assessment europe

Component reliability in GCSE and GCE

Sandra Johnson
Rod Johnson

Ofqual/11/4780
November 2010

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

Preface

This report is one outcome of a project commissioned by the Office of Qualifications and Examinations Regulation (Ofqual) in April 2010 (Contract no.OF-102), that focused on estimating the reliability of GCSE and GCE examination components using generalizability theory.

For the purpose of the project reliability was pre-defined as follows:

> *Reliability refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. A range of factors that exist in the assessment process can introduce unreliability into assessment results. Given the general parameters and controls that have been established for an assessment process – including test specification, administration conditions, approach to marking, linking design and so on – (un)reliability concerns the impact of the particular details that do happen to vary from one assessment to the next for whatever reason.*

Four potentially important sources of measurement error were identified alongside the given definition of reliability reproduced above, *viz*. occasion-related, test-related, marker-related and grading-related. There would be no opportunity within this project to explore the issue of occasion-related measurement error, and grading-related error contribution was beyond the project's scope. The project was rather intended to explore the impact on reliability of both test-related and marker-related factors. It was anticipated to generate reliability estimates for a range of GCSE and GCE components (units), with a particular focus on the quantification of the relative contributions from different sources of error to the overall error of measurement. The main activities of the research were expected to involve:

- the selection of a range of GCSE or GCE components for study;
- the identification of the major sources of error for the selected components;
- the compilation of the necessary data or the design of experiments to collect data for analysis;
- the development of a mechanism to quantify the overall reliability measures and the relative contributions from individual error sources to the overall error of measurement for the selected components;
- the estimation of standard errors of measurement (expressed in terms of raw score/standardised score and/or grade) for the selected components;
- the analysis, interpretation and reporting of the reliability evidence generated.

In the event, the project, during which numerous components from 2009 GCSE and GCE examinations were empirically explored, was for the most part constrained to focus on test-related factors only. This is essentially because all live examining for GCSE and GCE is single-marked, leaving no possibility within the 2009 datasets to explore marker effects. And while marker standardisation studies are routinely carried out by the examining boards, the design of these studies is typically too limited to offer scope for the simultaneous investigation of marker and question effects on the reliability of candidate scores. Despite this limitation, the analyses carried out for the project, which cover a variety of different examination subjects and a variety of differently structured component papers, will be of interest to anyone involved in the examining process.

Acknowledgements

This project depended critically on the availability for analysis of operational data from GCSE and GCE examinations. In the event, we were extremely fortunate to benefit from the collaboration of four of the six awarding bodies that offer GCSE/GCE examinations, or their Scottish equivalents, in the UK. All four awarding bodies supplied us with datasets that we specifically requested for analysis, and gave their permission to include the results of those analyses in this report. We are indebted to the following individuals and their organisations for their much-appreciated support: Michelle Meadows and Ian Stockford of the Assessment and Qualifications Alliance (AQA), Jeremy Pritchard and Linda To of Edexcel, Rob van Krieken and Susan Kirk of the Scottish Qualifications Authority (SQA), and Raymond Tongue and Jo Richards of the Welsh Joint Education Committee (WJEC). The Council for the Curriculum, Examinations and Assessments (CCEA) in Northern Ireland was also willing to participate in the project, but had at the time no question-level data available for analysis; CCEA began electronically recording response data at question level in 2010, just too late for inclusion here.

We also express our thanks to Qingping He of the Office of Qualifications and Examinations Regulation (Ofqual) and to Paul Black, Jo-Anne Baird and Gordon Stanley, members of Ofqual's Reliability Programme Technical Advisory Group, for their constructive comments on an earlier draft. Finally, we are happy to acknowledge the very efficient and helpful way in which Jo Taylor of Ofqual managed the scheduling of report drafting, review and publication.

Contents

Executive summary

The principal purpose of this project was to exemplify the potential of generalizability theory for research into the reliability of UK examinations, through applications using operational examination data. The intention was to explore the impact on component reliability at GCE and GCSE of test-based and marker-based factors.

Four examining boards agreed to collaborate with us and generously supplied us with over 50 different datasets from 2009 examinations, along with other relevant information about the component papers concerned. The papers whose datasets we requested were specifically selected by us to offer a variety of examination subject and paper structure. Unfortunately, with just one exception, we were unable to gain access to datasets that would permit the simultaneous investigation of both marker-based and test-based factors on reliability. Our analyses, therefore, have focused principally on test-based factors. Some interesting findings have nevertheless emerged.

The first feature worthy of note concerns mark distributions. While most of the component paper mark distributions were symmetrical, or nearly so, it was quite common for the underlying mark scale not to be fully used, and for the distribution to be relatively peaked. When mark scales are not fully used this must pose a problem for the validity and reliability of candidate grading. To accommodate greater study and assessment flexibility for qualification-seeking candidates, grading decisions are now made for unit papers separately, rather than for the examination a whole. The fact that sometimes relatively short intended mark scales are becoming even shorter achieved scales before being subdivided into seven parts to accommodate six grade boundaries must be a cause for concern.

Truncated mark distributions also impact on the reliability of individual candidate measurement, as do other factors, including marker and question variation. In our empirical analyses we computed two different indices of reliability: variance ratios, i.e. reliability, or generalizability, coefficients, and standard errors of measurement, from which confidence intervals around candidates' total marks can be calculated. The magnitude of reliability coefficients varied from one component paper to another, with interestingly low values in some cases. In general, 95% confidence intervals around candidates' paper marks spanned 20% to 40% of the underlying mark scale, but were much smaller for some unit papers.

Papers in which different questions carried different mark allocations were not uncommon, and caused problems for reliability interpretation and generalisation. And for some papers it proved impossible even to attempt to quantify reliability, because the paper itself, or a section within it, comprised one single question, leaving no scope for variance analysis; essay papers and writing assessments in language papers are particular examples.

The research findings raise issues in terms of the validity/reliability tension, and strongly suggest that research into the impact of paper structures on both assessment validity and assessment reliability be stepped up. Our initial investigations suggest that generalizability theory has a clear role to play in this research.

1 Introduction

1.1 The UK's external examining system

The school-level academic qualifications system in the UK is strikingly different from those of most countries in continental Europe, and indeed most countries in the developed world. Some countries offer no external examinations at all that lead to formal school leaving qualifications, certification being left entirely to the discretion of students' teachers, often with little or no guidance in the form of assessment criteria and no concern about comparability of standards across schools – the USA is a particular example. It is left to private testing organisations or to universities and colleges to undertake the task of assessing candidates for selection into employment or higher education. In contrast, countries that do offer formal school leaving qualifications typically do so on the basis of a national centrally-set examination, with every student candidate throughout the country attempting the same examination in any particular year. There might be subject options within the qualification as a whole and within the suite of examination papers, but for every subject within the range of offerings the same paper(s) will be attempted by all candidates choosing that option. Most continental European countries could be cited as relevant examples here.

In England, Wales and Northern Ireland the pre-university academic school leaving certificate is the now unitised *General Certificate of Education,* at Advanced or Advanced Subsidiary Level (A/AS level). School students typically achieve an AS qualification by passing two unit papers from four, with all four units leading to the A level. These school leaving qualifications are typically taken during Years 12 and 13 (17-18 year olds). Scotland has its own education and assessment system, with corresponding school leaving qualifications. A main difference between these UK qualifications and their apparent equivalents in other countries, such as the French *Baccalauréat*, the Italian *Maturità* and the Dutch *VWO*, is their narrower focus. The A/AS level remains for the most part a relatively specialised single-subject qualification. Even when students take three or four different A levels their joint subject coverage is rarely as broad as that of its continental equivalents. Students in principle have great freedom of choice in their A level subject combinations, even when constrained to some extent by higher education intentions and school timetabling and staffing issues. This variety of choice is extended even further to choices of different syllabus specification within the same subject, and to examinations in the same subject offered by more than one awarding body.

Another difference between the UK and many other countries in terms of school qualifications resides in the *General Certificate of Secondary Education* (GCSE). GCSEs are again single-subject qualifications, normally taken at the end of Year 11 (16 year olds). But at this stage students typically take a greater number of subject examinations in a greater variety of subjects than is the case one or two years later at A/AS level.

GCSEs and GCE A/AS levels are offered by a small subset of the 120+ awarding bodies that currently operate in the UK, and that between them offer over 6000 nationally accredited academic and vocational/occupational qualifications (see the *National Database of Accredited Qualifications* for full details: www.accreditedqualifications.org.uk). The subset of awarding bodies, whose members are often for historical reasons still referred to as 'examining boards',

comprises the Assessment and Qualifications Alliance (AQA), Edexcel, and Oxford, Cambridge and RSA Examinations (OCR) in England, the Welsh Joint Education Committee (WJEC) in Wales, and the Council for the Curriculum, Examinations and Assessments (CCEA) in Northern Ireland. Scottish qualifications are offered by the Scottish Qualifications Authority (SQA). An individual board can include literally hundreds of different qualifications in its annual offerings to candidates, and the set of offerings is under continual evolution.

The entire secondary-level examinations and qualifications system is regulated in England by the Office of Qualifications and Examinations Regulation (Ofqual), in Wales by the Department for Children, Education, Lifelong Learning and Skills (DCELLS), in Northern Ireland by the Council for the Curriculum, Examinations and Assessments (CCEA), and in Scotland by the Scottish Qualifications Authority (SQA). Ofqual also oversees the quality of the National Curriculum Assessment programme that continues to operate in the primary and lower secondary sectors in England, as well as the relatively new academic/vocational Diploma.

