

Figure 2 The mark distributions for the psychology unit papers 2009-2011



The critical question for this project is what can we say about assessment quality in this context, in terms in particular of marker reliability?

4 Generalizability theory

4.1 Overview of G-theory

G-theory is of particular interest in the context of this project because it is the only modern theory of measurement that has been developed specifically as a means of investigating and refining our understanding of measurement reliability (and, incidentally, also of validity, the two together being dubbed 'dependability'). As the originators of G-theory succinctly note: '*The question of "reliability" ... resolves into a question of accuracy of generalization, or generalizability*' (Cronbach, Gleser, Nanda and Rajaratnam, 1972, p.15, original italics). Specifically, the theory relies on notions of domain and replication to produce estimates of reliability which can be 'generalised' to a given defined universe. As Brennan notes,

> Generalizability theory enables the investigator to identify and quantify the sources of inconsistencies in observed scores that arise, or could arise, over replications of a measurement procedure. (Brennan, 2001a, p.2)

With particular reference to studies of marker reliability, the use of G-theory has a long and influential tradition. The seminal text of Cronbach et al. (*cit*), for example, contains an extended analysis of the interactions of raters (there called 'scorers') with other conditions of testing in the context of a test of communicative ability in aphasic patients (pp. 161-178), as well as an example of a classroom study involving observers rating both teachers and pupils simultaneously (pp. 194-215). In the literature devoted explicitly to rater reliability, the classic paper of Shrout and Fleiss (1979), cited 7,510 times[1], uses the conceptual apparatus of G-theory to disentangle the confusions surrounding six different uses and interpretations of the intra-class correlation coefficient as an index of reliability.

The conceptual basis of G-theory inherits from the classical tradition of reliability theory – with some terminological innovation – the definition of an observed score as the linear combination of true score and measurement error, the partition of observed score variance into true score variance and error variance, and the characterisation of reliability as a function of the proportion of observed variance not due to measurement error. Whereas conventional treatments deal only in two sources of variation in scores, however, observed scores in G-theory are decomposed into linear combinations of as many sources of assessment unreliability as are appropriate to the investigation – candidates, questions, raters, training teams, occasions – and, crucially, of their interactions.

[1] Google scholar search result.

Some of these components will be considered as contributing to the true score and some as engendered by measurement, but there are no fixed rules. Part of an investigator's skill lies in determining, within the context of their investigation, how these multiple score components should be partitioned into contributors to true score variance and error variance. In principle, 'random' factors are factors that are sampled, and effects associated with these contribute to measurement error, whereas 'fixed' factors are not sampled and therefore play no part in measurement error. For a full discussion of this issue see, for example, Cardinet, Johnson and Pini (2010).

To illustrate the principle, consider the complete linear decomposition of an observed score Y_{cmq} of candidate c rated by marker m on question q, as shown in Equation 1.

Equation 1

$$Y_{cmq} = \mu + \nu_c + \nu_m + \nu_q + \nu_{cm} + \nu_{cq} + \nu_{mq} + \nu_{cmq} + \varepsilon_{cmq}$$

The terms in ν with a single subscript on the right hand side of Equation 1 are *main effects*; those with more than one subscript are the interaction terms, representing the joint effect of components indexed by the subscripts as they act together to influence the observed score. The error term ε_{cmq} represents any effects that are not captured by the other right-hand-side terms; as such it is often called the *residual* term. Very few assumptions are needed to support practical application of Equation 1. The main ones are that all of the right-hand-side terms, except μ, which is always constant, when interpreted as random variables have expected value 0, so that $\mathcal{E}(Y_{cmq}) = \mu$, and are uncorrelated. No distributional assumptions are made at this stage beyond the requirement that the means and variances exist; in particular there are no normality distributional assumptions.

There are two circumstances under which distributional assumptions might be needed:

1. Many estimation techniques depend on defining a likelihood or prior distribution for the observations. Whenever possible, ANOVA-based estimation techniques, which only use least squares, are to be preferred. For unbalanced data, as long as the lack of balance is not outrageous, the so-called Henderson methods, otherwise known as analogous ANOVA, give satisfactory results without requiring parametric assumptions about the distribution of the observed data.
2. Although G-theory is not conceived within an inferential framework (it does not require F-tests to establish model fit, for example), it may be useful to construct confidence intervals around sample mean scores, most commonly based on a standard error of measurement. In such cases, there is normally a sufficient number of observations used to estimate the standard error of measurement, that we can

reasonably appeal to the Central Limit Theorem to justify defining the confidence interval (see, for example, Snedecor and Cochran, 1989).

4.2 Variance decomposition

On the basis of these few assumptions, we can define a decomposition of the observed score variance in terms of variance components corresponding to each of the effects in Equation 1:

Equation 2

$$\sigma_Y^2 = \sigma_c^2 + \sigma_m^2 + \sigma_q^2 + \sigma_{cm}^2 + \sigma_{cq}^2 + \sigma_{mq}^2 + \sigma_{cmq}^2 + \sigma_\varepsilon^2$$

The development of G-theory – as well as many of the principles of its application in practice – has been from the start heavily influenced by the theory of experimental design, which also has variance analysis as a central theme. Probably as a result of this close association, equations like Equation 1 are more likely to be referred to as *designs* than *models*, though it is also the case that hard-core G-theory advocates prefer to eschew the term 'model', perhaps to emphasise the paucity of assumptions needed to support a G-theory analysis.

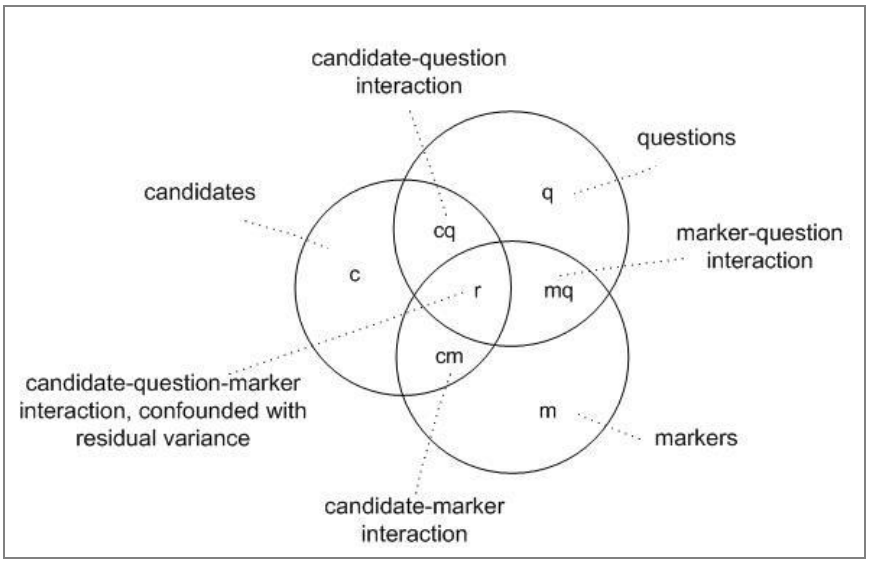
In practice, when we actually apply a design to real observations, it will almost always be the case that we only have available a single observation with which to estimate both ν_{cmq}, the interaction of one candidate with one question and one marker, and its variance, σ_{cmq}^2, which is clearly not sufficient. In fact, it will in general be impossible to disentangle the highest-order interaction effect from the residual error – the two effects are *confounded*. Consequently, we will usually write variance component equations like Equation 2 with a single term, which we notate σ_r^2, conflating the last two terms. The new look version of equation Equation 2 would be:

Equation 3

$$\sigma_Y^2 = \sigma_c^2 + \sigma_m^2 + \sigma_q^2 + \sigma_{cm}^2 + \sigma_{cq}^2 + \sigma_{mq}^2 + \sigma_r^2$$

Designs like Equation 1, and the resulting variance decomposition Equation 2, are quite simple by G-theory standards. The proliferation of subscripts and the strings of almost identical symbols can become impossible to read, and the information therein impossible to digest. An alternative representation, the variance component diagram, of which Figure 3 is an example, can be a very helpful visual aid in clarifying the relationship between sources of variation in a design.

Figure 3 The c x q x m design



The variance of a main effect is represented as a circle (occasionally as an ellipse when the topology becomes too complex). The intersection of two or more circles denotes the presence of an interaction between both or all of the intersecting effects. Two caveats: these are not Venn diagrams, and the size of a circle is not intended to bear any relation to the importance or magnitude of the component it represents. We make frequent use of these diagrams in the following exposition.

Depending on the aim of the analysis, certain variance sources are considered to be contributing to 'true' or valid variance and others to error variance, with some making no contribution to either. Classical reliability coefficients are most often constructed as *intra-class correlations*, ratios of true score variance to the combination of true score and error variance. Generalizability coefficients are a generalisation of the same idea, where designs like Equation 2 are the basis for deciding which linear combinations of variance components represent the valid variance and which the error variance. For further information on the conduct of a G-theory analysis see recent Ofqual reports by Johnson and Johnson (2012a, 2012b).

One of the innovations of G-theory is the recognition of a distinction between relative and absolute error variance. The measurement error variance for the case of 'relative' candidate measurement (typically norm-referenced applications) is a linear combination of the interaction variances involving candidates – in the design shown in Figure 3 these would be the candidate-marker interaction variance, the candidate-question interaction variance, and the confounded residual variance. The measurement error variance for 'absolute' candidate measurement is a linear combination of these same interaction variances, but this time with the main effect variances also added in.

Thus the error variances for candidate measurement are:

Equation 4

$$\sigma^2_{Rel}(candidates) = \sigma^2_{cm}/n_m + \sigma^2_{cq}/n_q + \sigma^2_r/n_m n_q$$

Equation 5

$$\sigma^2_{Abs}(candidates) = \sigma^2_m/n_m + \sigma^2_q/n_q + \sigma^2_{mq}/n_m n_q + \sigma^2_{cm}/n_m + \sigma^2_{cq}/n_q + \sigma^2_r/n_m n_q$$

where the denominators represent the sample sizes for the different effects.

Similarly, the measurement error variance for the case of 'relative' marker measurement (inter-rater reliability) is a linear combination of the interaction variances involving markers – in Figure 3 these are the candidate-marker interaction variance, the marker-question interaction variance, and the confounded residual variance (hence Equation 6). Again, the measurement error variance for 'absolute' marker measurement is a linear combination of these interaction variances and the between-candidate and between-question variances (Equation 7).

Equation 6

$$\sigma^2_{Rel}(markers) = \sigma^2_{cm}/n_c + \sigma^2_{mq}/n_q + \sigma^2_r/n_c n_q$$

Equation 7

$$\sigma^2_{Abs}(markers) = \sigma^2_c/n_c + \sigma^2_q/n_q + \sigma^2_{cq}/n_c n_q + \sigma^2_{cm}/n_c + \sigma^2_{mq}/n_q + \sigma^2_r/n_c n_q$$

From the estimated measurement error variance comes the standard error of measurement (SEM), and from the SEM the 95% confidence interval around a mean score or total score can be produced. What-if analyses can follow, in which the effect on the measurement error variance, SEM and confidence interval of changes in the measurement conditions can be predicted, by substituting previous factor sample sizes with alternatives. Thus, for candidate measurement we could explore the influence on total score precision of single-marking versus double-marking combined with changes in the numbers of questions in the component paper.

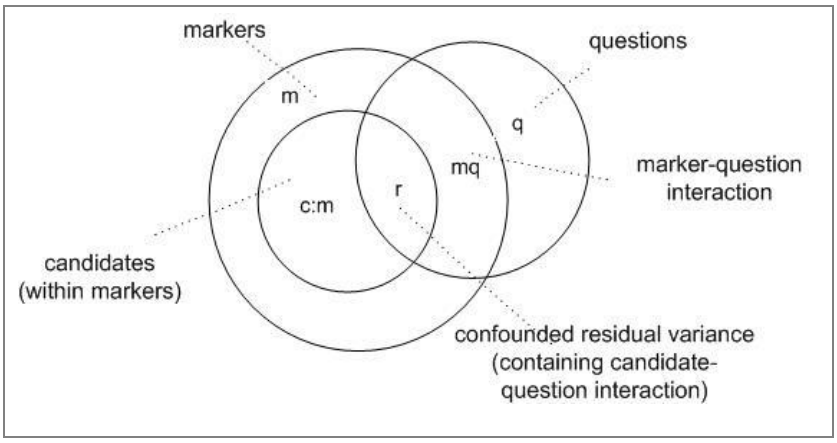
The what-if facility is a generalisation of the Spearman-Brown prophecy principle which allows prediction of the effect on reliability of changes in the conditions of assessment – for example, increasing the number of examiners marking any one question or script and/or changing the number of questions overall or within sections of the paper. Given the symmetry of the G-theory approach, the reliability of estimates of marker severity can equally be produced, and impacts on that reliability predicted should the number of

different questions or candidate scripts per marker be changed. We are thus able, provided data collection – the experimental design – has been organised to support extraction of the information we need, to apply the principles of G-theory to quantify the contribution of different sources of score variation to measurement error, and to use the results to identify the optimum future operational strategy for delivering and marking tests (see Wood, 1991; Johnson and Johnson, 2012a, 2012b; Bramley and Dhawan, 2012).

Thus far we have only considered the case of fully crossed designs, where all main effects and all interactions between the main effects are present. In the operational process, however, fully crossed designs typically do not hold. In particular, single-marking of scripts is the norm. This means that the mark data resulting from operational marking are not useful for estimating the reliability with which candidates will have been measured. The reason for this is that candidates are now *nested* within markers, so that we no longer have any way of quantifying the impact on candidate measurement reliability of any between-marker variation (which contributes to absolute measurement error) or of any candidate-marker interaction variance (which contributes to both absolute and relative measurement error).

With the notation c:m representing candidates nested within markers, Figure 4 gives the variance component diagram for the c:m x q design. All that can be done in this situation is to ignore the presence of markers in the assessment process altogether, so that 'markers' becomes a 'hidden factor' in the analysis, and simply look at the impact on candidate measurement error of test questions (using the c x q design, as in many of the application examples offered in Johnson and Johnson, 2012a). Ignoring markers, however, would reduce the validity of the reliability estimation, and cloud interpretation of the results.

Figure 4The (c:m) x q design



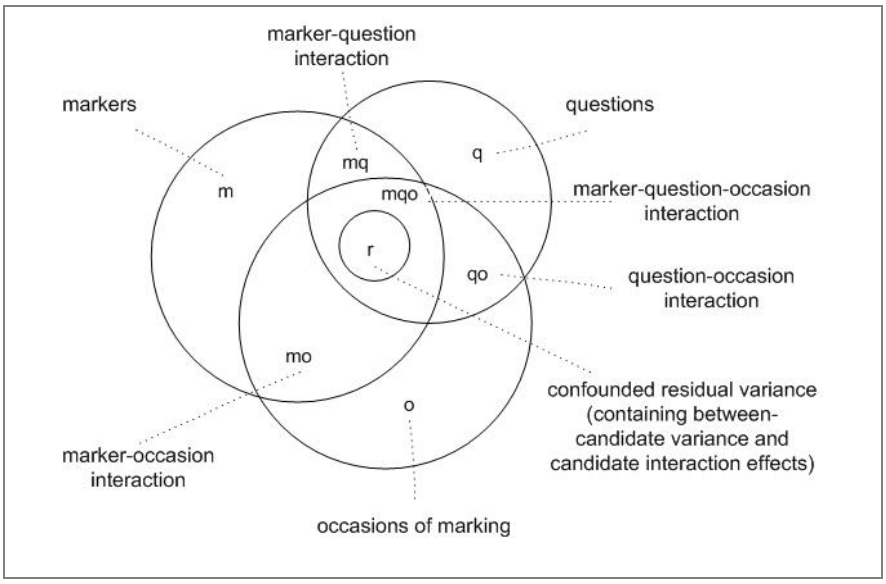
On the other hand, the single-marking of scripts does not necessarily of itself have the same devastating effect on marker reliability estimation, if it can be assumed that the candidates nested within markers are interchangeable samples (i.e. similarly representative of the whole component entry). Random allocation of scripts to markers should achieve this requirement, whereas the allocation of centres to markers, a common strategy in the days of paper-based scripts, would most likely not do so. If equivalent candidate samples *could* be assumed then the measurement error variance for relative marker measurement would be a linear combination of marker-question interaction variance, between-candidate (within marker) variance, and confounded residual variance. The between-question variance would be added to the linear combination for absolute marker measurement.

4.3 From script marking to clip marking

But the current situation is that scripts are no longer sent in their entirety to markers, but are decomposed into individual questions or part-questions before being delivered electronically to markers as 'clips' for online marking. For marker measurement the situation, while more complicated, is still accessible. Indeed, since randomised allocation of questions or part-questions is the norm, the previous potentially problematic effect of confounded marker-centre interaction is eliminated. Reliability analysis has simply become more challenging.

Despite analysis complexity, it would be possible to extend the model to explore the issue of potential marker drift over time. The new analysis design is shown in Figure 5. We have simply added 'occasions' to the picture. This extended design allows quantification of the contributions to score variation of, in particular, markers, questions, occasions and their interactions. This will in turn enable estimation of the reliability with which markers can be measured, overall and on different occasions. The findings have implications for evaluation of the efficacy of routine quality assurance procedures.

Figure 5The c:(moq) design



But can we say anything about the reliability with which candidates were assessed on an examination component? The answer is only if we have data from a situation where individual candidates were independently marked by more than one marker, as Bramley and Dhawan (2012) observe:

> From a measurement perspective, the ideal scenario for quantifying marker agreement is when two or more markers mark the same piece of work without knowledge of what marks the other markers have given to it (referred to as 'blind' double or multiple marking). The marks can then be treated as independent in the statistical sense which is usually an assumption of most common methods of analysing agreement. (Bramley and Dhawan, 2012, p.268)

Single-marked clip data from the operational process is not useful for this purpose. Backread data are not useful either, because although each clip is marked by two markers one of these markers, the team leader, has knowledge of the mark awarded by the marker whose marking performance is being checked. The two marks are not, therefore, independently awarded.

Seeded clip data, on the other hand, do have potential for this kind of analysis, given that seeded clips are independently marked by several, and sometimes all, of the regular markers. Seeded clips are distributed to the regular markers at various points in the operational marking process. The examiners mark these, as they do the other clips that are randomly allocated to them, unaware that they are seeded and therefore pre-marked by a principal examiner.

In principle, therefore, seeded clip data offer scope for estimation of the reliability with which candidates are measured, on individual part-questions and for the paper as a

whole. However, for reasons that will be explained as the results of seeded clip analyses are presented later in this chapter, the data that resulted from the seeded clip benchmark checks for each of the AS-level papers considered in this project were not ideal for the purpose of estimating component reliability. Analyses are nevertheless reported as illustrations of the potential value of seeded clip data for component reliability investigation, value that could be maximised by some adjustment to the scope and scale of benchmarking exercises in the future.

All of the reported G-theory analyses featuring markers were carried out using data from experienced regular markers only; the one or two new markers that participated in the marking in one or other year were excluded, as were supervisors, principal and chief examiners. In every case the unit of analysis was the mark given to a single clip by a single marker. Depending on their complexity and the amount of data involved, analyses were carried out using urGENOVA (Brennan, 2001b), mGENOVA (Brennan, 2001c) or SPSS (variance components procedure, for crossed designs with small datasets).

4.4 Geography G-study analyses and findings

As noted in Chapter 3, the geography unit paper carried the same structure and mark tariff pattern across the three years – only the specific question content changed, as did the time allowance, which was increased from 60 to 75 minutes from 2010. The paper comprised two sections, each presenting two three-part questions of which candidates were to choose one. Questions carried 35 marks, for a two-question paper total of 70 marks, with part-questions (e.g. q2a, q2b, q2c) carrying 10 or 15 marks.

In practice, given the question choice available to candidates, each geography paper actually embodied four pathways in terms of unit assessment, since candidates could choose to respond to any one of four question combinations: questions 1 and 3, 1 and 4, 2 and 3, or 2 and 4. Of the more than 10,000 candidates each year who took this unit paper, the majority (around 60%) chose question combination 2 and 4 in each year (Table 5), presumably having been prepared by their schools for the topics those questions related to.

Table 5 Candidate numbers for the geography unit paper

Year	*q1+q3*	*q1+q4*	*q2+q3*	*q2+q4*	*All**
2009	413	2,585	837	5,504	11,102
2010	415	2,590	800	6,028	10,844
2011	393	2,859	668	6,022	10,980

** Includes some resit candidates, and candidates who attempted at least one part-question over and above a valid combination of six*

4.4.1 Marker effects study for geography (operational data)

Table 6 presents the results of the analysis of the single-marked operational data for geography for each of the three years, looking across all 12 part-questions (or 'items') in each paper. The underpinning design is illustrated schematically in Figure 5, with the addition of markers nested within teams (imagine Figure 5 with an additional 'team' circle encompassing the marker circle). The primary intention was to investigate the possibility that there might be training team and marker drift effects at play, and if so to quantify them. A secondary objective was to estimate the reliability of marker means when calculated over different numbers of clips.

Sequential periods of marking were identified for individual markers so that the possibility of marker drift might be investigated. In general, during the one-month operational marking period, markers were allocated and marked several hundred clips (between 1,280 and 2,740 per marker in 2011) from across all the part-questions in the paper for the year in question. Different markers marked their allocated clips at different rates (which explains the wide range of clip numbers marked per marker) and over different sub-periods: for example one marker might have marked at a steady pace throughout the month while another might have marked almost all allocated clips within the first two weeks. It would make little sense, therefore, to look at the issue of possible marker drift by dividing the marking month into the same sub-periods for all markers, since some markers would be represented by very few clips for one or other of the four quarter-periods. In the interests of valid interpretation, therefore, four consecutive marking periods were identified for each marker individually. These were determined by dividing each marker's clip allocation into quarters over time (clip marking is routinely time-stamped), and rounding the resulting marking periods to one day.

These 'periods of marking' should not be considered as specific time periods, but rather as phases within each marker's personal marking schedule. The first phase, while

possibly different in real time from one marker to another, has the same meaning for all markers as the earliest marking period, and so on. For this reason we can say that markers are crossed with periods, as they are also crossed with questions (clips are nested within markers, but markers participate in the marking of all the questions in the paper). It is arguable whether 'periods of marking' should be considered in analysis as a fixed or a random factor. For ease of analysis we have treated it as random; in the event, given the low variance components associated with the factor 'periods' the same results would have arisen had we treated the factor as fixed.

Table 6 Variance breakdown for the single-marked operational datasets for geography

(> 60,000 single-marked clips per year, covering 12 part-questions (items); four teams of 5-8 markers each depending on year, four sequential marking periods)

	Component estimates			% contributions**		
	2009	2010	2011	2009	2010	2011
Teams	0.001	-0.039	0.028	-	-	-
Periods of marking*	-0.009	-0.004	0.001	-	-	-
Items	2.200	2.591	2.303	30	33	33
Team-period interaction	-0.018	0.010	0.005	-	-	-
Team-item interaction	0.023	0.010	0.023	-	-	-
Period-item interaction	0.028	0.003	-0.004	-	-	-
Team-period-item interaction	0.031	0.001	0.012	-	-	-
Markers within teams	0.282	0.457	0.206	4	6	3
Marker-period interaction	0.165	0.020	-0.001	2	-	-
Marker-item interaction	0.097	0.107	0.102	1	1	2
Marker-period-item interaction	-0.066	0.083	0.045	-	1	1
Residual***	4.582	4.576	4.286	62	58	61

** Each marker's operational marking period (identified by virtue of clip time-stamping) was roughly divided into four sequential 'periods of marking' in terms of clip counts*
*** Rounded percentage contributions of each component to the sum of estimated variance components; only contributions of at least one per cent are shown*
**** The residual variance will contain all candidate-related variance contributions, including interaction contributions, confounded with random error*

Perhaps the first feature to note in Table 6 is the presence of negative variance component estimates. In theory variances cannot be negative. In practice, when a population variance is zero, or very close to zero, then its sample-based estimate can be expected to be close to zero as well, in either direction. Where negative component estimates are much larger than zero this can indicate data inadequacy, i.e. too little data being over-exploited in an analysis involving too many factors. A common practice is to set small negative variance estimates to zero in follow-on calculations; we follow this practice here.

The second and related feature to note is that Table 6 provides no evidence of any marking period effect in the operational data, and neither is there any noticeable evidence of any team effect; this absence of evidence of team and period effects extended to every part-question in the geography papers. It should be noted that had there been strong evidence of a team effect, this would not necessarily have been interpretable as evidence of an influence of team leader training on team members, since it could just as readily have indicated pre-existing dispositions to relative severity/leniency on the part of the small number, 5-8, of markers who comprised each team in one year or the other.

In each year close to one-third of the total clip score variance is attributable to between-item variance, a phenomenon at least partly explained by the fact that some part-questions carried mark tariffs of 10 and others mark tariffs of 15. Between-marker differences make a very small contribution to total clip score variance in comparison, at between 3% and 6%. It should be noted that this is what is observed in the dataset. It can only be considered a reflection of genuine differences in markers' overall standards of judgement (severity/leniency) if it can be assumed that the sets of clips that different markers marked were randomly parallel, and therefore interchangeable. While the automated clip delivery system offered clips to markers in a random allocation process, and each marker did mark a relatively large number of clips, it must nevertheless be recognised that marker standards and clip batch difficulty are confounded in these data.

Around 60% of the clip score variance is residual variance, within which will be several confounded effects associated with the candidates, principal among them the between-candidate variance, the marker-candidate interaction variance and the candidate-question interaction variance, none of which can be isolated in the single-marked clip-level data.

Using the analysis results in Table 6, it might be interesting to look at the impact of clip numbers on the precision with which markers' mean scores might be measured. Table 7 provides this information for clip sample sizes of 5 (typical for quality checks in the operational process), 15 and 25 for each of the part-questions included in Table 7.

Table 7 Predicted precision of markers' mean clip scores in geography

	2009	*2010*	*2011*
*SEM associated with clip mean score**			
across 5 clips	1.09	1.06	1.00
across 15 clips	0.76	0.72	0.66
across 25 clips	0.67	0.63	0.56

** The relative measurement error variance was produced by dividing the residual variance by the relevant clip sample size, and adding to the result the marker-interaction variance components.*

In each year, for samples of just five candidates, 95% confidence intervals around markers' mean clip scores would be roughly ± 2 marks, reducing to around ± 1½ marks for samples of 15 clips, and reducing again to under ± 1½ marks for samples of 25 candidates. The figures would naturally vary somewhat across different part-questions.

Unfortunately, it is impossible using the operational data analysis results in Table 6 to provide estimates of the reliability with which candidates were measured by the component paper each year, since any variance contributions from candidates, and from candidate interactions with markers, questions, and so on, are confounded in the residual variance (as noted earlier, the single marking of clips renders it impossible to access this information). The best that we can offer in the circumstances are reliability estimates based on analyses in which all possible marker contributions to measurement error are ignored; in the analyses 'markers' is a 'hidden factor'. Since teams and periods of marking have proved to be irrelevant in reliability terms, these can be excluded from further consideration.

The fact that candidates had question choices within each paper demands a separate analysis for each of the possible pathways (questions 1+3, 1+4, 2+3, 2+4). The analysis design is complicated by the fact that the part-questions, the items, carried different mark tariffs: for each choice of question pair (each chosen pathway) four part-questions carry 10 marks each while the other two carry 15 marks each. This means that a multivariate composite score analysis is required (see Brennan, 2001a, section 10, for details; He, 2012, for an overview; Johnson and Johnson, 2012a, and Johnson, Johnson, Miller and Boyle, 2013, for example applications). The analysis design is therefore now c x (i:t) for each pathway, representing candidates crossed with items and items nested within tariffs. The analysis results are shown in Table 8.

Table 8 Composite score reliability for geography (relative generalizability coefficients)

(design c x (i:t), for candidates by items within tariffs; markers are a 'hidden factor', whose effect on reliability cannot here be separately quantified and taken into account)

	Pathways through the paper			
	q1+q3	*q1+q4*	*q2+q3*	*q2+q4*
2009	0.70	0.63	0.66	0.58
2010	0.60	0.56	0.50	0.56
2011	0.72	0.65	0.59	0.62

** 'Tariff section' weights in analyses were 0.571 and 0.429, respectively, for the 10-mark and 15-mark part-questions.*

In line with previous results for similar component papers in other subjects (Johnson and Johnson, 2012a) Table 8 shows modest reliability coefficients for the geography papers, at between 0.5 and 0.7, depending on year and question choice. These are relative generalizability coefficients; coefficients for absolute measurement are very slightly lower in all cases. Alternative analyses that recognised the actual labelled sections in each paper, with three part-questions per geography topic, produced similar results. The question is, how much lower might the coefficients become if we could include marker effects in the analysis? For a possible answer to this question we look now at what the seeded clip data might offer.

4.4.2 Seeded clip analyses for geography

In principle the set of seeded clip data would have been ideal for G-study analysis, following the design illustrated in Figure 3, with 'questions' simply replaced by 'part-questions'. With sufficient data, a single analysis for each year and pathway could have quantified the variance components for markers, part-questions and candidates, as well as for all interactions among these. This information could then have been used in the usual way to meaningfully quantify marker effects, as well as to estimate component reliability for the operational case of single-marking through a what-if analysis.

However, as noted in Chapter 3, despite the very large candidate entry for this subject each year, at over 10,000 candidates, providing over 60,000 clips for processing, the numbers of seeded clips that were used to quality check marker performance were very small each time: 138, 69 and 130, respectively, in 2009, 2010 and 2011, frequently with single-digit numbers of clips representing a part-question. Moreover, the clips were for the most part taken from different candidates; in other words, for any candidate there was usually just one clip, i.e. one part-question, represented in the seeded clip set. In

addition, not all the markers marked every clip; some clips were indeed marked by just a handful of the operational markers. Finally, the clips that had been selected for marker monitoring using clip seeding were specifically selected by the principal examiner as being of a particular kind. In other words, the clips were not randomly selected to reflect the total entry for a part-question nor to represent candidate subgroup such as borderline candidates; clips were specifically selected by the chief examiner for their particular qualities, perhaps because the candidate performances were uneven or in some other way difficult to mark.

These issues notwithstanding, for illustrative purposes we offer G-study analyses for the 2011 paper only, using data relating to question 2 and question 4, whose part-questions were represented by the largest number of clips in all three years, albeit at under 15 each (Table 9). Individual clips were marked by between three and all 37 markers. As a result, the part-question datasets were not only very small in size but the data matrices (candidates by markers) were sparse to varying degrees.

Table 9 Clip-level data availability for geography in 2011

Part-question (item)	Clips*	Markers/clip**
q2a	12	6-37
q2b	14	3-37
q2c	11	3-37
q4a	13	10-37
q4b	14	3-35
q4c	14	1-37

** The sets of clips were not from the same candidates*
*** Numbers of markers who marked individual clips*

We begin with the simple design c x m, i.e. candidates crossed with markers, for each of the six separate part-question datasets. For analysis purposes markers who marked fewer than five clips for a part-question and candidates whose part-question clip was marked by fewer than five markers were excluded. The analysis results are given Table 10. Note in every case the low between-marker variance contribution to total score variance compared with the contribution of the between-candidate variance, and, in particular, the very high contribution from residual variance, much of which can be assumed to be attributable to marker-candidate interaction variance. The similarity in

profiles between the first and second parts of each question compared with the third is also worth reflection.