1.2 Variety in examination structures and component formats

In all UK countries, qualifications are available in a wide variety of traditional and less traditional subjects, including, for example, history, French, mathematics, business studies, ICT, art and design, citizenship studies, drama, psychology. Examinations are now typically modular, with individual units often assessed in different ways within a single examination. Examination components might be uniquely written papers, comprising objective questions, constructed response questions, or a mixture of both. Written theory papers might alternatively be complemented by practical tests of one kind or another: for example by practical laboratory tasks in the sciences, instrumental performance in music examinations, or oral assessments in foreign language qualifications. In many subjects at GCSE teacher-assessed course work is also included in the final profile of attainment evidence that ultimately leads to an examination grade for the candidate concerned.

It might be useful at this point to offer some examples to illustrate the current variety of unit-based examination structure at the different examination levels, as well as the variety of component paper composition (the websites of the various examining boards offer full detail).

Among numerous examples of examinations that uniquely comprise written papers we can cite WJEC's Advanced Subsidiary GCE in Psychology, whose two mandatory papers contribute, respectively, 40% and 60% to the total qualification, and AQA's Advanced Subsidiary GCE in English Language (A), again with two mandatory written papers, contributing equally this time to the total qualification. Other qualifications can be based entirely on the evaluation of portfolio evidence. An example is Edexcel's GCE A level in Art and Design. This qualification comprises four independently assessed mandatory units: two coursework portfolios and two externally set assignment portfolios. Unit weightings in the qualification are 20% or 30%.

Still other examinations combine written papers with coursework, as does, for example, OCR's Advanced GCE in Media Studies: the equally weighted 4-unit qualification includes a mandatory coursework unit, two mandatory written papers

and a choice of written paper for the fourth unit. An example of an examination that combines written testing with task-based controlled assessment and practical demonstration is AQA's GCSE in Music (a controlled assessment is a teacher-supervised assessment of course work learning). Candidates are assessed for listening to and appraising music through aural and written examinations, for composing and appraising music through a written examination and task-based controlled assessment, for performing music though a practical demonstration and task-based controlled assessment, and for composing music through task-based controlled assessment. A third example is a CCEA GCSE in History. In this 3-unit examination two units comprise externally assessed written papers, the third being an internally assessed externally moderated controlled assessment. One of the written papers presents five options for prior study, from which teachers select two.

The OCR GCSE in Dutch, like many other language qualifications, assesses candidates' reading and writing skills in two separate written papers, speaking skills through an oral examination, and listening skills through an aural examination. As a final example we offer Edexcel's Advanced GCE with Advanced Subsidiary GCE (Additional) in Applied Information and Communication Technology. This has eight mandatory units and three optional units of which one must be taken: the nature of the units is not included in the qualification description. Two of the mandatory units are externally assessed, the rest being internally assessed and externally moderated by Edexcel. All units contribute equally to the qualification as a whole.

Written component papers within and across examinations show a similar variety of different forms, as the examples analysed in Chapters 4 and 5 illustrate. A paper might take the form of an objective test, presented on paper or online. Or it might comprise a series of short-answer questions, or a set of structured response questions sometimes requiring quite extended written responses, or a single essay question chosen from a given list of titles. Then again the paper might contain a mixture of different question types, collected together in sections. And while in some cases all questions are mandatory, in others there might be some question choice. Finally, while many papers award the same maximum marks to the different constituent questions, others award different maximum marks. Indeed, a frequent format is for there to be varying numbers of subquestions within questions, with both questions and subquestions carrying different mark allocations, presumably representing the views of subject specialists about the relative importance of different constituent content and skills as reflected in the underlying paper specification.

Practical components might offer choice of task to candidates, such as choice of topic to discuss with an oral assessor, or choice of instrument to play. Or assessors might allocate a task to candidates at random. Or all candidates in any given year will be expected to undertake the same common task. Either way any one candidate is usually given the opportunity to attempt one task only. Portfolios eliminate this task restriction to a great degree, but bring their own assessment challenges.

To add to this complexity, and notwithstanding the assumption that qualifications in the same subject at the same level from different examining boards are 'comparable', examinations in the same subject at the same level offered by different boards can also take different structures, and component papers can take different forms. Both examination structures and component forms can also change over time, within or

across boards, in response to government initiatives and individual board innovation. The situation is quite dynamic. The critical question is how reliable are the resulting examination components and their parent examinations?

1.3 Content of this report

In an earlier report (Johnson & Johnson 2009) we comprehensively outlined the application potential of generalizability theory (Cronbach *et al.* 1972, Shavelson & Webb 1991, Brennan 1992, 2001a), or G-theory, for estimating assessment reliability in the context of the kinds of tests and examinations that are currently used in the UK. It was not possible in that report to illustrate the potential of G-theory through application to real examination datasets. The remit of the project described in this current report was therefore to apply G-theory to UK operational examination data, to quantify and interpret the reliability of a selection of examination components.

The scope of the project did not extend to the issue of whole examination reliability. Indeed, in many cases this would be difficult to do, even impossible to do, using the variance analysis approach. This is because in order to use variance analysis to explore the impact of potentially influential factors on assessment reliability, and more specifically on measurement error, at least two observations of a factor would be required to feature within a dateset: two or more markers, two or more test questions, two or more alternative papers, and so on.

Where written component papers, such as an English essay paper, require candidates to answer one single question then clearly no variance component for questions can be produced. Similarly, when practical components require all candidates to attempt the same task, whether carrying out a physics experiment, playing a musical instrument or engaging in a one-to-one conversation in a foreign language, no between-task variance can be estimated and neither can interaction effects with tasks be explored. Equally, when the assessment evidence of any one candidate is marked by just one marker then no between-marker variance is available for analysis, from which assessment reliability might be estimated and results generalised. Since in any one year the candidates taking a practical component typically all attempt the same single task, and since all components, practical or written, are assessed by just one marker per candidate (or per candidate-question), it has not been possible in this project to look at anything other than written component papers, and only then at the impact of test-based characteristics on assessment reliability (this is with the exception of the small dataset analysed in Chapter 3, which does allow some investigation of the impact of markers as well as of questions on candidate outcomes).

Even among written papers there are examples where the variance analysis approach has been impossible to apply using archived operational data, or where application results would be of limited value. An example of the former is a GCSE French unit paper, that comprised three sections, one for listening, one for reading and one for writing, and in which the writing assessment required candidates to produce a single extended piece of writing on a topic chosen from three options. Units based on portfolio assessment would be another example, since a single portfolio would typically be evaluated by a single rater, generally the candidate's classroom teacher (a very small subsample of portfolios would normally be externally checked by teacher moderators). The solution is to organise designed studies, preferably during the qualification development phase when examination structures are designed, in order

to explore the potential influences of different variables on assessment reliability. The resulting findings could then be used to ensure that examinations have structures that are not only deemed by the responsible principal examiners to be acceptable in terms of assessment validity but which also lend themselves to ongoing reliability investigation.

Examples of component papers where generalizability analyses could be carried out, but where the results of the analyses would have limited value, include all those papers in which different test questions carry different total marks, i.e. are given differential weights in the paper total, and where there is just one single question with any particular mark total. Several examples are included in this report. While reliability indices can be calculated for such papers it is not clear how the results might be meaningfully generalised to past or future papers of similar structure, given the usual assumption that questions are sampled from a defined subject domain. If reliability is considered important to quantify for all component papers in the future then the rationales for some current paper structures would be worth exploring to evaluate prospects for modifications that would facilitate reliability investigation without unduly threatening assessment validity.

In this project we were fortunate to have the active support of four examining boards, all of which willingly supplied us with all the datasets that we requested: AQA, Edexcel, SQA and WJEC. Between them these boards supplied around 60 datasets, which emanated from 2009 GCSE or GCE examinations or their Scottish equivalents in a variety of different subjects. All the supplied datasets were analysed, and reports on each prepared for the supplying boards' information and use. In this current report we have specifically selected a subset of the datasets, to illustrate the application potential of G-theory in this context. To safeguard anonymity for all the boards, Scottish datasets are labelled as GCSE or GCE as most appropriate.

By design, the datasets that feature in the report offer a variety of component structure and of examination subject. This is partly to guarantee maximum scope for G-theory exemplification, but partly also to research possible variation in reliability outcome related to structure and subject. The datasets and their analysis results are described in Chapters 4 and 5. Chapter 4 focuses on the simplest analysis design, which, in the absence of marker information, explores the influence of examination questions on the reliability of candidate measurement. Chapter 5 takes the modelling a little further, extending consideration to the influence of subquestions as well as of questions on measurement reliability, along with the influence of candidate characteristics. In particular, Chapter 5 offers examples of the estimation of the reliability of composite scores for structured papers whose sections are distinguished by question format and weight.

Before presenting the application results, we offer in Chapter 2 an overview of the theoretical basis for the indices of reliability that we use in the applications, and in Chapter 3 we focus more specifically on G-theory itself, using real data to illustrate fundamental concepts.

2 Some theoretical background

In this chapter we offer a succinct overview of some of the technical background which underlies the models and analyses proposed in the remainder of the report. The exposition here is by design superficial, intended only to suggest what the technical issues might be, without proofs or detailed derivations. A more detailed review of the material in the first part of the chapter can be found in our earlier report (Johnson & Johnson, 2009).