Table 10 Generalizability analyses of seeded clip data for geography in 2011

Part-question	*q2a*	*q2b*	*q2c*	*q4a*	*q4b*	*q4c*
Mark tariff	10	15	10	10	15	10
No. markers/clip	6-31	12-31	26-31	10-31	27-30	30-31
No. clips/marker	7-11	6-11	6-10	8-13	12-13	11-12
Variance component estimates						
Markers	0.136	0.349	0.228	0.171	0.532	0.193
Candidates	0.821	1.411	3.180	1.440	2.267	2.225
Residual*	1.034	1.963	1.381	1.340	1.961	1.004
*% contributions to total variance (rounded)***						
Between-marker variance	7	9	5	6	11	6
Between-candidate variance	41	38	66	49	48	65
Residual variance	52	53	29	45	41	29

**The residual variance contains marker-candidate interaction variance*
***Calculated as percentages of the sum of variance components.*

With up to 30 or so examiners marking any one clip, candidate measurement at part-question level would be extremely reliable, with very high reliability coefficients for both relative and absolute measurement, and quite precise mean scores. The interesting question at this point, though, must be what happens to reliability in the regular operational procedure, when just one marker marks each clip? We can now use the G-study information presented in Table 10 to carry out what-if analyses to answer that question, though only at the level of part-questions. Table 11 provides the results.

Table 11 Estimated candidate reliability for single-marked geography items in 2011

	q2a	*q2b*	*q2c*	*q4a*	*q4b*	*q4c*
Reliability coefficients						
Relative measurement	0.44	0.42	0.70	0.52	0.54	0.69
Absolute measurement	0.41	0.38	0.66	0.49	0.48	0.65

As we can see from Table 11, the estimated reliability of candidate measurement for part-questions is universally low for single marking; only for part-questions q2c and q4c do the reliability coefficients approach conventionally acceptable values, even for relative measurement. Blind double marking would be predicted to improve the picture somewhat, with coefficients of around 0.8 appearing for q2c and q4c.

The final question must be what happens if we look beyond part-questions?

Had a sizeable number of the same candidates appeared in the different sets of part-question clips then the c x i(t) model analysed earlier using operational data (with markers a 'hidden factor'), and reported in Table 6, could have been extended to embrace marker effects in the model c x m x i(t). The resulting analysis would have given all the information about variance contributions needed to provide reliability estimates for the whole paper (for an example see Johnson et al., 2012). Unfortunately, no common candidates appeared in the clip sets for the six part-questions. Given this, there is no possibility of carrying out the required G-study analysis. The confounding of variables means that neither the large-scale operational data nor the small-scale seeded clip data offer any scope for providing valid estimates of the reliability of candidate measurement at component level for geography, since any estimates that could be produced would not simultaneously be able to take into account the contributions to unreliability of question and marker sampling, or of interaction effects between questions, markers and candidates.

4.5 Psychology G-study analyses and findings

Recall from Chapter 3 that each 100-minute psychology paper contained three sections of questions, with no question choice, for a maximum mark of 80. Section A comprised machine-marked multiple-choice questions for 10-12 marks in total. Section B, the largest contributor to the paper total mark with between 42 and 49 marks each year, presented a mix of short-answer and extended-response questions. Section C comprised two extended–response questions for a total of between 18 and 26 marks each year. In Section A the majority of questions were binary-scored with a question or two carrying two marks, while in Sections B and C mark tariffs varied across questions and part-questions. The trained operational markers marked those questions and part-questions in Sections B and C – the majority – that required an element of judgement in applying mark schemes, the remaining questions were marked by clerical staff.

Marker effects are the principal focus of interest in this project, so we exclude Section A from further consideration for the moment, along with all clerically marked data in Sections B and C. Following the pattern established for geography, we begin with analyses of operational data from the three years, before moving on to consider what the seeded clip data might offer in addition.

4.5.1 Marker effects study for psychology (operational data)

Following the strategy outlined earlier in this chapter for geography, to explore the possibility of marker drift, four sequential 'marking periods' were identified for each individual marker. These essentially divided each marker's complete marking period

into quartiles, in terms of the number of clips marked. The results of the global operational data analysis for psychology are shown in Table 12.

Table 12 Variance breakdown for the single-marked operational datasets for psychology

(> 60,000 single-marked clips per year covering 10 or more part-questions (items), three teams of 2-7 markers each depending on year, four sequential 'periods of marking')

	Component estimates			% contributions**		
	2009	2010	2011	2009	2010	2011
Teams	-0.001	0.005	0.016	-	2	-
Periods of marking*	0.003	-0.003	0.002	-	-	-
Items	1.555	1.207	1.677	46	36	50
Team-period interaction	0.009	0.008	-0.015	-	-	-
Team-item interaction	-0.008	0.001	0.010	-	-	-
Period-item interaction	0.005	0.007	0.000	-	-	-
Team-period-item interaction	-0.016	-0.008	0.014	-	-	-
Markers within teams	0.041	0.026	0.027	1	1	1
Marker-period interaction	-0.017	0.008	0.048	-	-	1
Marker-item interaction	0.067	0.045	0.031	2	1	1
Marker-period-item interaction	0.052	0.019	0.004	2	1	-
Residual***	1.667	1.988	1.533	49	60	46

** Each marker's operational marking period (identified by virtue of clip time-stamping) was roughly divided into four sequential 'periods of marking' in terms of clip counts*
*** Rounded percentage contributions of each component to the sum of estimated variance components; only contributions of at least one per cent are shown*
**** The residual variance will contain all candidate-related variance contributions, including interaction contributions, confounded with random error*

Table 12 shows that in 2009 and 2011 the total variance in clip scores was fairly evenly attributable to between-item variance and residual variance, the latter subsuming between-candidate variance, candidate interaction variances and random variance. Markers contributed little to clip score variance. There is evidence of small differences in markers' overall marking standards – but recall that these could be spurious if the sets of clips marked by individual markers differed in important respects (markers marked between 1,450 and 6,760 clips, with varying spread over items). There is also some evidence of marker interaction effects of different kinds (across marking periods, with items, and with items across periods), but these are also very small.

We can use the information in Table 12 to estimate the reliability with which markers could be measured in terms of their mean clip scores for varying-sized samples of clips,

as previously illustrated for geography. The results for samples of 5, 10 and 15 clips are shown in Table 13. In each year, for samples of just five clips, 95% confidence intervals around markers' mean clip scores would be roughly ± 1½ marks, reducing to around ± 1 mark for samples of 15 clips, and reducing again to under ± 1 mark for samples of 25 candidates. As mentioned earlier for geography, the figures would be different across the different questions and part-questions.

Table 13 Predicted precision of markers' mean clip scores in psychology

	2009	*2010*	*2011*
*SEM associated with clip mean score**			
across 5 clips	0.67	0.69	0.62
across 15 clips	0.48	0.45	0.43
across 25 clips	0.43	0.39	0.38

** The relative measurement error variance was produced by dividing the residual variance by the relevant clip sample size, and adding to the result the marker-interaction variance components.*

Crucially, as far as component reliability estimation is concerned, we do not have from this analysis any quantification of the contribution to total clip score variance of marker-candidate interaction, since this is one of those potential sources of variance that are confounded in the residual. Thus, as was the case for geography, the information in Table 12 is inadequate for providing any completely valid indication of the reliability with which candidates were measured by the unit papers. Once again, all we can do is treat all clips as though they had been marked without error, and carry out a composite score analysis on that basis. Table 14 presents the results. Component reliability for the psychology paper is estimated at around 0.75 per year, modest, but slightly higher than for geography. This is for relative reliability, and it does not reflect any possible marker contributions to unreliability in Sections B and C.

Interestingly, the least technically reliable section in the psychology paper was Section A, the multiple-choice test. This section was very easy for candidates in every year, section scores clustering in the top half of the short (11-point in 2011) mark scale (see Figure 6). Section A could probably have been eliminated without reducing the reliability of the paper as a whole, given that candidate spread is an asset in relative measurement contexts such as this. That said, there are likely to be sound curriculum reasons for retaining this section of the question paper.

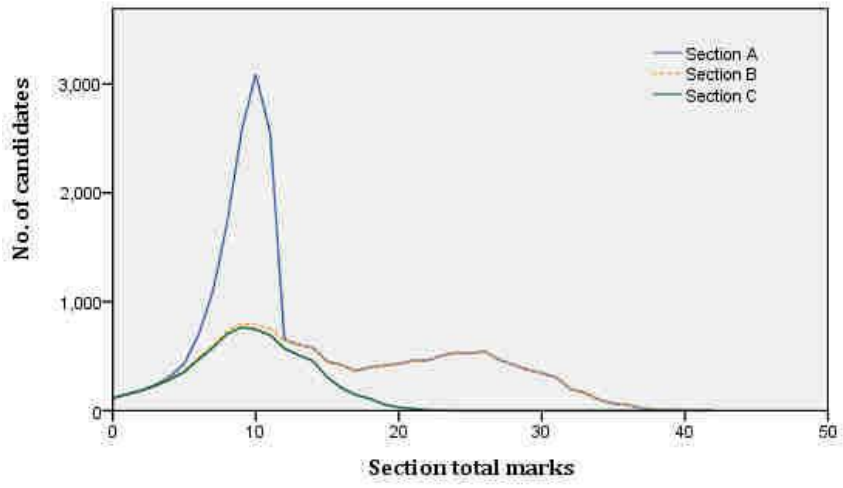
Table 14 Composite score reliability for psychology (relative generalizability coefficients)

(design c x (i:t), for candidates by items, i.e. questions or part-questions; markers are a 'hidden factor')

	*Paper sections **			*Whole paper*
	A	*B*	*C*	
2009	0.48	0.76	0.59	0.74
2010	0.46	0.76	0.62	0.77
2011	0.39	0.72	0.51	0.73

** For sections A, B and C, weights in analyses were: 0.150, 0.525 and 0.325, resp., for 2009; 0.163, 0.613 and 0.225 for 2010; 0.138, 0.550 and 0.313 for 2011.*

Figure 6Section score distributions for psychology in 2011



Section C, which comprised extended response questions, showed reliability coefficients of between 0.5 and 0.6, while Section B, the section of mixed response questions that carried most weight in the component, had coefficient values of over 0.7. Can the data from the seeded clip exercise throw more light on the reliability question, as far as marker effects are concerned?

4.5.2 Seeded clip analyses for psychology

Table 15 records the questions and part-questions for which seeded clip data were available for 2011, along with the numbers of clips that were seeded that year and the range of numbers of markers who marked individual clips within each set.

Table 15 Seeded clip statistics for psychology in 2011

Section	Item*	Clips**	Markers/clip***
B	q11a	20	1-17
	q12b	20	1-17
	q15a	17	12-17
	q16	17	14-17
	q17	18	4-17
C	q18	18	6-17
	q19b	17	15-17
	q19c	17	15-16

** Items (questions and part-questions) included in the seeded clip exercise*
*** There was little or no candidate overlap between the clip sets*
**** The numbers of markers who marked individual clips*

Table 16 presents the G-study results for Sections B and C in the 2011 paper (for some part-questions one or two clips were marked by a handful only of the markers; these clips have been dropped in the analyses). Between-marker variance made a generally low contribution to total clip score variance, the percentage contribution varying from just 1% to a maximum of 7%. With the exceptions only of questions 17 and 18, the highest contribution to clip score variation can be attributed to between-candidate variation, at 70% or higher (over 90% for question 11); the contribution from this source falls to just under 50% for question 17 and to under 30% for question 18. The residual variance in every case will contain the marker-candidate interaction variance.

Table 17 provides predicted reliability results for candidate measurement on questions and part-questions for the case of single marking, the routine operational marking pattern. We see that for most cases reliability coefficients, for both relative and absolute measurement, are between 0.70 and 0.80. For question 11 the coefficients are very high, as would be predicted from the almost negligible marker variance and the unusually low residual variance shown in Table 16 for this question. Questions 17 and 18, on the other hand, could be considered cause for concern. Reliability coefficients are extremely low in both cases, especially so for q18. Had this information been available ahead of time, these two questions could usefully have been targeted for blind double marking in the operational process.

Table 16 Generalizability analyses of seeded clip data for psychology in 2011

	Section B					*Section C*		
Part-question	*q11*	*q12b*	*q15a*	*q16*	*q17*	*q18*	*q19b*	*q19c*
Mark tariff	6	5	4	4	4	5	5	12
No. markers/clip	12-17	10-16	12-17	14-17	12-16	9-16	13-16	13-17
No. clips/marker	11-16	12-16	13-17	12-17	12-17	9-17	10-17	11-17
Variance component estimates								
Markers	0.011	0.067	0.067	0.048	0.043	0.120	0.052	0.569
Candidates	2.846	1.559	0.815	1.026	0.313	0.407	0.864	5.139
Residual*	0.158	0.474	0.277	0.257	0.312	0.879	0.257	1.284
*% contributions to total variance***								
Between-marker variance	1	3	5	3	4	9	4	7
Between-candidate variance	94	74	70	77	46	29	74	73
Residual variance	5	23	25	20	50	63	22	20

**The residual contains marker-candidate interaction variance*
***Calculated as percentages of the sum of component estimates*

Table 17 Estimated candidate reliability for single-marked psychology items in 2011

	q11	*q12b*	*q15a*	*q16*	*q17*	*q18*	*q19b*	*q19c*
Relative reliability	0.95	0.77	0.75	0.80	0.50	0.32	0.77	0.80
Absolute reliability	0.94	0.74	0.70	0.77	0.47	0.29	0.74	0.73

4.6 Data limitations

The single-marking of clips, like the single-marking of entire scripts that used to be the operational pattern, cannot provide appropriate data for the estimation of examination reliability; nor even simply for comparing marker standards unless the absolute comparability of delivered clip workloads can be guaranteed. Backread data cannot meet the requirements either, principally, though not only, because the two marks assigned to each clip are not independently given (the backreading supervisor having sight of the original operational marker's awarded mark). What is needed is a set of data comprising the marks independently awarded to the different elements of a paper for the same randomly selected candidates by all, or a large representative sample of, the operational markers.

Seeded clip exercises could, in principle, provide the required data, as well as continuing to serve ongoing quality assurance needs. It suffices to expand current

seeded clip exercises, whilst ensuring that all human-marked elements are similarly and adequately represented in the total clip set (in terms of numbers of clips, their representativeness element by element, and candidate commonality across elements).

To be specific, every human-marked question or part-question in an examination paper should have similar numbers of clips selected to represent it in a seeding exercise, even where the questions concerned might be attempted by a minority of the candidate entry (such as question 3 in geography). The clips should be randomly selected to represent the whole candidate entry or a specific subset of that, so that the results from the seeded clip analysis might be generalizable to some known larger group; 'unusual' question responses could continue to be selected by chief examiners from within this randomly chosen set for special review, or the set could be supplemented for additional quality assurance purposes (the supplementary clips not then included in the reliability analyses). Finally, the clips should derive from one group of candidates, so that across the questions there would be high commonality in candidate representation. In other words, clips could usefully be selected to facilitate analysis of the design c x q x m, meaning all the candidates in the study provide response clips for all the human-marked questions and part-questions in the paper, and these clips are independently marked by all, or a large representative sample of, the operational markers.

5 Rasch modelling

The Rasch model is a very restrictive model compared to other IRT models because it only estimates the difficulty of each item (part-question here) in addition to a single estimate for every person involved in the measurement process. Therefore, the Rasch model 'lacks' the flexibility of a two-parameter model (Birnbaum, 1968) which estimates the discrimination parameter for every item which is analogous to an item–total correlation from Classical Test Theory (Embretson and Reise, 2000). It also lacks the flexibility of other IRT models with even more parameters (e.g. a three-parameter model estimates a 'guessing' parameter for each question). Consequently, a Rasch model is often found to have worse model-data fit compared to other IRT models, which does not come as a surprise because it fits fewer parameters on the data.

We decided that the MFRM was appropriate for this project for a number of reasons:

(a) the Rasch model was considered to be theoretically appropriate for the purposes of this study and for the specific context of its datasets because it accepts candidates' raw scores as sufficient statistics for their ability estimates. This is in line with the tradition of the English examination boards that use raw scores to estimate the performance of candidates.

(b) In English public examinations, candidates are not rewarded for answering more difficult questions correctly and are not penalised for answering less difficult questions incorrectly. Therefore, in the context of this study it would not be ideal to use a model that used more than one parameter for items. Sticking with the Rasch model gives a monotonic relationship between the latent trait and the raw scores.

(c) The nature of the questions does not encourage guessing. Therefore a three-parameter model (incorporating a pseudo-guessing parameter) would be inappropriate for this study.

(d) The specific tests are not designed to be used as timed tests, although it is possible that time management is important during examinations.

(e) Experience with other datasets from English public examinations (e.g. Lamprianou and Boyle, 2004; Lamprianou, 2008; Ong, Williams and Lamprianou, 2011) suggested that the Rasch model might fit data of this kind reasonably well.

The next section presents the Rasch model and explains how we evaluated model-data fit.

5.1 The Many-Facets Rasch Model (MFRM)

Once we established that the use of a Rasch model would be theoretically reasonable for the purposes of this study, the natural choice among the Rasch models was the MFRM because it can handle additional facets of measurement beyond the questions and the examination candidates (for example, see Eckes, 2005; McManus, Thompson and Mollon, 2006). The Rasch model describes the interaction when an examination candidate encounters a test question (note that this is not to be confused with an 'interaction effect' in an analysis of variance context), estimating the abilities of candidates and the difficulties of the questions with reference to a single linear scale (i.e. the logit scale). This facilitates direct comparisons between the ability of a specific candidate and the difficulty of a specific question and allows us to estimate the probability of a correct response (or the probability of a specific score to be observed).

In line with Linacre (1994), Equation 8 illustrates the Rasch model in the case where a candidate *n* attempts to respond to question *i* which is scored on a scale from zero to *k*.

Equation 8

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_k$$

where

P_{nik} is the probability of candidate n being assigned on question i, the score k,

$P_{ni(k-1)}$ is the probability of candidate n being assigned on question i, the score $k-1$,

B_n is the ability of the candidate n,

D_i is the difficulty of question i,

F_k is the difficulty of being awarded score k instead of score $k-1$.

Each distinct score on a test question (marked from 0 to *k* points*)* is sometimes called a 'step'. According to Equation 8 the interaction between a candidate's ability, a question's difficulty and a step's difficulty define the probability of each score to be observed. Equation 8 corresponds to the Rating Scale model (Andrich, 1978; Wright and Masters, 1982) where all the questions on the test employ the same scale and it is assumed that the scale maintains the same meaning across the questions (e.g. a score of

5 has the same meaning for all questions). However, in the context of this study, different questions have their own rating scale and the equation may be re-written as:

Equation 9

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_{ik}$$

where F_{ik} indicates the difficulty of score k in relation to score $k-1$ for question *i*. This is the Partial Credit model (Wright and Masters, 1982).

Equation 9 employs only two facets: the candidates and the questions. However, the MFRM (Linacre, 1994) is just a natural generalisation that takes into consideration other facets that 'intervene' in the measurement mechanism. For example, a four-faceted approach may calibrate the severity of raters and the rater team effect which would be very convenient in the context of this study. It is usually hoped that the MFRM approach facilitates the removal of the influences of the various facets on the score obtained by the candidate (Linacre, 1994), thus providing a 'corrected' ability estimate. A four-faceted Partial Credit Rasch model involving the candidates, the questions, the raters and the rater teams may be presented in:

Equation 10

$$\log\left(\frac{P_{nijk}}{P_{nij(k1)}}\right) = B_n - D_i - C_j - T_m - F_{ik}$$

where

C_j is the severity of marker j,

T_m is the effect of being marked by a marker j, who is in team *m*.

Equation 10 models the candidates, the questions, the raters and the team effect, and could potentially be considered as the basis for the IRT analyses in this report, but we will also present simplified models for comparison.

5.2 Assumptions of the model

Since the MFRM is part of the large family of Rasch models, it makes a number of assumptions regarding the measurement context. First of all, the MFRM assumes that all the questions of the instrument of measurement (i.e. the test) tap onto a single latent dimension (i.e. measurement scale). This is the generic assumption of

'unidimensionality' and has been described in great detail by a number of researchers , including McDonald (1981), Lord (1980), Traub (1983), Dorans and Kingston (1985), and Stout (1987). Camilli, Wang and Fesq (1995) defined a test's dimensionality as the "number of latent variables that account for the correlation among item responses in a particular data set" (p.80). Thus, the assumption of unidimensionality demands that only one single latent trait or ability θ is measured by a test (Lord, 1980). However, because in real life situations the assumption of strict unidimensionality is difficult to achieve, it is possible that across a single test one prominent common factor (hopefully the one targeted by the test constructors) and some minor specific factors ('noise') may coexist. In this view, many researchers have repeatedly supported the idea that examinees' responses to test questions are multiply determined (Humphreys, 1981; Traub, 1983; Yen, 1985). For this purpose, Stout (1990) described a situation of 'essential unidimensionality' where the resulting data may be considered to be unidimensional for any practical intent and purpose.

Another main assumption of the Rasch model is that of local independence. Local independence demands that when the abilities of persons are constant throughout the test, their responses to any pair of items are statistically independent (see Jannarone, 1986). Essentially, this means that the response of a person on an item, or the experience of actually attempting an item, does not affect the interaction of that person with the next item (i.e. the ability of the person remains constant across the test).

Other, less important, or at least less frequently discussed, assumptions are that the test is not timed and that there is no guessing, copying, cheating and other practices that interfere with the measurement process. In contrast to other IRT models, it is assumed that for the Rasch model all items have the same discrimination (Lord, 1980). This is a strong assumption that has attracted a fair amount of criticism. However, in common with other assumptions of the Rasch model, there are statistical tests that can be used to investigate the degree for which it holds for a specific dataset.

In accordance with, and as a result of, the above assumptions, it is further assumed that there will be no interaction effects at play, that is:

- Item difficulty order is the same for all candidates
- Rater severity order is the same for all candidates
- Rater severity order is the same for all items.

5.3 MFRM Analyses

The MFRM analysis was carried out using Facets (Linacre, 2011), a standard software package for this kind of analysis. The datasets were formatted according to the requirements of the software and command control files set up appropriately.

Evaluation of the MFRM model-data fit was conducted through inspection of the Infit and Outfit Mean Squares (Wright and Stone, 1979), which identify general aberrance rather than any specific type of misfit. This is often an advantage because a fit statistic that focuses only on a specific type of aberrance may not have enough power to identify other types of aberrance (see Klauer, 1995; Lamprianou, 2010). However, had we been specifically focused upon a particular type of misfit, we might have decided to use a different fit statistic; for example, Karabatsos (2003) describes almost 40 different fit statistics in the context of Rasch models.

Use of Infit and Outfit Mean Squares, both of which are approximately chi–square distributed (Wright and Mok, 2000), is encouraged by a large body of literature (e.g., Smith, 1991, 2000; Wright and Masters, 1982; Wright and Mok, 2000). However, no universal cut-off values have been agreed. Karabatsos (2000) suggested that the distributional properties of the two statistics could differ significantly across datasets with different characteristics (e.g. when the distribution of item difficulties differs). In the past, different researchers have used slightly different ranges of acceptability for these fit statistics. For example, some (e.g., Engelhard, 1992, 1994; Lunz, Wright and Linacre, 1990) have set the range of acceptance for examinee and rater fit at 0.6 to 1.5. The range of acceptance for question fit is often set at 0.7 to 1.3; however, these are just rules of thumb and may not be appropriate for all datasets in every context, as Karabatsos (2000) notes.

Assessing model-data fit is very important, because in the case where fit is unacceptable the properties of invariance of the Rasch model do not hold, and it is risky then to assume that model estimates are correct in the context of a sparse data matrix. Although the effects of person misfit on the quality of measurement of the other facets in the model has not been investigated thoroughly, there is some research that shows that the inclusion of misfitting response patterns of persons in a Partial Credit Rasch analysis can affect the estimates of the other facets as well. For example, Curtis (2004) used empirical and simulated data to investigate the effect of removing the misfitting response patterns of persons from a Partial Credit Rasch analysis. Curtis stated that,

> The analyses undertaken on two real data sets in this study have shown that the inclusion of responses that underfit the Rasch measurement model, and that may reflect carelessness in responding, increase the standard errors of item estimates, reduce the range of item locations on the scale, and reduce the inter-threshold range within items. Thus, the inclusion of misfitting cases compromises the measurement properties of the scale formed by the instrument … Together, these findings suggest that it is important to examine person fit as well as item fit in the analysis of data sets and to remove, at least

> for the purposes of calibration, those cases as well as items, that reveal substantial misfit. (Curtis, 2004, p. 141)

Having said that, it is still not clear how persons' misfitting response patterns may affect the quality of rater measurement. An issue that arose when attempting to analyse the four-facet data in this project was that the raters were nested within teams, with the consequence that the datasets consisted of several disconnected subsets. If the data can be separated into non-overlapping blocks when grouped according to one or more variables, then the data are said to be disconnected. In practice, even though some linking might exist, it could be insufficient to establish a common scale across the blocks. Although candidates' work could have been marked by examiners from different teams, Facets interpreted teams as being disconnected subsets.

According to the Facets Manual (Linacre, 2011)

> The only effect of "disconnection" is on the relationship between measures in different subsets. Under these circumstances, **Facets reports one of the infinite number of different possible solutions, all equally likely.** Only measures in the same subset are directly comparable. A separate set of vertical rulers is produced for each disjoint subset. (Linacre, 2011, p.299).

According to this interpretation, it is not straightforward to use the four-facets model results to investigate rater effects or to compare rater severities. We therefore dropped the team facet and ran a three-facets model (in effect, we dropped the T_m effect from the model of Equation 10), which does not 'suffer' from the effect of the subset 'disconnection'.

To avoid the problem of disconnected subsets, it was also important to run the analyses separately for every year group because:

(a) there are no common test questions between the tests from different years,

(b) even in the case where the same candidate took a test in two consecutive years, it would not be defensible to assume that that person retained the same knowledge, motivation or test-taking behaviour across the period (i.e. we cannot treat this person as being 'the same person' across tests), and

(c) we cannot assume that the raters retain the same rating behaviour across years because this is one of our research questions.

It was confirmed that our data design was consistent with types of design which were shown to be appropriate for this type of rater analysis (Linacre, 1997; Lunz et al., 1990).

To summarise,

(a) We analysed the datasets from different years separately

(b) We ran a four-facets model to measure the differences between the teams of raters, and then dropped the team facet in order to run a simpler model with no disconnected subsets in order to focus on the rater statistics.
(c) For each of the analyses, we evaluated the model-data fit using the criteria mentioned above in order to assess the validity of the results.

The data from both geography and psychology proved problematic for MFRM analysis, but geography more so than psychology. For this reason we begin by presenting analysis results for psychology.

5.4 Rasch findings – psychology

For the purposes of estimating the impact on individual candidates' scores of any differential marker severity, 'operational' datasets were generated. For each of the three years, subsets were generated which contained only the final scores awarded to candidates. Therefore, no duplicate records, backread records or seeded clip records were initially included in the datasets. A repeat analysis that incorporated the seeded clip records to enhance the degree of marker connectivity in the dataset produced similar findings. Also, clerically-marked and computer-marked part-questions were excluded from the analyses because these would not provide information relevant to differential rating severity. For psychology, the 'operational' datasets had the characteristics recorded in Table 18. In effect, in each dataset there was only one score per part-question per candidate, and this was the mark that was used for the calculation of the final total test score that was used for later grade classification. Each of the three datasets for each subject was analysed separately. A three-facets model was estimated using candidates, items and examiners as facets.

Table 18 Operational subsets for the psychology data in the Rasch analyses

Year	*No. candidates*	*No. items*	*No. examiners*	*No. records in operational subset*
2009	5448	16	20	5448*16 = 87168
2010	7149	16	26	7149*16 = 114384
2011	7679	14	28	7679*14 = 107506

For all three datasets item fit was satisfactory, with the Infit Mean Square being within the rule-of-thumb boundaries of 0.7 to 1.3; only in one case was the Infit Mean Square marginally beyond the limit of 1.3. The Outfit Mean Square for items was (as expected) slightly larger, but again within the limits for all but one item. The separation index of the items was around 50 for all three psychology datasets. In each case, the distribution of item difficulties matched the distribution of candidate abilities, indicating a well-targeted assessment instrument, maximising the psychometric precision of the ability measures.

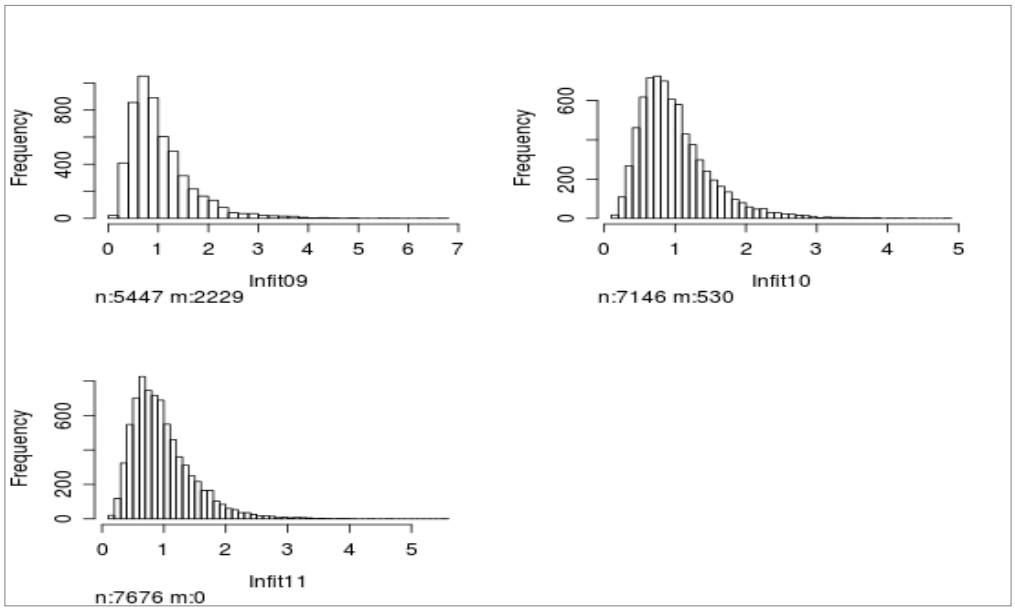
Rater fit was also found to be satisfactory, with Infit Mean Square statistics being within the range 0.7 to 1.3, with just one exception for the 2009 dataset. The standard errors of rater measures were very small and a satisfactory separation index of around 7 to 8 was estimated for all three datasets.