2.1 The basic model

In the field of educational and behavioural measurement we frequently have to deal with observations that arise naturally from grouped data: pupils sharing the same teacher, questions on the same paper, scripts evaluated by the same marker, are some of many possible examples. A typical model for this class of situations is

[2.1] $Y_{ij} = \mu + A_i + R_{ij}$,

Essentially, the *j*th observation on the *i*th group can be broken down into an overall mean value (μ), a component due to the influence of the *i*th group (A_i), and a component arising from random fluctuations in the measurement process (R_{ij}), typically called a *residual*.

It is customary to add some standard assumptions to the basic model [2.1]. One of these is that the expected values of A_i and R_{ij} are zero, so that the expected value of Y_{ij} must be μ. Conventionally, also, the A_i and R_{ij} are assumed to be linearly independent, and hence uncorrelated among themselves. Note that we do not need to make any further distributional assumptions about the A_i and R_{ij}, other than that their variances exist. In particular we do not need the assumption that the R_{ij} are normally distributed.

A specific consequence of the linear independence assumption is that the covariance of A_i and R_{ij} is zero. It follows, writing σ_Y^2 for $Var(Y_{ij})$, σ_A^2 for $Var(A_i)$ and σ_R^2 for $Var(R_{ij})$, that

[2.2] $\sigma_Y^2 = \sigma_A^2 + \sigma_R^2$

a very important result to which we shall have reason to refer frequently throughout the report.

Equation [2.1] above has been used to model many situations of interest to measurement specialists and assessment practitioners, including the reliability of potentially parallel tests, or the consistency of raters' judgements. It is also the standard one-way random effects analysis of variance (ANOVA) model (see, for example, Chapter 13 of Snedecor & Cochran, 1989). We shall see how all these different views on the same basic equation fit together into a theory of measurement reliability.

2.2 Classical reliability

Rewriting [2.1] with X_{ij} for Y_{ij}, T_i for $\mu+A_i$ and E_{ij} for R_{ij}, we arrive at the familiar equation

[2.3] $$X_{ij} = T_i + E_{ij}$$

which underpins all of so-called 'classical' test theory. In [2.3], X_{ij} conventionally represents the observed score of candidate i on test j, T_i the latent, unobservable 'true score' of candidate i, and E_{ij} the 'error' involved in measuring candidate i relative to test j.

We have covered the history and applications of [2.3] extensively in our companion report (Johnson & Johnson, 2009), so will cover only the main points briefly here. For the purposes of this short development we revert to the notation of [2.1], in the interest of consistency with the rest of the report.

Using the standard assumptions, it can be shown relatively easily that the squared correlation between the observed score and the true score is equal to the ratio of the true score variance to the total variance (Lord & Novick, 1968, p.57). Using our notation, we have

[2.4] $$\rho_{YA}^2 = \frac{\sigma_A^2}{\sigma_Y^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_R^2} = 1 - \frac{\sigma_R^2}{\sigma_Y^2}$$

Intuitively, it is reasonable to assert that the closer a candidate's observed score comes to that candidate's true score the more trustworthy we can assume the test will be, so that the (squared) correlation between true score and observed score could, in principle, be a useful indicator of test reliability.

From another perspective, we can think of the same quantity as being a measure of the proportion of variability in the observed score which is not due to measurement error.

Thus, [2.4] gives a powerful insight into the nature of test reliability (assuming we accept [2.1] and the standard assumptions), allowing us to reason interchangeably in terms of score correlations and score variance.

An alternative perspective on test reliability came from the notion of *parallel tests*. Suppose we have two measurements Y and Y', which have the same mean (true score) and the same variance, each otherwise satisfying the standard assumptions. Then it can be shown that the correlation between the two measurements Y and Y' is also equal to the squared correlation between the observed score and the true score (Lord & Novick, 1968, p.58). In symbols

[2.5] $$\rho_{YA}^2 = \rho_{YY'}$$

It follows from [2.5] that

[2.6] $$\sigma_A^2 = \sigma_{YY'},$$

the covariance between two parallel measurements is equal to the potentially unobservable variance between true scores.

Using the idea of parallel tests to develop a methodology for manipulating true scores is, of course, one thing. Actually determining how to construct a pair of parallel tests is another. A variety of suggestions have been made over the years about possible strategies for achieving parallel tests, with more or less success. However, the principle proved a very fruitful one, and the construction of reliability indices based on the sample correlation between parallel sets of questions came to be a staple of measurement practitioners over many decades.

2.3 The intraclass correlation

We have already observed that the model [2.1] implies that the observations Y_{ij} are grouped together into different instantiations, or *levels*, of the variable *A*, designated by values of the subscript *i*.

Consider any two distinct observations on model [2.1], say Y_{ij} and Y_{ik}. It is not difficult to show that, because of the linear independence assumptions, the covariance of Y_{ij} and Y_{ik} is equal to σ_A^2, the variance between groups, while they each have the same variance $\sigma_A^2 + \sigma_R^2$. Their correlation, known as the *intraclass correlation*, is thus

$$[2.7] \qquad \rho_I = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_R^2}$$

Thus, the more that observations from the same level of *A* are correlated, the higher will be the value of ρ_I.

This is effectively the same result as [2.4]. Indeed, as pointed out, *inter alia*, by Shrout and Fleiss (1979), most reliability coefficients are actually versions of the intraclass correlation.

The intraclass correlation was first described by Fisher (1925). It has since been widely used in studies on inter-rater reliability, though for much of the time relatively independently of the long tradition of work on test reliability as formulated in the previous section.

In fact, Coefficient Alpha, probably the most extensively reported measure of test reliability (*cf* Cortina, 1993, p.98; Hogan, Benjamin & Brezinski, 2000), is itself a form of intraclass correlation. Although this fact is not evident from Cronbach's (1951) original exposition, it was eventually recognised by Cronbach himself (see, for example, Cronbach & Shavelson, 2004, p.396).

2.4 ANOVA estimation

The chapter in which Fisher (1925) described the intraclass correlation (Chapter 7 in our version, which is the 1946 10th edition) also introduced the method which later came to be known as the *analysis of variance*, now universally abbreviated as *ANOVA*. Interestingly enough, Fisher's first use of the analysis of variance was as a device for computing estimates of the components of variance, used in their turn for estimating expressions of the form of [2.7].

ANOVA estimation is of fundamental importance to the understanding of the methodologies employed throughout the rest of this report, so it is worth dedicating a few paragraphs to it here.

First, we need some more notational conventions. We start with a sample of observations Y_{ij} from the model [2.1] , where we observe a levels from the variable A, with n observations drawn from each level, giving a sample of total size an. Thus, $1 \leq i \leq a$ and $1 \leq j \leq n$. To denote aggregates over a given subscript, we write a dot in the place of that subscript. As is usual, we use a bar above the variable name to represent averaging. So we can write

[2.8a] $\overline{Y}_{\cdot\cdot} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j}^{n} Y_{ij}$, the average of all an observations,

[2.8b] $\overline{Y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{i\cdot}$, the average of all n observations with the same value of i.

Just as in [2.2.] we can decompose the (population) variance, σ_Y^2 , so we can break down the sum of squared deviations of a sample of Ys from their sample mean $\overline{Y}_{\cdot\cdot}$ into a component based on the averages for each level of A and everything else.

[2.9] $$\sum_{i}^{a}\sum_{j}^{n}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i}^{a}\sum_{j}^{n}(Y_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 + \sum_{i}^{a}\sum_{j}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2 ,$$

because the cross-product term reduces to zero:

[2.9a] $$\sum_{i}^{a}\sum_{j}^{n}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)\left(Y_{i\cdot} - \overline{Y}_{\cdot\cdot}\right) = 0$$

It will be helpful to abbreviate the names for these sums of squares to

[2.10a] $SSA = \sum_{i}^{a}\sum_{j}^{n}(Y_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$

[2.10b] $SSR = \sum_{i}^{a}\sum_{j}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2$

[2.10c] $SST = \sum_{i}^{a}\sum_{j}^{n}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2$

SSA is often called the *between groups* sum of squares, and SSR the *within groups*, or *residual*, sum of squares. SST is the total sum of squares.

Taking expected values of [2.10] we find that

[2.11a] $\mathfrak{E}\, SSA = (a-1)(n\sigma_A^2 + \sigma_R^2)$

[2.11b] $\mathfrak{E}\, SSR = a(n-1)\sigma_R^2$

[2.11c] $\mathfrak{E}\, SST = (an-1)\sigma_Y^2$

Finally, we remove the expectation operators from the left hand sides, and treat the variances as if they were the corresponding estimators. This yields

[2.12a] $\hat{\sigma}_R^2 = \dfrac{SSR}{a(n-1)}$

[2.12b] $\hat{\sigma}_A^2 = \dfrac{SSA/(a-1) - \hat{\sigma}_R^2}{n}$

The estimators in [2.12], known as ANOVA estimators, have the useful property of being unbiased (by definition, because their expectation is defined as the corresponding population parameter).

Given suitable estimators for the variance components of [2.2], we can substitute them for their population counterparts in [2.7] to provide a sample estimate of the intraclass correlation, and hence, equivalently, of the simple classical reliability coefficient, $\hat{\rho}_I$.

[2.13] $\hat{\rho}_I = \dfrac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_R^2}$

2.5 Extending the model

The simple model [2.1] is very constrained, effectively restricting the range of applicable situations to those involving a single grouped variable (candidates in a test, scripts assigned to a marker, questions on a paper, and so forth).

However, given that [2.1] can be considered as a simple case of a random-effects ANOVA model, there is no reason in principle why the right hand side should not be extended to include more than one variable, as well as interactions between them.