Candidate fit is recorded in Figure 7. For all three datasets, the average Infit Mean Square was very small (0.98 to 1.05) but the standard deviations were large, indicating that a non-negligible number of candidates had large Infit Mean Square statistics. Overall, 7.8% of the 2009 candidates, 4.6% of the 2010 candidates and 4.2% of the 2011 candidates had Infit Mean Squares larger than 2. Such large values are outside the usual acceptable boundaries for the Rasch model, indicating a degree of misfit for the candidates concerned. However, since the aim of our study is to investigate the quality of measurement for the raters, the high fit statistics for a percentage of candidates is not alarming in itself. It may, however, have implications for the estimates of other effects, including rater effects, in the model.

One of the aims of our study was to investigate the effect of any differential rater severity on candidates' marks. One index of the rater effect is the distance between the Rasch measures of the most and the least lenient raters (i.e. the range of Rasch measures for the raters). This measure can also be expressed in terms of the standard deviation of the candidate ability distribution. Another measure of rater effect is the ratio between the standard deviation of the rater measures and the standard deviation of the candidate measures. The results for the two indices for each of the three psychology datasets are shown italicised in Table 19.

According to Table 19, the effect for candidates of having their script (assuming their whole script was marked by a single rater) marked by the most lenient rather than the most severe rater can be extremely large (a range of 0.84 to 1.26 logits). In other words, this effect may be the equivalent of more than one standard deviation of the candidate ability distribution, suggesting detrimental effects on reported test scores. However, this is the worst-case scenario for the potential effect on *individuals*, where a script is marked by the most lenient rather than the harshest rater. A more conservative scenario of the *overall* rater effect is the ratio of the standard deviation of rater measures over the standard deviation of candidate measures. This index, again, is very large and ranges from 0.22 to 0.26 logits.

Figure 7Histograms - Infit Mean Square fit for psychology candidates (all 3 years)



However, the two indices just described, i.e. range and standard deviation of rater measures divided by the standard deviation of candidate measures, do not represent reality, because in practice different raters marked the responses of particular candidates to different part-questions. Therefore, it is possible to assume that because of randomisation in the allocation of clips to raters, the overall effect of raters on individual candidate test scores could be very small, because the chances are that some of the part-questions responded to by a candidate were marked by less severe and others by more severe raters. To investigate this possibility, we can use a third index, which is computed by subtracting candidates' 'fair scores' from their 'observed scores'. This gives an indication of the magnitude of the Rasch model 'correction', estimated for each candidate overall. In effect, it is the net Rasch estimated impact of the marking process on each candidate. Since all candidates in any year attempted the same questions, the only differential measurement effect is that of the differential severity of the raters.

Table 19 Indices of rater effects for the three psychology datasets

Measures and Indices (in logits)	*2009*	*2010*	*2011*
Range of rater measures	0.93	0.87	0.58
SD of rater measures	0.19	0.18	0.15
SD of candidate measures	0.74	0.74	0.69
Range of rater measures over SD of candidate measures	*1.26*	*1.18*	*0.84*
SD of rater measures over SD of candidate measures	*0.26*	*0.24*	*0.22*
Rater separation* (="true" SD/Root Mean Square Error)	10.11	7.57	7.18
Candidate separation* (="true" SD / Root Mean Square Error)	2.47	2.46	2.05
Candidate reliability (similar to Cronbach's alpha)	0.86	0.86	0.81

** a measure of the spread of the estimates relative to their precision*

Turning to the consistency of the rater measures, it was found that raters were partly consistent in the severity rank ordering. For example, nine raters marked in both 2009 and 2010. The correlation between their measures was zero; however, this is because one data point which was beyond the main pattern of the data (an outlier) had a large impact (see Figure 9).

Figure 8 shows the distribution of the third index mentioned in the previous paragraph, which is computed by subtracting candidates' 'fair scores' from their 'observed scores'. The effect of the raters seems to be substantially bigger (measured in raw scores equivalents) for the 2009 psychology dataset, although it is not negligible for the other years. The effect is the smallest for the 2011 dataset.

According to the Rater Separation index (see Table 19), the raters demonstrated substantially different severity. The spread of the rater estimates is wide compared to the standard error of their estimates. Candidate Separation is much smaller, suggesting that the psychology tests did not differentiate significantly between candidates.

The mean absolute rater effect (in units of raw score, i.e. marks) was 6.0 for year 2010 (SD=1.9, N=5448), 5.8 for year 2010 (SD=1.3, N=7149) and 1.5 for year 2011 (SD=1.0, N=7679). We use the absolute effect because it is immaterial whether the score of a specific candidate was positively or negatively affected (some were positively affected and some negatively affected). The absolute effect is relatively large and may be presented as a percentage of the range of the actually observed scores for each dataset.

For example, for the 2009 dataset, the mean absolute rater effect was 6 marks over a range of 61 marks (minimum observed score was 0 and maximum observed score was 61). This corresponds to 9.9% of the observed mark range. The equivalent percentage for the 2010 dataset is 10.0% and for the 2011 dataset is 2.9%.

Figure 8Distribution of rater effects for the three psychology datasets

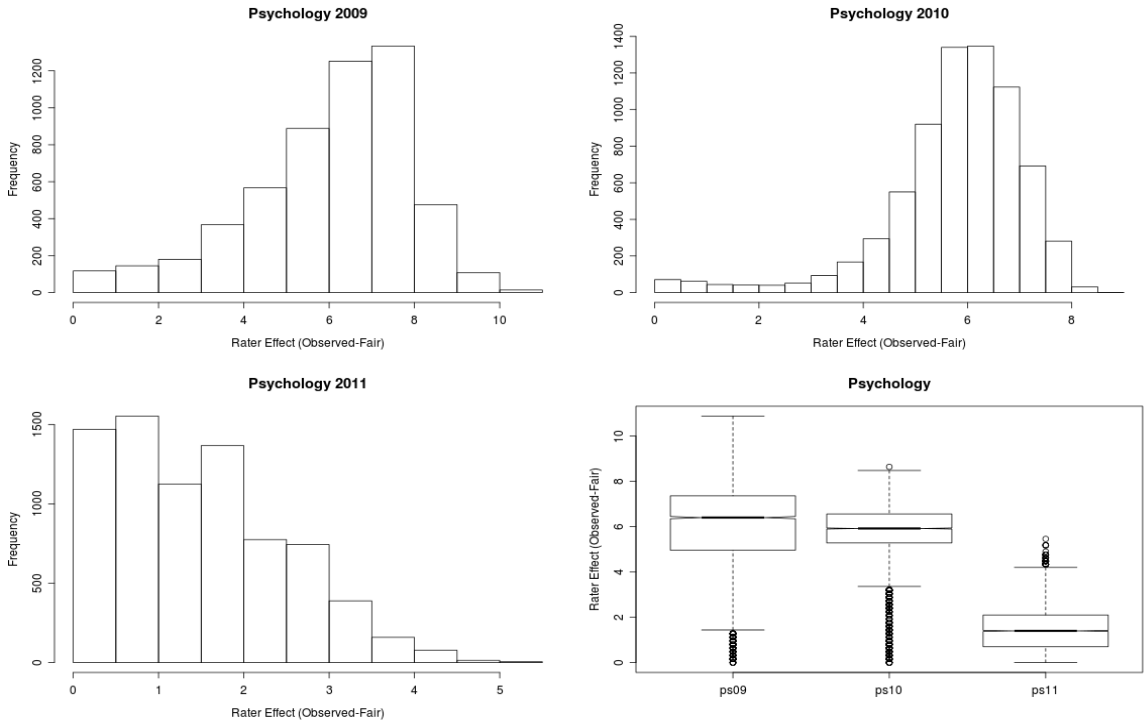
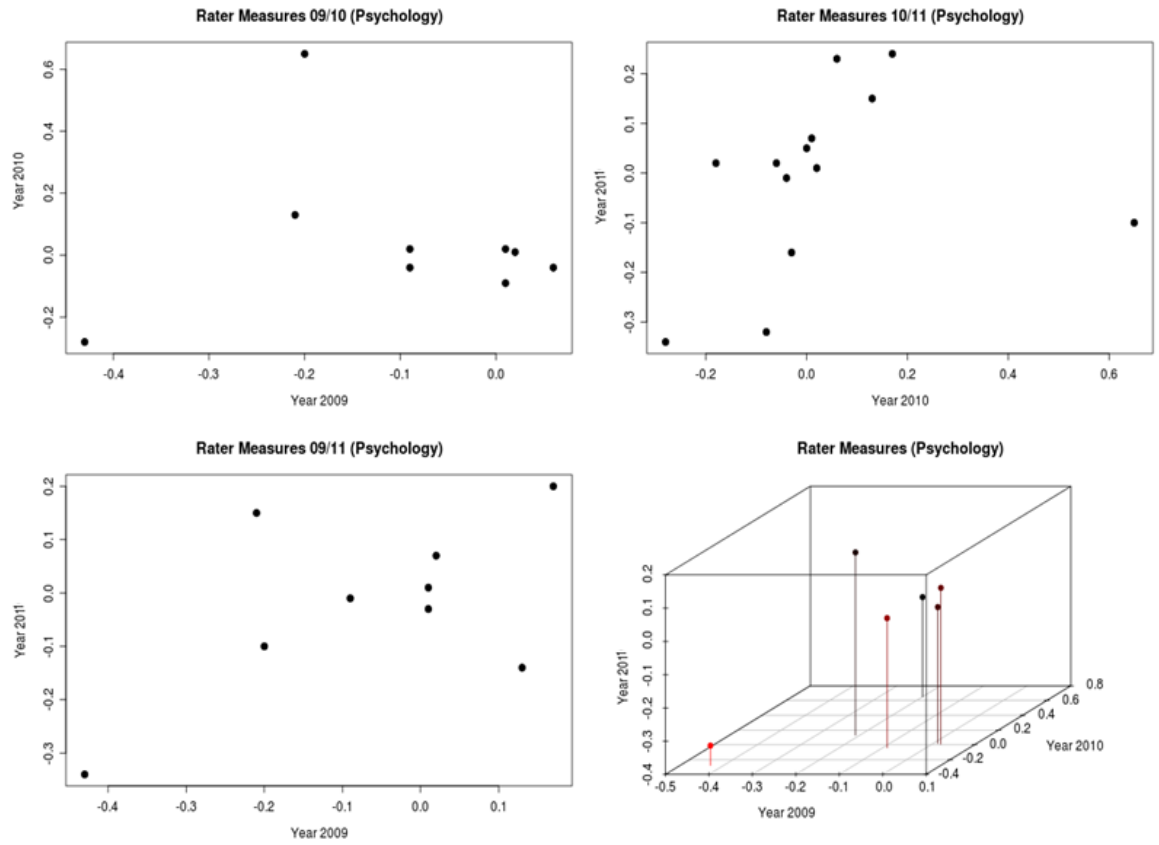


Table 20 records the correlation between the rater measures for different combinations of years. These correlations are non-negligible if one bivariate outlier is removed from two of the datasets. With such low numbers of examiners common between years, these findings cannot be taken as general indications about the stability of rater effects.

Table 20 Correlation between rater measures for different years

	Number of raters	*Correlation between measures*
2009 vs 2010	9	0.00 (0.69 if one outlying point is removed)
2010 vs 2011	13	0.28 (0.77 if one outlying point is removed)
2009 vs 2011	9	0.57

Figure 9Correlation of rater measures across years (psychology data)



5.5 Rasch findings – geography

Following the procedure for psychology, 'operational' datasets for each of the three years were generated for the MFRM analyses of the geography data, by excluding duplicate records, backread records and seeded clip records. In each dataset there was only one score per part-question per candidate. Each of the three datasets for each subject was analysed separately. A three-facets model was estimated using candidates, items and examiners as facets.

A complication in the datasets for geography that was not present in the datasets for psychology is that candidates in geography had some question choice whereas in psychology all candidates in any year took the same set of questions. As far as candidates were concerned, therefore, the response data matrix was particularly sparse in terms of both markers and items (each candidate had observations for just six of the 12 items in each paper, and for up to six only of the 30+ operational markers). Not surprisingly, perhaps, the MFRM analyses for geography revealed a higher degree of candidate misfit than did the analyses discussed earlier in this chapter for psychology: 9.5% for 2009, 10.6% for 2010 and 9.7% for 2011. Even though model fit was good for

items and raters, this substantial candidate misfit throws doubt on the validity of the MFRM results. Nevertheless, for the record we present in Table 21 the computed indices of rater effects for the geography data.

Table 21 Indices of rater effects for the three geography datasets

Measures and Indices (in logits) GEOG	*2009*	*2010*	*2011*
Range of rater measures	0.90	0.91	0.88
SD of rater measures	0.19	0.22	0.20
SD of candidate measures	0.50	0.48	0.55
Range of rater measures over SD of candidate measures	1.80	1.90	1.60
SD of Rater Measures over SD of candidate measures	0.38	0.46	0.36
Rater separation* (="true" SD/Root Mean Square Error)	9.19	8.29	7.29
Candidate separation* (="true" SD / Root Mean Square Error)	1.63	1.54	1.69
Candidate reliability (similar to Cronbach's alpha)	0.73	0.70	0.74

** a measure of the spread of the estimates relative to their precision*

According to the Rater Separation index (see Table 21), the raters demonstrated substantially different severity, as the psychology raters had also done. The spread of rater estimates is too wide in comparison with the standard error of their estimates. Candidate Separation is much smaller, and even smaller than for psychology, suggesting that the geography papers differentiated less well among candidates than did the psychology papers. Yet the computed candidate reliability is markedly higher than that found in the generalizability analyses described in the previous chapter, which were constrained to treat clip scores as free of marker error.

6 Multilevel modelling

Multilevel modelling (Goldstein, 2010; Snijders and Bosker, 2011) takes into account the complex structure of data by explicitly modelling hierarchical and, in this case, non-hierarchical, cross-classified relationships. This set of techniques is also known as hierarchical linear modelling (Raudenbush and Bryk, 2002), random effects and mixed-effects models. Here, for example, we have individual scores awarded by examiners to scanned parts of candidates' work (clips). The clips analysed here are double-marked (backread data), and sometimes multiply marked (seeded clip data), which leads to a cross-classified data structure. Marks at level 1 in the data hierarchy are nested within examiners at level 2 in the data structure, and can simultaneously be viewed as nested within clips, also conceptually at level 2 of the data structure. More formally, marks at level 1 are said to be nested within the cells of an examiner-by-clip two-way cross-classification.

Multilevel models can be seen as an extension to the linear regression model. In a standard regression model, we are interested in the relationships between the independent (predictor) variables and the dependent variable. The residual variance is treated as error in a regression and is not analysed, but there are interesting and useful effects to be gained by such analyses. Therefore, in multilevel modelling, the residual variance is partitioned into separate variance component parts, associated with the different levels in the data structure. This is known as the random part of the model and the variance of units at different levels is presented, although the software will also allow us to estimate an effect for each unit at a given level. So, in our data, it will be possible to estimate the precision with which individual clips were marked.

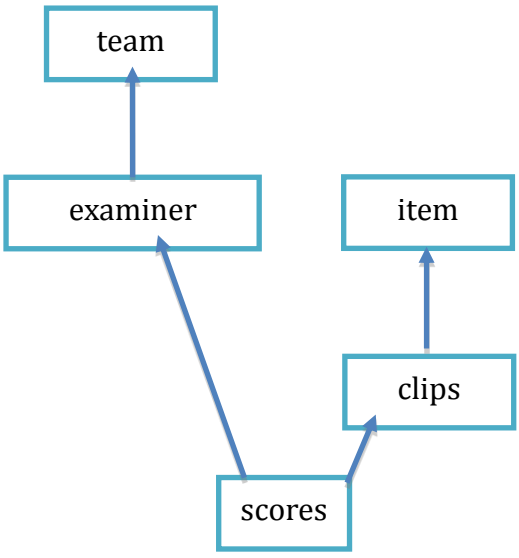
Prior to the introduction of multilevel models, researchers had few options about the analysis of complex structured data. One approach was to analyse aggregated data, but it is now widely understood that inferences reached from aggregated data cannot be generalised to the lower levels of analysis. Robinson (1950) showed that the higher the percentage of black people in a neighbourhood in the US, the higher the proportion of people with extreme right-wing views. However, one should not infer that black people had more extreme views than white people. There is not a statistical connection between the individual's characteristic (being black) and political views. Instead, this is related to the neighbourhoods' characteristics – the aggregate level. This is known as the ecological fallacy. With multilevel modelling, relationships at lower levels and aggregate levels can be analysed simultaneously.

6.1 Structure of the data for multilevel analyses

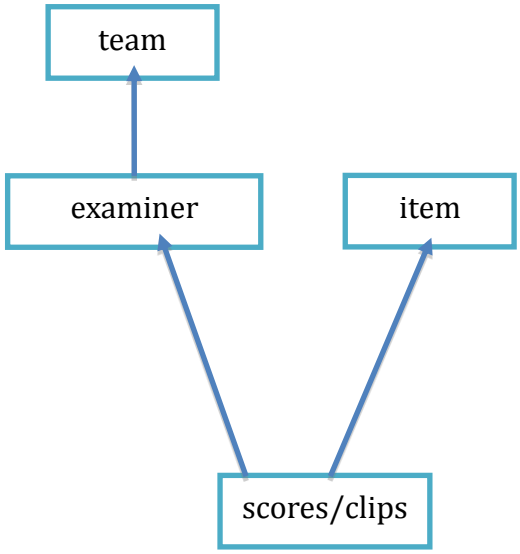
Figure 10 shows the structure of the data using a classification diagram (Browne, Goldstein and Rasbash, 2001). Each 'node' in the diagram represents a classification in the model. An arrow between two nodes indicates a nested relationship, while an absence of an arrow indicates a crossed relationship. The data are slightly different for each monitoring system. We will take the seeded script system first (Figure 10a), as it is more complex. At the lowest level is the score given to part of a candidate's work. These scanned parts of a candidate's script, called 'clips', are seeded and multiply marked in the seeded clip monitoring system. As such, the clips are cross-classified with examiners. Clips are nested within items (or 'part-questions') and examiners also multiply mark items, so examiners are also cross-classified with item. Finally, examiners are nested within teams. Candidates are not modelled in these analyses because we did not have the whole work of a candidate – clips were each just a part of their written work.

Figure 10 Classification structure

(a) Seeded monitoring system



(b) Backread monitoring system



To be more precise, in the geography data analysis, raters were nested within teams, as there were 15 teams.[2] However, multilevel modelling assumes that the units at each level are randomly sampled from a population, and it is generally expected that there should be at least 20 units at each level. Thus, 15 teams is really too low a number for this analysis. In the psychology data there were only 12 teams across the three years.

[2] Examiners were members of multiple teams when they marked in more than one year.

Given the small variation due to team effects found in the G-theory analyses reported in Chapter 4, this is not a big loss to the current research.

Turning to the backread monitoring system, we can see that the data structure is slightly simpler (Figure 10b). Here, clips are not generally multiply marked because there will tend to be a single supervisor check on these scores. This means that clips and scores are confounded in these data and it is not possible to separately identify their variance components.

In the multilevel analyses of the psychology datasets, only data relating to clips that had been multiply-marked by examiners was included, and clerically-marked or machine-marked data were removed. Therefore, all of the scores that had single marks were excluded, as those data do not contain information about the accuracy of the marks in an MLM context. Of 52 examiners in the dataset, only 23 marked in all three years. Nine examiners marked in two years and 20 marked in only one year.

6.2 Multilevel analyses

Data were modelled using the MLwiN software (Rasbash, Steele, Browne and Goldstein, 2009; Browne, 2009). A number of exploratory models were investigated, but for simplicity one model is presented and the findings of other models are referred to where appropriate. Equation 11 follows the classification notation of Browne et al (2001), which avoids the proliferation of subscripts that arises in models with many classifications. The response variable y_i is the i-th score.

The model includes fixed effects for the intercept and separate binary indicators for whether the clip was from the seeded clip monitoring system and the year in which the clip was marked. There is also a fixed effect for the mark that the candidate's work was awarded. This is deemed 'correct' and will be the supervisor's mark in the backread monitoring system and the principal examiner's mark in the seeded clip monitoring system.

There are five classifications in the random part of the model: score, clip, item, rater and training team, which are numbered from one to five. Hence, the '(2)', '(3)', '(4)', and '(5)' superscripts and subscripts identify random effects that are associated with each of the classifications above the absolute score difference level where the '(1)' superscripts and subscripts are implicit for convenience of notation.

Equation 11

$$y_i = \beta_1 seeded_i + \beta_2 backread_i + \beta_3 year10_i + \beta_4 year11_i + \beta_5 imark * backread * 2009_i + \beta_6 imark * backread * 2010_i + \beta_7 imark * backread * 2011_i + \beta_8 imark * seeded * 2009_i + \beta_9 imark * seeded * 2010_i + \beta_{10} imark * seeded * 2011_i$$

$$+ \left(u^{(5)}_{0\text{training team}(i)} + u^{(4)}_{0\text{rater}(i)} + u^{(3)}_{0\text{item}(i)} + u^{(2)}_{0\text{clip}(i)} + e_{0i}\right) seeded_i$$

$$+ \left(u^{(5)}_{1\text{training team}(i)} + u^{(4)}_{1\text{rater}(i)} + u^{(3)}_{1\text{item}(i)} + e_{0i}\right) backread_i$$

$$training\ team(i) \in \left(1, \dots, \mathrm{J}^{(5)}\right), \qquad rater(i) \in \left(1, \dots, \mathrm{J}^{(4)}\right)$$
$$item(i) \in \left(1, \dots, \mathrm{J}^{(3)}\right), \qquad clip(i) \in \left(1, \dots, \mathrm{J}^{(2)}\right)$$
$$i = 1, \dots, \mathrm{N}$$

$$\begin{pmatrix} u^{(5)}_{0\text{training team}(i)} \\ u^{(5)}_{1\text{training team}(i)} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(5)} & \\ 0 & \sigma^2_{u1(5)} \end{pmatrix} \right\}$$

$$\begin{pmatrix} u^{(4)}_{0\text{rater}(i)} \\ u^{(4)}_{1\text{rater}(i)} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(4)} & \\ \sigma^2_{u01(4)} & \sigma^2_{u1(4)} \end{pmatrix} \right\}$$

$$\begin{pmatrix} u^{(3)}_{0\text{item}(i)} \\ u^{(3)}_{1\text{item}(i)} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(3)} & \\ 0 & \sigma^2_{u1(3)} \end{pmatrix} \right\}$$

$$u^{(2)}_{0clip(i)} \sim \mathrm{N}\left(0, \sigma^2_{u0(2)}\right)$$

$$\begin{pmatrix} e^{(1)}_{0(i)} \\ e^{(1)}_{1(i)} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0(1)} & \\ 0 & \sigma^2_{u1(1)} \end{pmatrix} \right\}$$

The classification function 'rater(i)' denotes the rater associated with the i-th score. Raters are indexed from 1 to $\mathrm{J}^{(4)}$ and $\mathrm{u}^{(4)}_{0\text{rater(i)}}$ and $\mathrm{u}^{(4)}_{1\text{rater(i)}}$ are the different effects that the rater has on the i-th score depending on which monitoring system was being used. The covariance at level 4, $\sigma 2_{u01(4)}$, shows the extent to which examiner effects were affected by the monitoring system. That is, a positive covariance would indicate higher rater effects in the seeded clip system and vice versa. For the sake of model simplicity, covariance is presented only for the most interesting level in this research – the raters.

The classification functions and random effects for the other classifications are similarly defined. For the backread system, there is no clip effect as each clip is scored by only one rater in that system. The clip effect is therefore confounded with the level 1 residual. All random effects are assumed normally distributed, independent across classifications and independent of any predictor variables included in the model.

6.3 Multilevel modelling findings for geography

Table 22 shows the multilevel model for the geography backread and seeded data for all three years included in the study. We will interpret the fixed part of the model first. A dummy variable was entered, representing the seeded clip data for 2009. It shows that, with the other effects in the model taken into account, examiners were on average 2.188 marks more lenient (χ^2_1=80.84, p<0.001) than the mark given to candidates (imarks). For 2009 backread data, examiners were on average 0.035 of a mark more lenient than the 'correct mark', but this average was not significantly different from the correct mark (χ^2_1=1.37, p=0.24). We can see that the seeded clips were, on average, marked slightly more leniently than the backread scripts (2.188-0.035=2.153) and this difference was statistically significant (χ^2_1=156.78, p<0.001).

Two further dummy variables show the average score for 2010 (-0.114) and for 2011 (0.062) compared with the average score for 2009. Scoring was significantly more severe on average in 2010 compared with 2009 (χ^2_1=7.58, p=0.006), but there was not a significant difference between 2009 and 2011 (χ^2_1=2.09, p=0.15). It can be seen that these average differences were very small, even where they were statistically significant.

The variable for the 'correct mark' given to candidates (imarks) was interacted with year and with monitoring type to produce six variables. As would be expected, the relationship between marks awarded by examiners and the correct score was highly significant under the seeded and backread systems (a joint test summarised the effect of the six variables: χ^2_6=580406.06, p<0.001). Marking under the backread system was significantly more associated with the correct scores than under the seeded system (χ^2_1=145.90, p<0.001). This is most likely to be accounted for by the fact that the supervisor can see the original examiners' marks in the backread system. There were no significant differences between years for the seeded data (a joint test summarised the three contrasts: χ^2_3=2.15, p=0.14). For the backread data, there was a very high relationship between correct score and score awarded by markers (0.989, 1.003 and 0.985 in the three years respectively). This was significantly higher in 2010 than in

2009 (χ^2_1=21.59, p<0.001) or 2011 (χ^2_1=32.12, p<0.001), but the difference between 2009 and 2011 was not significant (χ^2_1=1.49, p=0.222).

Table 22 Multilevel model for geography multiply-marked data (2009-11)

Parameter	Estimate	Standard error
Fixed part		
Seeded scripts 2009 ($\beta_1 seeded_i$)	2.188	0.243
Backread scripts 2009 ($\beta_2 backread_i$)	0.035	0.030
2010 ($\beta_3 year10_i$)	-0.114	0.041
2011 ($\beta_4 year11_i$)	0.062	0.043
imark 2009 seeded ($\beta_8 imark * seeded * 2009_i$)	0.677	0.028
imark 2010 seeded ($\beta_9 imark * seeded * 2010_i$)	0.719	0.035
imark 2011 seeded ($\beta_{10} imark * seeded * 2011_i$)	0.674	0.031
imark 2009 backread ($\beta_5 imark * backread * 2009_i$)	0.989	0.002
imark 2010 backread ($\beta_6 imark * backread * 2010_i$)	1.003	0.002
imark 2011 backread ($\beta_7 imark * backread * 2011_i$)	0.985	0.002
Random part		
Seeded: team variance ($\sigma^2_{u0(5)}$)	0.019	0.013
Backread: team variance ($\sigma^2_{u1(5)}$)	0.003	0.001
Seeded: rater variance ($\sigma^2_{u0(4)}$)	0.195	0.042
Seeded and backread: rater covariance ($\sigma^2_{u01(4)}$)	0.013	0.005
Backread: rater variance ($\sigma^2_{u1(4)}$)	0.003	0.001
Seeded: item variance ($\sigma^2_{u0(3)}$)	0.365	0.132
Backread: item variance ($\sigma^2_{u1(3)}$)	0.001	<0.001
Seeded: clip variance ($\sigma^2_{u0(2)}$)	0.769	0.074
Seeded: residual variance ($\sigma^2_{u0(1)}$)	1.545	0.025
Backread: residual variance ($\sigma^2_{u1(1)}$)	0.214	0.002

Turning to the random part of the model, we can see that most of the variation is at the lowest level. This is the standard pattern and simply means that the errors were largely due to idiosyncratic marker responses to individual parts of candidates' performances. In fact, for the backread data, nearly all of the variance is at the lowest level.

Working upwards, we see that clips (0.769) varied significantly in terms of how accurately they were marked (χ^2_1=109.20, p<0.001). In other words, some were more difficult to mark than others in the seeded data, where the effect of clips could be estimated because they were multiply marked.

Although there were statistically significant effects of team (χ^2_1=3.104, p=0.039), examiner (χ^2_1=15.11, p<0.001) and item (χ^2_1=8.02, p<0.002) in the backread data, they were all small (0.003, 0.003 and 0.001 respectively). Team, examiner and item all had larger variances for the seeded data. The team effect was not significant (χ^2_1=2.13, p=0.07), but the examiner and item effects were significant (χ^2_1=21.05, p<0.001 and χ^2_1=7.68, p<0.003 respectively).

There was significantly more variance for rater effects in the seeded data compared with the backread data (χ^2_1=20.50, p<0.001). Figure 11 shows the estimated rater effects and their confidence intervals. On average, raters had a small impact upon candidates' marks and few examiners had estimated effects that were significantly different from zero, as the confidence intervals cross the zero line in most cases. Note that the scale is very different on the two graphs and that the rater effects are larger for the seeded data, although still a maximum of about one mark on average.

Rater effects covaried (0.013) significantly when calculated for seeded and backread data (χ^2_1=8.41, p<0.002). When the modelled covariance is converted to a correlation, it shows a moderate relationship (0.54; see plot in Figure 12).

With so little spread in the rater effects for the backread data, a lack of correlation between the two sets of estimated rater effects is not necessarily cause for alarm. A further multilevel model was created in which rater effects were modelled separately in the random part of the model for 2009, 2010 and 2011. Correlations of rater effects between years were again found to be moderate (0.50 to 0.53).