From Chapter 3 onwards we introduce a variety of models involving two or more factors. To facilitate the discussion, we introduce here some of the standard ANOVA terminology. In the ANOVA, the right-hand-side variables are generally called *factors* or *effects*, and their possible values are known as *levels*. In the random-effects model, the observed levels of a factor are considered to be sampled from a very large universe of possible levels (which, like potential test items, might not all pre-exist). It is also possible for the levels of a factor to comprise a relatively small, fixed and predetermined set of values, like types of examination centre (school, college, workplace, …), or question formats (multiple choice, short answer, extended response, …). Factors of this type are called *fixed effects*, or sometimes *fixed factors,* when all of the few possible levels are included in the model, or *finite random factors* when some of the few available levels are sampled. Models which include both fixed and random effects are called *mixed models*. Because of the close association, initiated by Fisher, of ANOVA methodology with the development of the theory and practice of the design of experiments, the terms *model* and *design* are often used interchangeably, as they are in this report.

2.6 Unbalanced data

In Section 2.4 we reviewed the definition of ANOVA estimators for the variance components in a simple one-way random effects model with *n* observations at each level of the single factor. In reality, though, we cannot always guarantee that the number of observations per level will be the same – this might be by design or

because data are missing. Datasets with equal numbers of observations at each level are called *balanced*; those where this is not the case are called *unbalanced*. The problem generalises beyond single-factor models to any number of factors and their interactions.

Unfortunately, the technique of ANOVA estimation, equating sums of squares to their expected values, does not work for unbalanced data. When data are unbalanced the number of possible estimation equations exceeds the number of parameters to be estimated, and there are no clear criteria for choosing a suitable subset of the estimation equations – a neat summary of the problem is given by Verhelst (2000). Most ANOVA-style estimation methods applied to unbalanced data in use today are due to Henderson (1953). The Henderson methods are examined in considerable detail in Chapter 5 of Searle, Casella and McCulloch (2006).

Although they are in principle not unique, the Henderson estimators do have some ostensibly desirable properties. Like all ANOVA estimators they are unbiased. Unlike other estimation methods, notably those which depend on maximising likelihoods, they require no distributional assumptions. And because they have closed form analytical solutions they can be computationally very efficient, and produce results very fast.

The Henderson methods, however, may not be applicable in all cases. They can be particularly problematic when the model includes fixed factors or when the data are unbalanced with respect to nesting. They also have the slightly disturbing property that estimates of variance components – population parameters which by definition are always positive – can be negative. In effect, this property, which goes hand in hand with unbiasedness, is relatively easy to understand intuitively. For if the estimator is unbiased then sometimes it must yield values which are less than the parameter being estimated. If the true value of that parameter is very close to zero, we should not be surprised if occasionally we come up with estimates which are negative.

If for some reason we cannot, or do not wish to, use ANOVA estimation, we must rely on other methods which have no closed, analytic solution, and so rely on iterative techniques to try to converge on a stable estimate. The estimation strategies which have traditionally been favoured for linear modelling problems are based on maximising a likelihood function of the parameters of interest in the light of a particular set of observations. Two major issues might arise with this, and similar, approaches.

One is that, in order to specify the likelihood function, we need to make distributional assumptions about random variables in the model, assumptions which are not necessary for ANOVA estimation. The default assumption is of normality, which may not always be appropriate for the data in question.

The second issue has to do with computation time. It is certainly true that computer hardware is increasing in power all the time, even as numerical techniques are also becoming more sophisticated and efficient. Nonetheless it is still the case that, while solving the ANOVA estimation equations, even for very large data sets, can be virtually instantaneous on a modern desktop workstation, an iterative solution using general purpose software on the same data and the same computer can take hours (or

even days!). We consider briefly the question of suitable software in the next, and final, section of this chapter.

2.7 Software

Estimating examination reliability is then, in the perspective of this report, a question of computing a ratio of linear combinations of variance component estimates. While this sounds simple enough, in reality there are many practical difficulties.

In the first place, once we go beyond the simple one-way random effects model, there is more to a reliability coefficient than the ratio of 'true score' variance to total variance. We need to study the testing situation carefully, so as to determine which components contribute to true score variance, which count potentially as 'error' or 'noise' and which can be discarded altogether. Because each examination situation is potentially unique, we cannot necessarily rely on finding a ready-made 'cookbook' solution that can be applied straight out of the box.

We need also to be conscious of the fact that some models are more tractable computationally than others. It may be, for example, that the ideal model for a particular set of examination data might be too difficult to set up, or that the associated computation is too lengthy or too heavy for the computer available.

There are two software packages, both in the public domain, which are designed to be used for computing reliability coefficients on the basis of a restricted set of linear models. These are (a) various versions of GENOVA (Crick & Brennan, 1983; Brennan, 2001b; Brennan, 2001c) and (b) EduG (Cardinet, Johnson & Pini, 2010).

GENOVA, with its variants urGENOVA and mGENOVA, is a package of freeware programs designed by Robert Brennan to carry out a range of generalizability analyses for quite a large class of models. The theory behind GENOVA is set out in comprehensive fashion in Brennan (2001a). The software, which runs on Windows PCs and Macintosh Power PCs, is downloadable from http://www.education.uiowa.edu/casma/computer_programs.htm.

All of the GENOVA suite of programs estimate linear model parameters using ANOVA estimation, either orthodox ANOVA for balanced designs or Henderson's method I for certain unbalanced models. They all produce a substantial amount of information, including standard ANOVA tables and model parameter estimates, as well as reliability coefficients and selected *what if?* analyses ('D-studies' in generalizability theory terminology) as described in Brennan (2001a).

GENOVA works on balanced, complete designs (those where all interactions are specified) only; urGENOVA can also handle unbalanced, complete designs for random-effects models; mGENOVA implements so-called *multivariate generalizability*, where a factor in a limited class of balanced or unbalanced random-effects designs can be crossed with the levels of a fixed factor.

The GENOVA programs can handle data sets of unlimited size with extremely fast processing times. Provided the model you want is available they offer an extremely efficient solution, as well as producing automatically not just variance component estimates but also various reliability coefficients. However, they make little or no

concession to user-friendliness. They are all written in Fortran and use a command-line interface which is still described in the documentation in terms of 'control cards'. They have acquired over the years a reputation, somewhat undeserved, of being difficult to use.

EduG was designed as part of a project to popularise generalizability theory, in particular by providing software which would be easier to use than GENOVA. It handles a narrower range of models than the full GENOVA suite, effectively the same balanced, complete designs as GENOVA itself, with none of the extra features of urGENOVA or mGENOVA. Like GENOVA it uses ANOVA estimation and is consequently very fast. It has a more up-to-date and more forgiving user interface than GENOVA. For balanced data sets it is probably preferable. EduG is available as freeware at http://www.irdp.ch/edumetrie/englishprogram.htm.

If GENOVA or EduG are for some reason not suitable, there exist many general-purpose statistical packages which have routines designed for the extraction of variance components. SPSS, for example, has GLM, which uses ANOVA estimation, and MIXED, as well as VARCOMP, a subset of MIXED, which uses (restricted) maximum likelihood. SAS, similarly, has VARCOMP, with options for ANOVA estimation or a variety of iterative methods. The public domain software R (R Development Core Team, 2005) provides a number of packages for treating linear models, notably the mixed-model package `lme4`. Another useful option for handling variance component estimation from linear models with nesting is MLwiN, though its treatment of complete (i.e. fully crossed) designs could be somewhat laborious (Rashbash, Steele, Browne & Goldstein, 2009, Section 18.3).

These alternative packages for the most part use iterative solutions to converge on a set of variance component estimates. Our experience has been that they are considerably slower and more resource hungry than the ANOVA-based GENOVA and EduG, as well as requiring users to construct their own reliability coefficients from estimated components of variance. All of the analyses described in the main body of this report were carried out using EduG and/or GENOVA.

3 The variance analysis approach: generalizability theory

3.1 Partitioning variance and quantifying contributions

In operational examination situations the number of candidates taking any particular unit will count in the hundreds for low-entry subjects and in tens of thousands for high-entry subjects. For the purposes of illustrating some possible G-theory applications we begin by considering a modest response dataset, which emanated from a random subsample of the candidates entered for a GCSE history examination in 2007. The 2-section examination paper had a time allowance of 1 hour 45 minutes. Depending on their period of study, candidates were to answer three multi-part constructed response questions, each worth 25 marks for a 75-mark paper total. One question was compulsory while the other two were chosen by the candidates from three given options. All 30 candidates in the subsample to be considered here responded to the same three questions.

The candidate (or script) subsample formed the basis of a marker standardisation exercise, in which all candidates' responses were independently marked by a total of 95 individuals: the principal examiner, who set the paper, five team leaders, and 89 assistant examiners. The marking study was actually more than a regular standardisation exercise, since it was also designed to compare conventional with electronic marking. Just under half the assistant examiners marked the scripts in conventional paper format while the rest marked script images electronically. The paper was not tiered, candidates' performances being assumed to indicate appropriate grades across the full A to G grade range. The explicit and detailed mark scheme, which had been devised by the principal examiner with subject specialist consultation, was reviewed by the principal examiner and the team leaders before use in the marking study, and where necessary tightened. In the study, markers were instructed first to identify an appropriate 'level' for each subquestion response, using a best fit level description scheme, and then to award an appropriate mark for the response from within the given narrow mark range for the level.

The outcome of the independent marking was a 360 by 95 matrix of candidate-subquestion by marker scores. Variation in both the relative performances of the candidates and the relative 'difficulty' (for this group of candidates) of the questions would account for much, though not all, of the observed variation in the dataset as a whole. Other contributions would arise from the influence of interaction between candidates and questions, inter-marker and intra-marker variation, unidentified 'hidden' factors and random fluctuations. The essence of G-theory is to quantify the contributions of identifiable factors to the total observed score variance, so that this information can be used (a) to estimate the apparent reliability of this particular examination (through a 'G-study' analysis), and (b) to predict how reliability might change should candidates of similar type be required to answer more, or fewer, questions of similar style in a future examination, and/or be marked by a single marker or independently by several different markers (*what if,* or 'D-study', analysis).

For purposes of G-theory exposition we here selectively use subsets of the whole data matrix to illustrate analysis models, or designs, of varying degrees of sophistication. We start with the question-level data for a single one of the 95 markers. The smaller dataset is a 30×3 matrix of candidate-question marks (or scores), varying in value in

the range 0 to 25. The design in this case is represented by $c \times q$ (or cq), where c represents candidates, q represents questions, and $\times$ indicates that candidates are 'crossed' with questions, i.e. that all the candidates attempt all the questions. Both candidates and questions are considered to be 'random' factors, in the sense that the candidates and questions that actually feature in this particular examination are in theory random samples of those that might have featured in the present, the past or the future.

The mark or score for candidate c on question q, which we denote as Y_{cq}, can be expressed as a linear function of candidate, question and confounded residual effects as follows:

[3.1] $$Y_{cq} = \mu + (\mu_c - \mu) + (\mu_q - \mu) + (Y_{cq} - \mu_c - \mu_q + \mu)$$

where μ is the overall mean candidate-question score, $(\mu_c - \mu)$ is the 'effect' associated with candidate c, i.e. the deviation of candidate c's mean question score from the overall mean score; $(\mu_q - \mu)$ is the effect associated with question q, i.e. the deviation of question q's mean candidate score from the overall mean score; and the remaining term is the confounded residual effect – confounded by virtue of the fact that we have here a repeated measures design, in which there is just one single candidate-question score in each cell.

Representing the candidate effect as A_c, the question effect as B_q, and the confounded interaction effect as $(AB)_{cq,e}$, we can rewrite equation [3.1] as:

[3.2] $$Y_{cq} = \mu + A_c + B_q + (AB)_{cq,e}$$

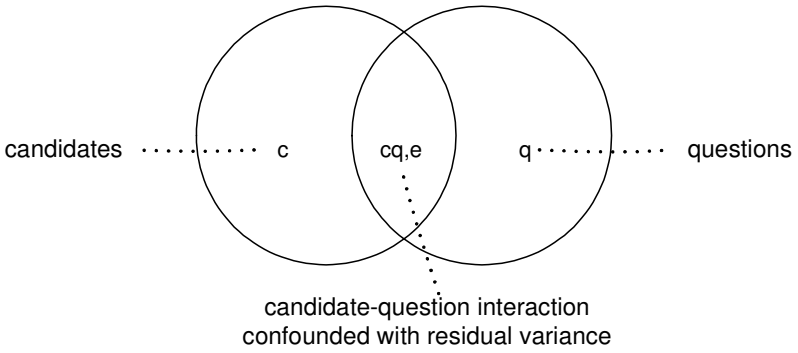
If we now make the usual assumption that all effects are linearly independent, so that all covariances on the right hand side are equal to 0, then by subtracting μ from both sides of equation [3.1], and squaring and summing the squares on both sides, we can partition the total variance, σ_Y^2, into between-candidate variance, σ_c^2, between-question variance, σ_q^2, and confounded residual variance, $\sigma_{cq,e}^2$ (as illustrated in Figure 3.1):

[3.3] $$\sigma_Y^2 = \sigma_c^2 + \sigma_q^2 + \sigma_{cq,e}^2$$

Note that in $\sigma_{cq,e}^2$ the 'e' represents contributions to the residual variance from unidentified 'hidden' factors as well as random fluctuations.

Using ANOVA we can now quantify the different variance component estimates in the expression for σ_Y^2, as shown in Table 3.1 (note that variance component estimates, like all sample-based estimates, will themselves be subject to error; for further details see Brennan 2001a, Chapter 6). The ANOVA table for the subset of 90 candidate-question scores (30 candidates by three questions) is shown as Table 3.2. It is the estimated variance components that are in turn used to estimate measurement errors and reliability coefficients.

Figure 3.1
Partitioning of total score variance* for the 2-factor design c × q



** Note that the residual variance comprises contributions from all unidentified 'hidden' factors as well as random fluctuations*

Table 3.1
ANOVA table* for the $c \times q$ design

Source of variance	*SS*	*df*	*MS*	$\hat{\sigma}^2$
Candidates	SS_c	n_c-1	$MS_c = SS_c/(n_c-1)$	$\hat{\sigma}^2_c = (MS_r - MS_c)/n_q$
Questions	SS_q	n_q-1	$MS_q = SS_q/(n_q-1)$	$\hat{\sigma}^2_q = (MS_r - MS_q)/n_c$
Confounded residual	SS_r	$(n_c-1)(n_q-1)$	$MS_r = SS_r/[(n_c-1)(n_q-1)]$	$\hat{\sigma}^2_r = MS_r$
Total	SS_T	$n_c n_q$-1		

** For notational convenience we here substitute with r the more explicit, but more cumbersome, cq,e in the confounded residual terms. The circumflex diacritics ('hats') in the last column indicate that the variance components are ANOVA estimators.*

Table 3.2
ANOVA results for 30 candidates attempting 3 questions

Source of variance	*SS*	*df*	*MS*	$\hat{\sigma}^2$	*% contribution**
Candidates	1788.3222	29	61.6663	18.2529	71
Questions	49.3556	2	24.6778	0.5923	2
Confounded residual	400.6444	58	6.9077	6.9077	27
Total	2238.3222	89			

** Percentage contributions of the estimated variance components to the total variance, where the total variance is the sum of the components.*

Note in Table 3.2 the very high percentage of total variance that can be attributed to between-candidate variation, and the very low contribution of between-question variance. The average test score (0-75 mark scale) for the 30 candidates marked by this one marker was 46.6 with a standard deviation of 13.6. The mean question score per candidate was 15.5, with a standard deviation of 4.5 and a range of 6.3 to 23.7. The mean question score for the three questions over the 30 candidates was also 15.5, but with a relatively small variation in marks from one question to the other: 16.0, 14.5 and 16.0. The confounded residual accounts for over a quarter of the total observed variance. Much of this contribution can be assumed to be attributable to

candidate-question interaction, i.e. to inconsistency in the performances of individual candidates over the three test questions.

The relative sizes of the different variance components will vary from one test paper to another and from one candidate group to another, depending on the composition of the candidate group and the nature of the set of test questions put to them. In this particular example the test paper succeeded in spreading candidates across the mark scale, and the three questions which had been set and used in the live examination without any form of pretesting showed little variation in relative difficulty for that candidate group, reflecting both the question setting skill and experience of the principal examiner and the relative stability in the group characteristics of candidate entries from year to year (in terms of history ability and examination preparedness).

But now let us see how the variance component information in Table 3.2 is used in reliability estimation, always recognising that three test questions is rather a small number to use as a basis for such estimation and for subsequent generalisation.

3.2 Relative candidate measurement: 2-factor crossed model

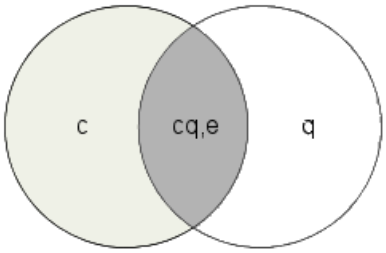
When we calculate a reliability coefficient we first identify what we consider to be 'true score' or 'valid' variance, and simultaneously what we consider to be the contributions to measurement error variance. A reliability coefficient is then simply the ratio of valid variance to the sum of valid and error variance. In other words, the coefficient once calculated indicates the proportion of 'total' variance that is valid variance (note that this total variance is not necessarily synonymous with the total variation in the original matrix of candidate-question scores, the 'observed score' variance, since some sources of score variation contribute neither to valid nor to error variance).

When we are focusing on how well we are measuring examination candidates the valid variance will be the between-candidate variance, $\hat{\sigma}_c^2$. In the simple $c \times q$ design there are only two other contributions to the total variance. These are the between-question variance and the confounded residual variance, which subsumes candidate-question interaction.

If the purpose of the measurement is to rank candidates relative to one another on the measurement scale (i.e. on the 0-75 total test score scale in this case) then the difficulty or easiness of the three questions will have no part to play in measurement error, leaving only the confounded residual variance to worry about. This situation is illustrated in Figure 3.2, the variance attribution diagram, in which the sector representing the between-question variance is unshaded to indicate its passive presence in the total observed score variance.

The estimated error variance for this situation of relative candidate measurement is $\hat{\sigma}_r^2 / n_q$, or $\hat{\sigma}^2(\delta)$ in G-theory notation. This is the usual expression for the variance of a sample mean, the sample mean in this case being candidate c's mean question score. In this example $\hat{\sigma}^2(\delta)$ = 6.9077/3 = 2.3.

Figure 3.2
Valid, passive and measurement error variance for relative candidate measurement for the design $c \times q$



** Light shading indicates valid variance and darker shading the source of measurement error variance; passive variance, that contributes neither to valid nor to error variance, is unshaded.*

G-theory provides its version of a reliability measure, a form of intraclass correlation called the generalizability coefficient, notated by Cronbach *et al* (1972) as $\mathcal{G}\rho^2$, which for the simple 2-factor crossed model is identical with Cronbach's α. In this report we use the simpler notation Γ (gamma), rather than $\mathcal{G}\rho^2$, to denote the generalizability coefficient. In the basic $c \times q$ model Γ is given by:

[3.4] $$\Gamma = \hat{\sigma}_c^2 / (\hat{\sigma}_c^2 + \hat{\sigma}_r^2 / n_q) = 0.89$$

So, the generalizability coefficient for relative measurement is in this case a high 0.89, indicating that 89% of the observed score variance can be attributed to valid, between-candidate, variance.