Figure 11 Rater effects for the seeded (top) and backread data (bottom) – geography

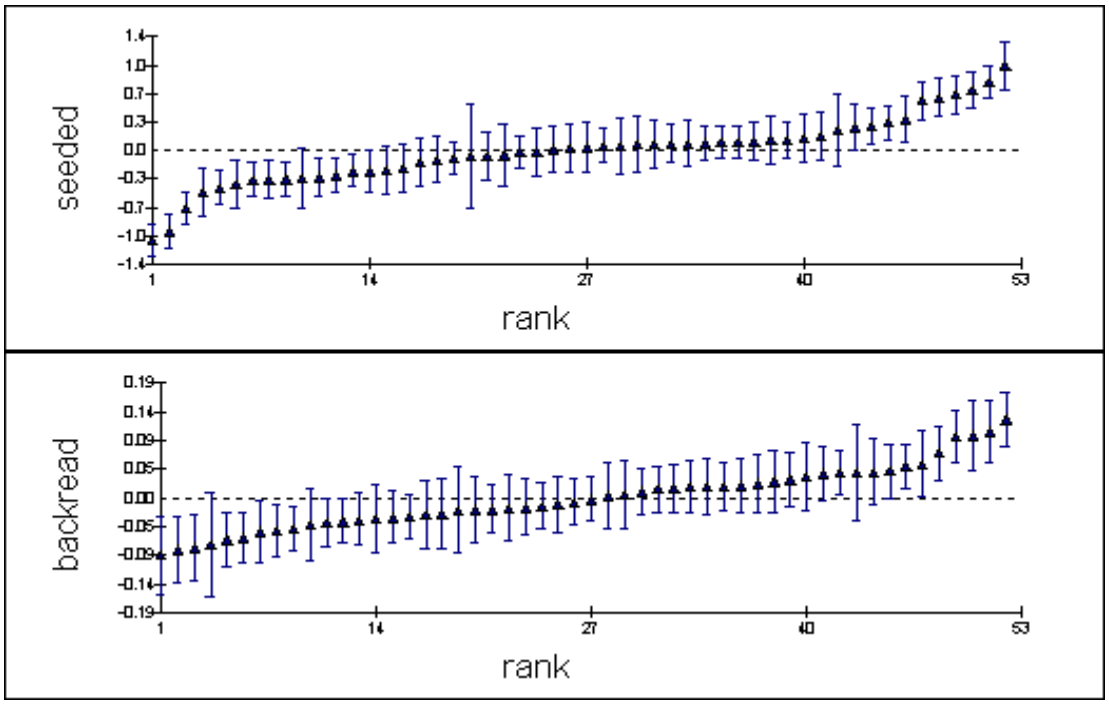
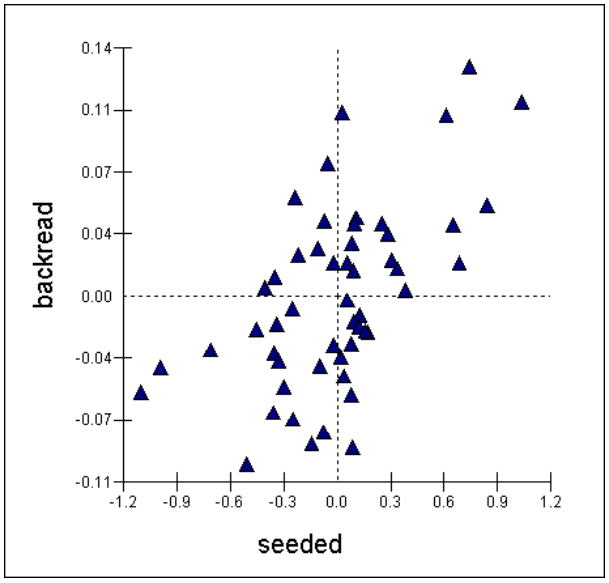


Figure 12 Plot of rater effects for each monitoring system - geography



6.4 Multilevel modelling findings for psychology

Table 23 shows the multilevel model for the psychology data, which was simpler than for geography due to issues in getting the model to converge. Specifically, in the random part of the model there is no team level and the item level is not split by monitoring system.

Table 23 Multilevel model for psychology multiply-marked data (2009-11)

Parameter	**Estimate**	**Standard error**
Fixed part		
Seeded scripts 2009 ($\beta_1 seeded_i$)	0.407	0.137
Backread scripts 2009 ($\beta_2 backread_i$)	0.416	0.156
2010 ($\beta_3 year10_i$)	-0.113	0.173
2011 ($\beta_4 year11_i$)	-0.138	0.242
imark 2009 seeded ($\beta_8 imark * seeded * 2009_i$)	0.707	0.021
imark 2010 seeded ($\beta_9 imark * seeded * 2010_i$)	0.740	0.035
imark 2011 seeded ($\beta_{10} imark * seeded * 2011_i$)	0.670	0.033
imark 2009 backread ($\beta_5 imark * backread * 2009_i$)	0.750	0.006
imark 2010 backread ($\beta_6 imark * backread * 2010_i$)	0.755	0.010
imark 2011 backread ($\beta_7 imark * backread * 2011_i$)	0.731	0.012
Random part		
Seeded: rater variance ($\sigma^2_{u0(4)}$)	0.038	0.042
Seeded and backread: rater covariance ($\sigma^2_{u01(4)}$)	0.011	0.004
Backread: rater variance ($\sigma^2_{u1(4)}$)	0.011	0.003
Item variance ($\sigma^2_{u0(3)}$)	0.480	0.117
Seeded: clip variance ($\sigma^2_{u0(2)}$)	0.303	0.189
Seeded: residual variance ($\sigma^2_{u0(1)}$)	0.606	0.160
Backread: residual variance ($\sigma^2_{u1(1)}$)	0.522	0.005

There was not a significant deviation from the correct mark on average for the backread (0.416, χ^2_1=0.431, p=0.512) or the seeded clip (0.407, χ^2_1=0.326, p=0.568) monitoring systems, and there were no average differences in scoring accuracy for the three years studied (-0.113 and -0.138 average differences for 2010 and 2011 respectively,

χ^2_2=0.463, p=0.793). Just as for the geography data discussed above, the relationship with the correct mark was strong in psychology (0.670-0.755), but this was lower than in geography. The relationship between the correct marks (imark) and examiner scores was not significantly different between the seeded and backread systems (χ^2_1=2.418, p=0.120), but there was a significant difference in this relationship between years (χ^2_3=11.439, p=0.010). From observation of the coefficients, it can be seen that the relationship with imark was highest for both systems in 2010.

As with geography, a lot of the residual variation is at level 1, but there were not significant differences in residual variation between the seeded and backread systems in the psychology analysis (χ^2_1=0.286, p=0.296). Residual variance associated with clips (0.303) was not significant (χ^2_1=2.574, p>0.05), but all other random effects were significant.

In the psychology model we find significant variation associated with questions (0.480), which we did not find for geography. In common with geography, there were significant differences between the rater effects calculated in the two monitoring systems (0.011 for the backread system and 0.038 for the seeded clip system; χ^2_1=10.182, p<0.001). Looking at the two sets of rater effects (Figure 13), we see that the rater effects are small in psychology (y-axis values) and that most of the examiners were not significantly different from zero, as the 95% error bars cross zero.

Across the two monitoring systems, the correlation of rater effects for the 52 examiners was low (0.54; see plot in Figure 14). A separate model was also created to model the covariance of rater effects across years, but the correlations were found to be very low (0.04-0.08). This is likely to be due to the small number of raters who marked across years and the outliers identified in the MFRM analysis (see Table 20).

Figure 13 Rater effects for the seeded (top) and backread data (bottom) - psychology

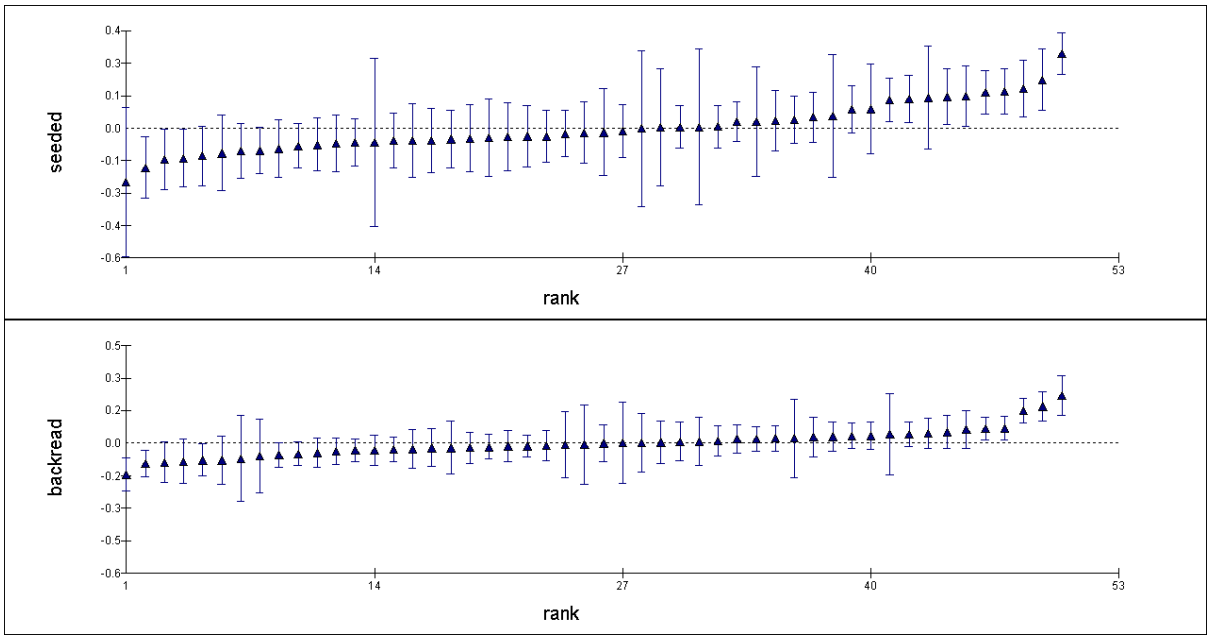
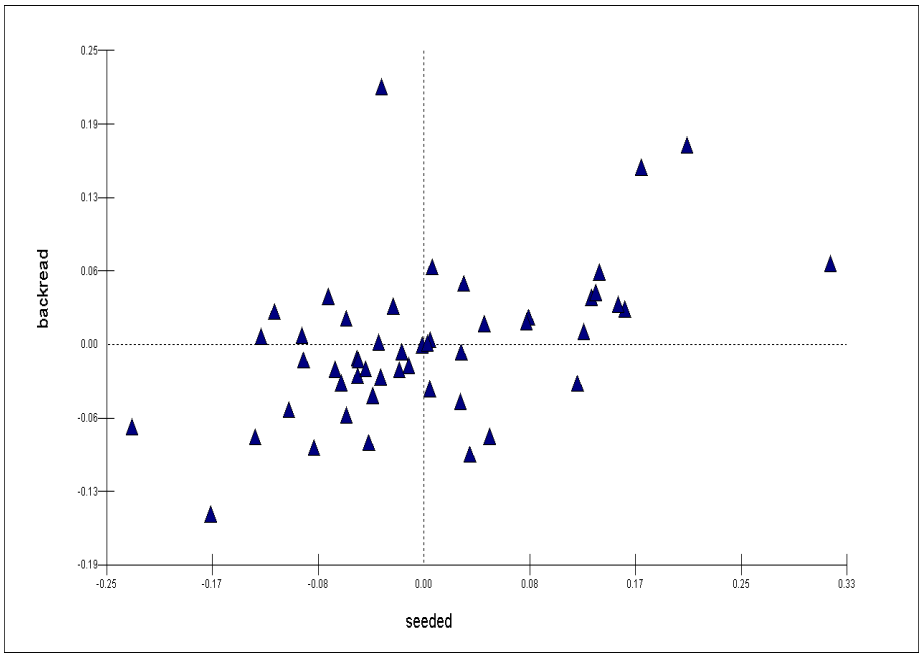


Figure 14 Plot of rater effects for each monitoring system - psychology



7 Reflections and conclusions

7.1 The impact of inter-rater reliability

Our analyses would lead to different conclusions about the impact of inter-rater reliability upon candidates' scores. The G-theory and multilevel models implied that the average difference between raters would be around one or two marks for clips. In other words, inter-rater reliability was a comparatively small effect. However, the MFRM analyses, which estimated the effects at the level of the whole question paper, implied that for psychology the average effect of raters was approximately 6 marks for the 2009 and 2010 data and 1.5 marks for the 2011 data. Without whole scripts being scanned and included as part of the monitoring data, it was not possible to conduct question paper level analyses using these other approaches. Indeed, it is possible that the MFRM results could be affected by the lack of whole scripts being multiply-marked, as the data were disjointed and sparse. Candidates misfitted the Rasch model to a large extent. Interaction effects were revealed in the G-theory analysis, and Rasch modelling is only robust to this up to a point.

7.2 Stability of rater effects: intra-rater reliability

Our measures of rater effects are crucial to performance monitoring of examiners, but we noted in Chapter 2 that previous research has indicated that they may be unstable (e.g. Lamprianou, 2006; Leckie and Baird, 2011; Myford and Wolfe, 2009). The current data allowed analysis of stability of rater effects within a marking series, across the two ways of monitoring marker performance (backread and seeded clips), and over years. The G-theory analyses found little evidence of variability of rater effects over an examination series. This was not investigated using the other two methods.

The MFRM investigated stability of rater effects for the psychology and geography examiners across years and found that they were unstable. Using multilevel modelling, this was also found for the psychology examiners. However, there were very few examiners in these analyses, so the findings might not be generalisable. Further, it was noted that the removal of outliers in the MFRM analysis would have resulted in moderate correlations. For the geography data, moderate correlations of rater effects were found between years using multilevel modelling. Thus, though inconclusive, some evidence of the stability of rater effect was found. Given the small magnitude of rater effects in the multilevel models, we might have expected the correlations to be small.

Stability of the estimates of rater effects between monitoring systems was investigated using multilevel modelling. With these analyses, the number of examiners included was greater as every examiner was subjected to both systems of monitoring. Moderate

correlations were found, which could again be seen as encouraging, given the small magnitude of rater effects in these analyses.

7.3 Question effects

The impact of questions (or items) could not be investigated using MFRM, but it was found to be a significant source of variance in the generalizability and multilevel analyses. As the item tariffs varied, it is difficult to draw conclusions from this finding, but it might be assumed that some questions would be more difficult to mark accurately than others – the generalizability analyses of seeded clip data would support this assumption. To fully explore such an issue, we would need a more extensive and representative dataset. Limitations of the operational data for studies of this kind are outlined below.

7.4 Training group effects

Baird, Leckie and Meadows (*in submission*) found that the team in which examiners underwent standardisation training for marking had an impact upon rater severity. In that study, marker training was conducted face-to-face. The current study represented a contrast, in that training took place remotely and electronically (with the exception of the psychology examination in 2009). Without face-to-face discussion, we might expect less of an impact of the teams in which markers were grouped. There was still a possibility that the feedback given by supervisors (team leaders) through the monitoring process could have impacted on their severity, so this research set out to investigate any team effects. MFRM was not a suitable technique for this research question because teams were disjoined subsets of the data (examiners were assigned to one team only). No evidence of team effects was found in the G-theory analyses for either subject in any year, and they were found to be very small in the multilevel analysis for geography. Team effects were not investigated using multilevel modelling for psychology due to the low number of training teams, although it would be possible to investigate them as a fixed effect in the model.

7.5 The three approaches to rater effect investigation

We set out in this project to apply to the response data from two typical AS component papers the three main approaches to analysing rater effects in use today, *viz*. G-theory, Rasch analysis and multilevel modelling, to see how they might complement each other in addressing the research aims.

G-theory has a very rich conceptual framework and has been traditionally used to address measurement issues in the wider context of educational assessment (e.g. Lamprianou and Christie, 2009; Johnson and Johnson 2012a). As a measurement method, Brennan (2001a) illustrated how G-theory can be used to quantify and explain

the consistency of observed scores, and how it can be very useful in the case of multifaceted designs (such as we see in the current research). By 'facets' we mean the kind of replications included in the reliability study, in particular raters, questions and occasions and their interactions. G-theory has been compared to a special case of IRT models (i.e. the Many-facets Rasch model - MFRM) and has been shown to have many common characteristics, but also some differences (e.g. Sudweeks, Reeve and Bradshaw, 2005; Smith and Kulikowich, 2004). Interestingly, Sudweeks et al. (2005) found that the two methods gave similar results for a simple dataset, although they were reported in slightly different ways. Their dataset was also fully crossed, whereas the data analysed in the current research was often a sparse matrix. They concluded that the results of the two methods could complement each other. Smith and Kulikowich (2004) reached similar conclusions and stated that

> … both measurement techniques agree on the relative magnitudes of variation among the facets but differ on how to handle the sources of variation. (p.617)

One could interpret Linacre and Wright (2002) as saying that the main aims of the two methods are: (a) in G-theory to estimate the variance produced by each facet, whereas, (b) in MFRM to estimate 'adjusted' measures for each of the elements of the facets. More recently, Kim and Wilson (2009) used empirical data to illustrate that:

> The view that G theory and the MFRM are alternative solutions to the same measurement problem, in particular, rater effects, is seen to be only partially true. G theory provides a general summary including an estimation of the relative influence of each facet on a measure and the reliability of a decision based on the data. MFRM concentrates on the individual examinee or rater and provides as fair a measure as it is possible to derive from the data as well as summary information such as reliability indices and ways to express the relative influence of the facets. (Kim and Wilson, 2009, p.408).