But what can we say about the likely precision of individual candidate scores? For this we need to compute the standard error of measurement, $\hat{\sigma}(\delta)$. In this example $\hat{\sigma}(\delta)$ = 1.52. This is the standard error of measurement for a candidate's mean question score, which we will denote more specifically as SEM_{ms}, to distinguish it from SEM_{ts}, which is the more appropriate SEM to consider when it is the precision of candidates' total test scores that is of concern. To compute SEM_{ts}, we simply multiply SEM_{ms} by n_q (because the variance of the sum of a set of independent variables is equal to the sum of the variances). Here our estimate of the variance of a generic candidate-question score is just the residual variance estimate $\hat{\sigma}_r^2 = \hat{\sigma}_{cq,e}^2$, so that the estimated variance of the sum of a candidate's question scores is $n_q\hat{\sigma}_r^2$, which is equivalent to $n_q^2(\hat{\sigma}_r^2 / n_q)$. The square root of this expression, $n_q SEM_{ms}$, is SEM_{ts}.

In this example, SEM_{ts} is 1.52 × 3, or 4.55. The corresponding margin of error is 8.92. Thus, despite the high alpha value we have margins of error around candidates' test scores that are over 10% the length of the 0-75 test score scale, giving 95% confidence intervals almost 25% of that length. In practice, margins of error will differ from one section of the scale to another. These conditional margins of error can be calculated in a number of ways (see Brennan, 2001a, Chapters 5 and 10; Raju, Price, Oshima & Nering, 2007), including by analysing the response data for candidates with test scores within any given range of values.

When we have response data from a sample of questions we can proceed to explore how changes in the size of that sample might impact on measurement error variance, generalizability coefficients and margins of error (the *what if?,* or D-study, analyses). For this simple model, in which the only variable whose sample size can be changed is 'questions', we need only to substitute different values of n_q in the relevant algebraic expressions to predict the new indicator values. Changing n_q from its current value of 3 to values of 4, 5 and 6 gives the estimates shown in Table 3.3.

Table 3.3
Estimated changes in Γ, SEM and margin of error of increases in numbers of questions* ($\hat{\sigma}_r^2 = 6.9077$)

No. questions	*Mark scale*	Γ	SEM_{ts}	ME_{ts}	ME_{ts} *as % of the mark scale*
3	0-75	0.89	4.6	9.0	11.8
4	0-100	0.91	5.3	10.4	10.3
5	0-125	0.93	5.9	11.6	9.2
6	0-150	0.94	6.4	12.5	8.2

** In the simple $c \times q$ design Γ is equivalent to α*

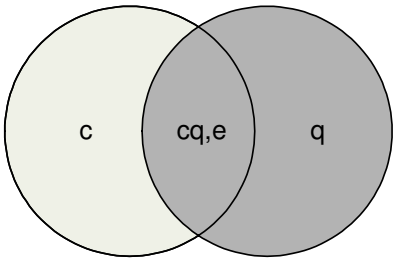
As Table 3.3 shows, increasing the number of test questions would result in an increase in the value of Γ, and would reduce the error margin as a percentage of the total test score scale. For example, doubling the number of questions from three to six, i.e. doubling the length of testing time per candidate, would increase the value of Γ, or α in this case, from 0.89 to 0.94. It would also reduce the margin of error by around three and a half percentage points in terms of the length of the test score scale: from 9 marks for a 3-question 0-75 mark scale (11.8% of the scale) to 12½ marks for a 6-question 0-150 mark scale (just over 8% of the scale). Whether this relatively modest increase in precision would justify doubling the testing time for candidates would be a question for debate.

3.3 Absolute candidate measurement: 2-factor crossed model

Whenever we are concerned with applying cut scores to mark distributions in order to classify candidates in merit terms, and when those cut scores are not determined *a priori* to divide a distribution into fixed proportions of candidates, then we are in the business of 'absolute' measurement. It is no longer sufficient to know how well we can distribute candidates relative to one another on the test score scale – we now need to know how much confidence we can place in the actual scores that individual candidates achieve, or, in other words, we need to know the degree of precision that we can attach to those absolute scores. At this point we can no longer ignore the levels of difficulty of the questions that we have used to form our test, unless those questions are the only ones that matter, which is rarely the case – if it were, then the same questions would be used in every examination in that subject. The questions, as before, implicitly represent a sample from some larger domain of questions that could have been based on the subject specification concerned and which could have been set and used in the particular examination paper.

We now, therefore, have two potential sources of measurement error, the confounded residual variance and the between-question variance. Figure 3.3 illustrates this new situation.

Figure 3.3
Valid variance and measurement error variance for absolute candidate measurement for the design c×q



** Light shading indicates valid variance and darker shading sources of measurement error variance*

The estimated error variance for this situation of absolute candidate measurement, i.e. $\hat{\sigma}^2(\Delta)$, is given by $\hat{\sigma}_q^2 / n_q + \hat{\sigma}_r^2 / n_q = (\hat{\sigma}_q^2 + \hat{\sigma}_r^2) / n_q$. From Table 3.2 we see again that $\hat{\sigma}_q^2$ has an estimated value of 0.5923, while $\hat{\sigma}_r^2$ has estimated value 6.9077. $\hat{\sigma}^2(\Delta)$, therefore, has value 2.5.

The 'absolute' G coefficient, Φ, is given by:

[3.5] $$\Phi = \hat{\sigma}_c^2 / (\hat{\sigma}_c^2 + \hat{\sigma}^2(\Delta)) = 0.88$$

In this case the absolute G coefficient barely differs in value from the relative coefficient, a fact explained by the very low between-question variation in this example. Had there been no between-question variation at all, or had the questions been considered the only important ones to set, making questions a 'fixed' factor with 'passive' variance, then the value of Φ would have equalled that of Γ (Φ can never have a value higher than Γ). As before, the square root of the error variance is the standard error of measurement for a mean candidate question score, i.e. SEM_{ms}. Multiply by n_q, here 3, and we find the SEM_{ts}, which is 4.74. The margin of error is therefore 9.29, only slightly higher than for relative measurement. Table 3.4 shows the likely effect of increases in question numbers on the G coefficient and error estimates.

Table 3.4 confirms that for this particular set of data there is an almost indiscernible difference in the results for absolute compared with relative measurement for this one marker. But we can only generalise the analysis results for this one marker, since another marker could have produced a different set of outcomes, and a third a different set again. In the next section we extend the model to look simultaneously at both question and marker impact on candidate outcomes and on assessment reliability.

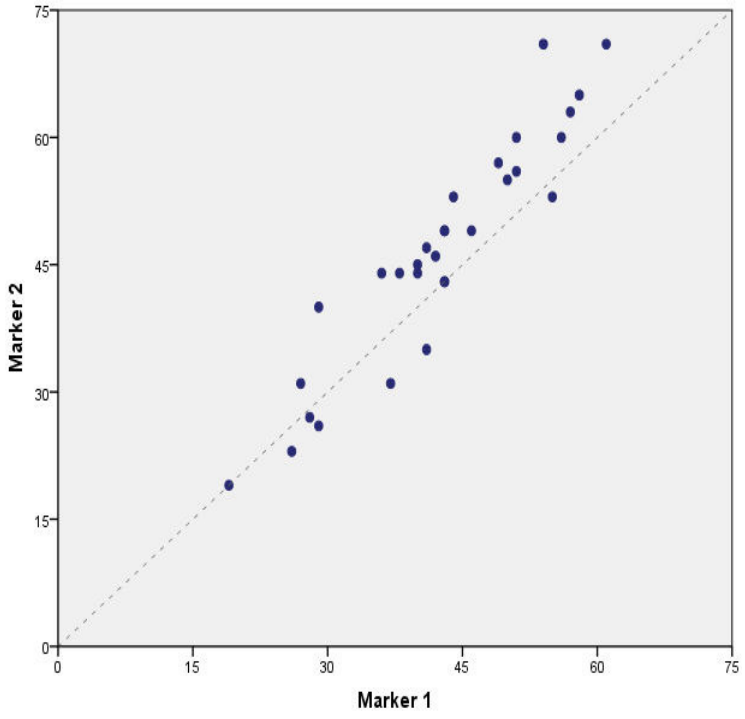
Table 3.4
Estimated changes in Φ, SEM and margin of error of increases in numbers of questions (with single marking)
($\hat{\sigma}_q^2 = 0.5923,\ \hat{\sigma}_r^2 = 6.9077$)

No. questions	*Mark scale*	Φ	SEM_{ts}	ME_{ts}	ME_{ts} *as % of the mark scale*
3	0-75	0.88	4.7	9.2	12.1
4	0-100	0.91	5.5	10.8	10.7
5	0-125	0.92	6.1	12.0	9.5
6	0-150	0.94	6.7	13.1	8.7

3.4 A 3-factor crossed model: candidates, questions and markers

Let us add another degree of realism to the GCSE History example, by considering the additional impact of differences in marker standards and marking consistency on assessment outcomes. Despite involvement in standardisation exercises we can expect different markers to exhibit greater or lesser degrees of difference in their overall marking standards and in their marking consistency. Figure 3.4, for example, compares the total marks given to each of the 30 candidates by two different markers selected at random from within the group of 40 markers who marked the paper-based scripts.

Figure 3.4
The mark allocations of two different markers for the 30 candidates
(two of the points represent two candidates each)



We see from Figure 3.4 that while the script rank orders produced by the two markers are roughly the same, which would surely be expected given the spread in marks, the fate of some of the candidates in terms of marks and grades would be quite different depending on which of the two markers their work had been marked by. The two most extreme examples are the candidate given just under 30 marks by marker 1 and around 40 marks by marker 2, and the candidate given around 55 marks by marker 1 and more than 70 marks by marker 2. There is clear evidence in Figure 3.4 that the marking standards of marker 2 were in general more lenient than those of marker 1 (this is inter-marker variation), with the exception of candidates in the bottom section of the mark scale. But marker 2's standards were even more lenient, or, equivalently, marker 1's standards were even more severe, for some of the candidates compared with others (this is intra-marker variation, or marker-candidate interaction).

The purpose of marker standardisation exercises is to reduce between-marker variation to a minimum. When such exercises are undertaken before live marking begins then any marker still operating after training outside some tolerance limit with respect to markers in general and to lead markers in particular are rejected. When ongoing monitoring of marker standards is carried out through script seeding then markers might again be rejected at any point, or their results might be adjusted up or down by an appropriate amount to bring their standards into line. It is much less easy, impossible even, to handle marker-candidate interaction in this way.

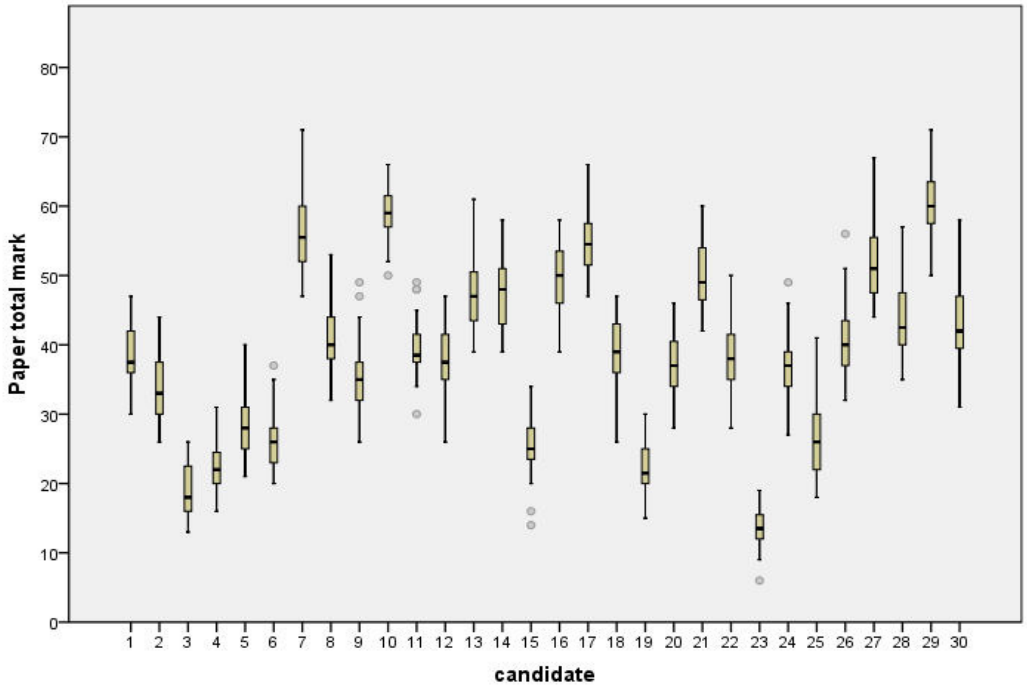
Even should all the marks awarded by marker 2 be adjusted downwards this would have little useful effect on the fate of many of the candidates. Several candidates would still have gained more marks had they been marked by marker 2 as opposed to marker 1, while several other candidates would have done less well. Whatever the size of the difference in marks this difference cannot be predicted when it varies across candidates. And for some candidates the difference could result in a different final grade award. There is no way that marker-candidate interaction can be detected in single-marker live marking. If such interaction is revealed in prior marker reliability studies, such as this one, and is large enough to warrant concern, then the only way to deal with it is to at least double mark candidates' responses to examination questions. Unless, of course, the interaction can be attributed to specific examination questions, in which case further standardisation for those particular questions could be useful.

Figure 3.5 shows the variations in total marks awarded to each of the 30 candidates by the 40 different markers (while this is the picture pre-standardisation the post-standardisation pattern was barely changed). We see in Figure 3.5 some quite wide mark spreads for some candidates, with no obvious relationship to overall test performances. This is again evidence that while candidate rank orders might be similar from one marker to another there are location shifts for some candidates from one marker to another within those rank orders: in other words, there are marker-candidate interaction effects at play here.

G-theory can make a useful contribution to the prior evaluation of marker effects, in particular for indicating which subject components would benefit from double or triple marking in place of single marking. For illustration we analyse the marking results for the 40 individuals who marked the scripts of the 30 GCSE History

candidates in the conventional way, i.e. on paper. For each of the 30 candidates we now have marks for each of the three test questions from 40 different markers.

Figure 3.5
Variations in marks awarded to candidates by the 40 independent markers



The analysis model becomes slightly more complicated than before, because in addition to potential question effects we now also have potential marker effects, i.e. inter-marker variation, as well as potential marker-question interaction effects and marker-candidate effects, resulting from intra-marker variation.

The linear model representing the score achieved by candidate *c* for question *q* as judged by marker *m* is now:

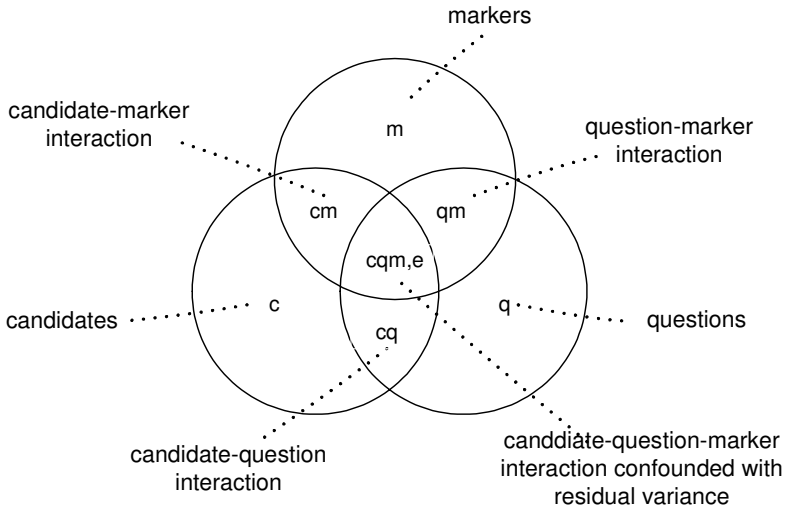
$$[3.6] \quad \begin{aligned} Y_{cqm} = {} & \mu + (\mu_c - \mu) + (\mu_q - \mu) + (\mu_m - \mu) \\ & + (\mu_{cq} - \mu_c - \mu_q + \mu) + (\mu_{cm} - \mu_c - \mu_m + \mu) + (\mu_{qm} - \mu_q - \mu_m + \mu) \\ & + (Y_{cqm} - \mu_{cq} - \mu_{cm} - \mu_{qm} + \mu_c + \mu_q + \mu_m - \mu) \end{aligned}$$

where μ is the overall mean candidate-question-marker score, $(\mu_c - \mu)$ is the candidate effect, $(\mu_q - \mu)$ the question effect, $(\mu_m - \mu)$ the marker effect, $(\mu_{cq} - \mu_c - \mu_q + \mu)$ the candidate-question interaction effect, etc.

Making the usual linear modelling assumption, i.e. that all effects are independent, we can show as before that the total variance can be partitioned into seven components as follows (and as illustrated in Figure 3.6):

$$[3.7] \quad \sigma_Y^2 = \sigma_c^2 + \sigma_q^2 + \sigma_m^2 + \sigma_{cq}^2 + \sigma_{cm}^2 + \sigma_{qm}^2 + \sigma_{cqm,e}^2$$

Figure 3.6
Variance partition for the 3-factor design c × q × m



Unless the markers who mark the candidates' work are the only ones that could do the job, then markers are by default sampled from a larger pool of potential markers, just as the test questions are implicitly sampled from a larger pool of potentially relevant questions. Both variables are therefore potential contributors to measurement error for candidate assessment. For *relative* candidate measurement it is only interactions between one or both of these two factors and candidates that are error contributors, whereas for *absolute* measurement the main effects, inter-marker variation and inter-question variation, as well as question-marker interaction also count. Figure 3.7 illustrates the situation for relative candidate measurement, while Figure 3.8 does so for absolute candidate measurement.