MLM techniques have also been used in the context of educational assessment in order to address issues of repeated measures or facets. Interestingly, it has been shown that various Rasch models may be conceptualized as special cases of Generalized Mixed Effects Models (e.g. Kamata, 2001) – or else Multilevel Models. For example, Ong et al. (*in press*) used Multilevel Models to apply a dichotomous Rasch model to assessment data in order to investigate Differential Bundle Functioning.[3] The researchers were initially inspired by Doran et al. (2007) who illustrated how easily a Rasch model may be conceptualized as a special case of Generalized Mixed Effects Models using a generic statistical package instead of one of the more 'traditional' and specialised packages (also see Johnson, 2007). Very recently, DeBoeck et al. (2011) have shown how various Rasch

[3] A bundle of items is not identical to a testlet, hence the use of this term here.

models may be easily fitted in the same context and using the same generic software. Indeed, from a strictly algebraic point of view both the designs of G-theory and the models of the Rasch family are fairly evidently special cases of multilevel models (Briggs and Wilson, 2007; see also the insightful discussion in Verhelst and Verstalen, 2001). From this perspective, the main practical difference between fitting and interpreting a Multilevel Model, a Rasch model or a G-theory design using more 'traditional' software probably has to do with differences in the philosophy, tradition, the richness of the available software, the underlying sampling and distributional assumptions, the estimation methods, and the presentation of results.

From the preceding discussion, one might have expected that the results of the three methods would not be very different – but differences were indeed observed. Although this was not expected, it was consistent with some other studies that found very different results between some of those methods. For example, Macmillan (2000) used large, sparse matrices of examination data, like the datasets we analysed, and compared the results from classical test theory, generalizability theory and many-facets Rasch modelling. He found that

> CTT and MFRM indicated substantial variation among raters; the MFRM analysis identified far more raters as different than the CTT analysis did. In contrast, the GT rater variance component and the Rasch histograms suggested little rater variation. (Macmillan, 2000, p.167).

Also, Lynch and McNamara (1998) found "striking", as they called them, differences between the results of a many-facets analysis and a generalizability analysis in a similar setting:

> A somewhat striking difference between the two approaches is clear in the contrasting results for the interaction effects in the two analyses. (Lynch and McNamara, 1998, p.176)

Linacre (2002) presents another example where striking differences were found between the results of generalizability theory and many-facets Rasch measurement, writing that "This [the results] challenges the G-Theory finding that rater differences can be safely ignored." (p. 499).

In addition to differences due to the software, the estimation method, the different assumptions of each model, the reporting style and the philosophical paradigm behind any analyses, it is possible that differences occurred because each group of researchers working with one of the three methodologies approached the global task from different

research perspectives (e.g. different groups of researchers may use different subsets of the data). Overall, we could generally suggest that the Generalized Mixed Effects Models follow the regression approach, and therefore fit nicely with the explanatory research paradigm. Since they allow for fixed and random regression weights for items as well as for persons, they provide a complete toolbox for explanatory modelling. On the other hand, Rasch models fit in the latent variable paradigm where there is no room for an independent variable or causal status. G-theory, for its part, was specifically developed to address issues of replication-based measurement reliability. Multilevel modelling is designed to address complexity in data and, in particular, hierarchical data. This technique does not have theoretical underpinnings with regard to reliability, but is best associated with classical test theory.

7.6 The craft of analysis

As with all statistical analysis, the choices to be made regarding the selection and preparation of the data and the setting up of the model are based upon statistical principles, theoretical issues and the craft knowledge of the researcher. Given the same three-year datasets for each of the two subjects, application of each of the techniques has involved quite different selections from that data to serve different analysis aims.

Backread records were not analysed using the G-theory or Rasch modelling techniques, mainly because the two marks for each clip were not independently awarded. In direct contrast, in the multilevel models, only clips that had been multiply marked were retained for analysis. Separate G-theory analyses were conducted on the seeded clip data in those cases where there were sufficient multiply marked clips to justify this. For the G-theory and Rasch analyses, annual datasets were analysed separately, but all three years' data were analysed together in the multilevel model.

To investigate stability of rater effects, marking phase was included as a variable in the G-theory analyses. In the Rasch analyses, rater stability over years was investigated and in the multilevel models, stability over monitoring system and years were pursued. Using each technique, the research decisions tended to focus upon the ways in which the research questions can best and most easily be addressed. In G-theory, partitioning rater effects into chunks of variance to compare is an obvious choice. For IRT, correlating the rater severity parameters between years is a possibility (although the number of raters was small). Using multilevel modelling, it was possible to investigate the stability of rater effects over years and monitoring systems.

7.7 Limitations of the study

Many of the limitations of the research relate to the fact that we analysed operational data that were generated for the purpose of allocating scores to candidates and

monitoring marker performance. A designed study could overcome some of these limitations, but these can be costly. For all of the following reasons, we need to be cautious about the likely dependability of our findings.

Our datasets were drawn from only two AS question papers across three examination series. As previously noted, there were a limited number of examiners involved in marking those examinations. We did not have complete scripts multiply-marked in the backread and seeded monitoring data. Analyses were not therefore conducted at question paper level, but at the level of clips. A further problem then arises with the lack of a standard mark tariff for clips. Moreover, scanned clips were not drawn to be representative of the distribution of scores in the candidate population, but were selected for other operational purposes. Clips were not equally representative of questions and there was variation in the number of times clips were multiply marked.

Neither were the single-marked data ideal for estimation of reliability, as clips are not randomly allocated to examiners. To be clear, the system *does* randomise allocation of clips to markers, but it can only do this on the basis of the pool of *available* clips at any given time. Markers also work at different speeds, and as a result there can be a large variation in the number of clips evaluated by each of them. Interpretation of differences in rater effects is impossible in analyses of the single-mark data, as candidate ability and rater effects are confounded.

7.8 Implications for future quality monitoring processes

To enable more dependable estimates of the reliability of marking, consideration could be given to the way in which clips are selected and allocated to markers. Scanning the work of whole candidates to use as seeded clips would enable estimates of reliability to be made at question paper level as opposed to merely at clip level. Selecting clips to be equally representative of items and representative of the distribution of candidate scores would also aid the capacity for research into the reliability of the question papers. Operational processes might preclude these modifications in their entirety, but adapting systems to produce better representation of the question papers, scores and candidates would be beneficial for future research on reliability using these operational data.

8 References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika,* 43: 561-573.

Baird, J., Greatorex, J. and Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice,* 11: 331-348.

Baird, J., Leckie, G. and Meadows, M. (*in submission*) Effects of training groups upon examiner accuracy.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M.Lord and M.R. Novick (eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, M A: Addison-Wesley.

Bramley, T. and Dhawan, V. (2012). *Estimates of Reliability of Qualifications*. Chapter 7 in Opposs and He (2012), pp.217-320.

Brennan, R.L. (2001a). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R.L. (2001b). *Manual for urGENOVA version 2.1*. Iowa testing programs occasional papers, no. 49.

Brennan, R.L. (2001c). *Manual for mGENOVA version 2.1*. Iowa testing programs occasional papers, no. 50.

Briggs, D.C. and Wilson, M. (2007). Generalizability in Item Response Modeling. *Journal of Educational Measurement*, 44: 131-155.

Browne, W.J. (2009). *MCMC Estimation in MLwiN, v2.13*. Centre for Multilevel Modelling, University of Bristol.

Browne, W.J., Goldstein, H. and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1: 103-124.

Cammilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement,* 32: 79-96.

Cardinet, J., Johnson, S. and Pini, G-R. (2010). *Applying Generalizability Theory using EduG*. New York: Routledge.

Congdon, P.J., and McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37: 163-178.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education,* 36: 1-21.

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. New York: Wiley.

Curtis, D. D. (2004). Person Misfit in Attitude Surveys: Influences, Impacts and Implications. *International Education Journal,* 5: 125-144.

DeBoeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F. and Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software,* 39: 1-28.

Doran H., Bates D., Bliese P. and Dowling, M. (2007). Estimating the Multilevel Rasch Model: With the lme4 Package. *Journal of Statistical Software,* 20: 1-18. URLhttp://www.jstatsoft.org/v20/i02/.

Dorans, N.J. and Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement,* 22: 249-262.

Eckes, T. (2005). Examining Rater Effects in Test DaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly,* 2: 197–221.

Embretson, S.E. and Reise, S. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5: 171–191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement,* 31: 93–112.

Goldstein, H. (2010). *Multilevel statistical models*. 4th Edition, London: Arnold.

Greatorex, J. and Bell, J.F. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education,* 23: 233-255.

Harik, P., Clauser, B.E., Grabovsky, I., Nungester, R.J., Swanson, D. and Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, 46: 43-58.

He, Q. (2012). Estimating the reliability of composite scores. Chapter 12 in Opposs and He (2012), 523-556.

Hoskens, M. and Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden state examination. *Journal of Educational Measurement*, 38: 121-145.

Humphreys, L.G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das and N. O'Connor (eds), *Intelligence and Learning* (pp. 87-102). New York: Plenum.

Ipsos MORI (2009) *Public perceptions of reliability in examinations*. Available online at: http://www.ofqual.gov.uk/files/2009-05-14_public_perceptions_of_reliability.pdf.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika,* 51: 357-373.

Johnson, M.S. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R. *Journal of Statistical Software,* 20: 1-24.

Johnson, M., Nadas, R. and Bell, J.F. (2010). Marking essays on screen: an investigation into the reliability of marking extended subjective texts. *British Journal of Education Technology*, 41: 814-826.

Johnson, S. and Johnson, R. (2012a). Component reliability in GCSE and GCE. Chapter 6 in Opposs and He (2012), pp.141-216.

Johnson, S. and Johnson, R. (2012b). Conceptualising and interpreting reliability. Chapter 11 in Opposs and He (2012), pp.459-522.

Johnson, S., Johnson, R., Miller, L. and Boyle, A. (2013). *Reliability of Vocational Assessment: An evaluation of level 3 electro-technical qualifications*. Coventry: Office for Qualifications and Examinations Regulation.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38: 79-93.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement,* 1: 152–176.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education,* 16: 277-298.

Kim, S.C. and Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalisability theory and the many-facet Rasch model. *Journal of Applied Measurement,* 10: 408-23.

Kingsbury, F.A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research,* 1: 377-383.

Klauer, K.C. (1995). The assessment of person fit. In G. H. Fischer and I. W. Molenaar (eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 97–110). New York: Academic Press.

Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. *Journal of Applied Measurement,* 7: 192-200.

Lamprianou, I., (2008). High Stakes Tests with Self-Selected Essay Questions: Addressing Issues of Fairness, *International Journal of Testing,* 8: 55 – 89.

Lamprianou, I., (2010). The practical application of Optimal Appropriateness Measurement on empirical data using Rasch Models. *Journal of Applied Measurement,* 11: 409-423.

Lamprianou, I. and Boyle, B. (2004). Accuracy of measurement in the context of mathematics National Curriculum tests in England for Ethnic Minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement,* 41: 239-260.

Lamprianou, I. and Christie, T. (2009). Why school based assessment is not a universal feature of high stakes assessment systems? *Educational Assessment, Evaluation and Accountability,* 21: 329-345.

Leckie, G. and Baird, J. (2011). Scoring severity: effects of rater drift, rater experience and scale use, *Journal of Educational Measurement,* 48: 399 – 418.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement* (2nd ed.). Chicago: MESA Press

Linacre, J.M. (1997). MESA Release note #3. Online at http://www.rasch.org/rn3.htm; accessed January 3rd, 2012.

Linacre, J.M. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, 3: 484-509.

Linacre, J.M. (2011). *Facets Computer Program for Many-Facet Rasch Measurement,* version 3.68.1. Beaverton, Oregon: Winsteps.com.

Linacre, J.M. and Wright, B. (2002). Construction of measures from many-facet data. In E. Smith and R. Smith (eds.) *Introduction to Rasch Measurement. Theory, Models and Applications.* (pp. 296-321) Maple Grove, MN: JAM Press.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hilldale, NJ: Laurence Erlbaum.

Lunz, M.E. and Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13: 425-444.

Lunz, M.E., Wright, B.D. and Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education,* 3: 331–345.

Lynch, B.K. and McNamara, T.F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. Language Testing, 15: 158-180.

Macmillan, P.D. (2000). Classical, Generalizability and Multifaceted Rasch detection of interrater variability in large, sparse sets. The Journal of Experimental Education, 68: 167-190.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology,* 34: 100-117.

McManus, I.C., Thompson, M. and Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education,* 6:42.

Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability.* Manchester: AQA.

McVey, P.J. (1975). The errors in scoring examination scripts in electronic engineering. *International Journal of Electronic Engineering Education,* 12: 203-216.

Murphy, R.J.L. (1979). Removing the scores from examination scripts before re-scoring them: Does it make any difference? *British Journal of Educational Psychology,* 49: 73-78.

Myford, C.M. and Wolfe, E.W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46: 371-389.

Ong, Y-M., Williams, S.J. and Lamprianou, I. (2011). Exploration of the Validity of Gender Differences in Mathematics Assessment Using Differential Bundle Functioning. *International Journal of Testing,* 11: 1-23.

Ong, Y-M., Williams, S.J. and Lamprianou, I. (*in press*). Exploring differential bundle functioning in mathematics by gender: the effect of hierarchical modelling. *International Journal of Research & Method in Education*.

Opposs, D. and He, Q. (eds) (2012). *Ofqual's Reliability Compendium*. Ofqual/12/5117. Coventry: Office of Qualifications and Examinations Regulation. ISBN 978-0-85743-016-8.

Pinot de Moira, A., Massey, C., Baird, J. and Morrissy, M. (2002). Marking consistency over time. *Research in Education*, 67: 79-87.

Powers, D. and Kubota, M. (1998). Qualifying essay readers for an online scoring network (OSN). *Research Report Educational Testing Service.* Princeton, NJ: Educational Testing Service.

Rasbash, J., Steele, F., Browne, W. J. and Goldstein, H. (2009). *A User's Guide to MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* London: Sage Publications.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15: 351-357.

Royal-Dawson, L. and Baird, J. (2009). The impact of teaching experience upon marking reliability in Key Stage 3 English. *Educational Measurement: Issues and Practice*, 28: 2-8.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes,* 8: 13-17.

Shohamy, E., Gordon, C.M., and Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76: 27-33.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin,* 86: 420-428.

Smith, E.V. and Kulikowich, J.M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement,* 64: 617-639.

Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement,* 5: 541–565.

Smith, R.M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement,* 1: 199–218.

Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods*. 8th edition. Iowa State University Press.

Snijders, T.A.B. and Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications.

Stout, W.F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika,* 55: 293-325.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika,* 52: 589-618.

Sudweeks, R.R., Reeve, S. and Bradshaw, W.S. (2005). A comparison of generalisability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing,* 9: 239-261.

Suto, W.M.I., Nadas, R. and Bell, J.F. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education,* 26: 21-51.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (ed.) *Applications of Item Response Theory*. British Columbia:

Educational Research Institute of British Columbia.

Verhelst, N.D. and Verstralen, H.H.F.M. (2001). An IRT model for multiple raters. In M. Boomsma, M.A.J. van Duijn and T.A.B. Snijders (eds.), *Essays on item response theory* (pp. 88–108). New York: Springer-Verlag.

Vidal Rodeiro, C. (2007). Agreement between outcomes from different double-marking models. *Research Matters: A Cambridge Assessment Publication*, 4: 28-34.

Wood, R. (1991). *Assessment and Testing*. Cambridge: University of Cambridge Local Examination Syndicate.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, B.D. and Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement,* 1: 83–106.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

Yen, W.M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika,* 50: 399-410.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2013

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346