The expression for the estimated relative measurement error variance is now more complex than before, with three contributing terms:

[3.8] $\hat{\sigma}^2(\delta) = \hat{\sigma}^2_{cq} / n_q + \hat{\sigma}^2_{cm} / n_m + \hat{\sigma}^2_r / (n_q n_m)$

while the relative G coefficient is given as usual by:

[3.9] $\Gamma = \hat{\sigma}^2_c / (\hat{\sigma}^2_c + \hat{\sigma}^2(\delta))$

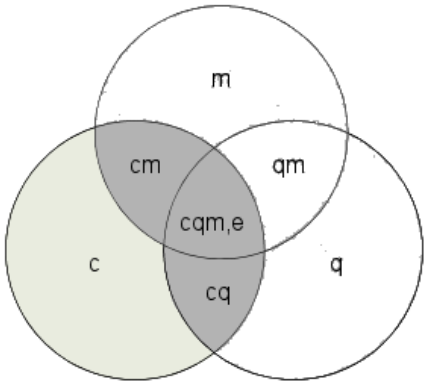
The expression for the estimated absolute measurement error variance is even more complex than that for its relative counterpart, with six contributing terms:

[3.10] $\hat{\sigma}^2(\Delta) = \hat{\sigma}^2_q / n_q + \hat{\sigma}^2_m / n_m + \hat{\sigma}^2_{qm} / (n_q n_m) + \hat{\sigma}^2_{cq} / n_q + \hat{\sigma}^2_{cm} / n_m + \hat{\sigma}^2_r / (n_q n_m)$

The absolute G coefficient is given by:

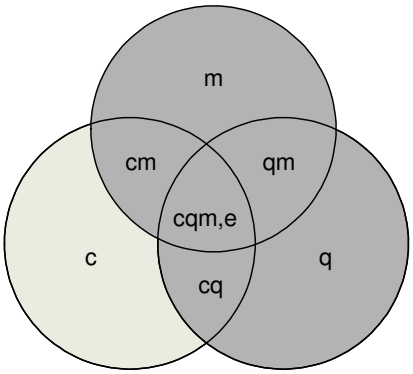
[3.11] $\Phi = \hat{\sigma}^2_c / \left(\hat{\sigma}^2_c + \hat{\sigma}^2(\Delta)\right)$

Figure 3.7
Valid, passive and measurement error variance for relative candidate measurement for the design c×q×m



** Light shading indicates valid variance, darker shading sources of measurement error variance, and no shading passive variance.*

Figure 3.8
Valid variance and measurement error variance for absolute candidate measurement for the design c×q×m



In order to calculate the values of measurement error and of reliability coefficients we need estimates of the contributing variance components – shown in Table 3.5.

Table 3.5
ANOVA results for 30 candidates attempting three questions, all candidate scripts marked independently by 40 markers

Source of variance	*SS*	*df*	*MS*	$\hat{\sigma}^2$	*% contribution**
Candidates	56726.3167	29	1956.0799	14.9181	62
Questions	2060.2217	2	1030.1108	0.7115	3
Markers	4575.0611	39	117.3093	1.1360	5
Candidates by questions	9555.7117	58	164.7536	4.0610	17
Candidates by markers	3927.0389	1131	3.4722	0.3858	2
Questions by markers	1085.4006	78	13.9154	0.3867	2
Confounded residual	5235.9994	2262	2.3148	2.3148	10
Total	83165.7500	3599			

** Percentage contributions of the estimated variance components to the total variance, where the total variance is the sum of the components.*

By introducing markers into the model we have been able to isolate some of the variance contributions that were hidden in the earlier analysis, but that contributed in one way or another to the main and interaction effects. In particular, while we see that, as before, by far the largest component is the between-candidate variance, with a 62% contribution to total variance, of the other components the candidate by question interaction variance is the most important, with a 17% contribution, followed by the confounded residual variance, with a 10% contribution. Interestingly, the between-marker variance and the interaction effects between markers and candidates and between markers and questions are relatively small – this might not have been the case for other GCSE subjects and papers.

Using the estimated variance component values given in Table 3.5 we find that the relative measurement error variance is 1.3826, with SEM_{ms} equal to 1.18, SEM_{ts} equal to 3.54 (note that SEM_{ts} is still given by $n_q SEM_{ms}$, because we are summing mean candidate-question-marker scores over questions only, and averaging over markers), and margin of error equal to 6.94. The value of Γ is 0.92. For this particular dataset the corresponding values for the absolute measurement error and of Φ are, respectively, 1.6514 and 0.90. The SEM_{ms} is 1.29 and the SEM_{ts} is 3.87.

These very positive values for reliability coefficients and SEMs are based on the results of averaging over 40 independent marker judgements. In live examining situations a single marker per script is the norm. So what can we say about likely test score precision for single marking? Table 3.6 provides the response: the new predicted values of reliability coefficients, SEMs and MEs have been produced simply by substituting different values of n_m (1 to 4) into the expressions above.

Table 3.6 looks at the impact of using between one and three markers per script, and shows that the greatest benefit is seen between single and double marking.

Table 3.6
Estimated changes in reliability coefficients, SEM_{ts} and margins of error of changes in numbers of markers

Relative measurement						
No. questions	*No. markers*	*Mark scale*	Γ	SEM_{ts}	ME_{ts}	ME_{ts} *as % of mark scale*
3	4	0-75	0.90	3.8	7.4	9.7
3	3	0-75	0.90	4.0	7.8	10.3
3	2	0-75	0.89	4.2	8.2	10.8
3	1	0-75	0.86	4.8	9.3	12.2
Absolute measurement						
No. questions	*No. markers*	*Mark scale*	Φ	SEM_{ts}	ME_{ts}	ME_{ts} *as % of mark scale*
3	4	0-75	0.87	4.4	8.6	11.3
3	3	0-75	0.86	4.6	9.1	12.0
3	2	0-75	0.84	5.0	9.8	12.9
3	1	0-75	0.79	6.0	11.8	15.5

Note also in Table 3.6 the predicted values of the reliability coefficients and other statistics for the case of single marking, compared with the results given in the

previous sections for analysis of the data for one only of the markers. The degree of fit between individual markers' actual results and those predicted for a generic single marker will naturally differ. The values for relative measurement for the particular marker featured earlier happen to be closely in line with the prediction from the larger G-study. For absolute measurement this one particular marker's performance is more positive than the generic marker prediction.

We can extend this *what if* analysis to embrace simultaneous changes in the numbers of markers and questions, as shown in Table 3.7 for combinations of one or two markers and three or six questions (although it should be noted that three questions is a very small sample to use for prediction – it would have been better if the original marking study had involved more questions).

Table 3.7
Estimated changes in reliability coefficients, SEM_{ts} and margins of error of changes in question *and* marker numbers

Relative measurement						
No. questions	*No. markers*	*Mark scale*	Γ	SEM_{ts}	ME_{ts}	ME_{ts} *as % of mark scale*
3	2	0-75	0.89	4.2	8.2	10.8
3	1	0-75	0.86	4.8	9.3	12.2
6	2	0.150	0.93	6.2	12.1	8.0
6	1	0-150	0.91	7.2	14.1	9.3
Absolute measurement						
No. questions	*No. markers*	*Mark scale*	Φ	SEM_{ts}	ME_{ts}	ME_{ts} *as % of mark scale*
3	2	0-75	0.84	5.0	9.8	12.9
3	1	0-75	0.79	6.0	11.8	15.5
6	2	0-150	0.89	8.0	15.7	10.4
6	1	0-150	0.84	10.0	19.6	13.0

Interestingly, while an increase in either marker or question numbers results in an increase in score precision for relative measurement, for absolute measurement the picture is less predictable. Here, double marking of a 3-question paper has the same impact on score precision as single marking of a 6-question paper, when scale length is taken into account. This difference between relative measurement and absolute measurement is attributable to the presence in absolute measurement error of the between-marker variation.

3.5 Nesting within marking teams: a hierarchical model

The 40 markers considered in Section 3.4 were actually divided into eight marking teams, with each team of five markers led by a different team leader. We can explore any influence that membership of a marking team might have on marking behaviour by incorporating marking team as a nesting variable for markers. Figure 3.9 illustrates the variance partition in this case.

Figure 3.9
Partitioning of total score variance for the 4-factor nested design c × q ×(m:t)

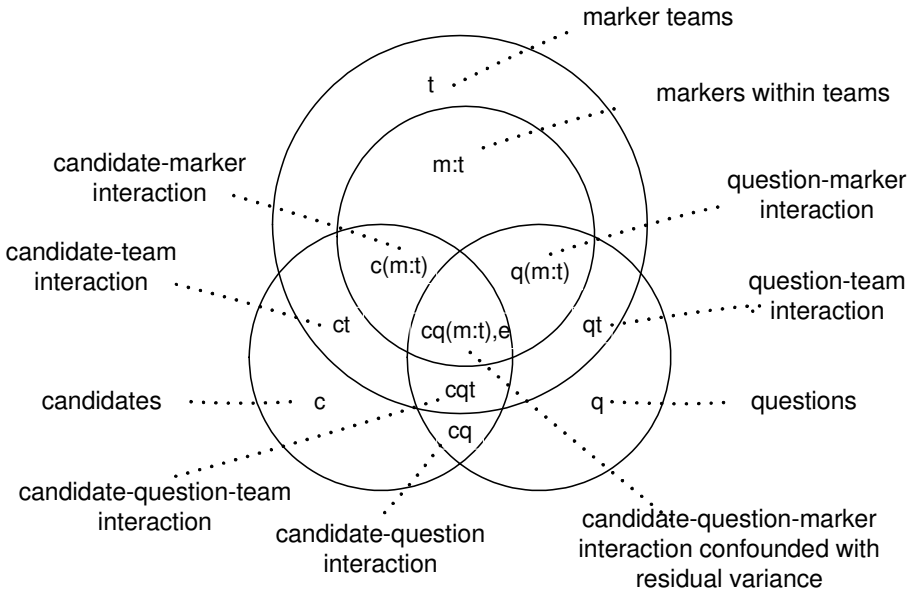


Table 3.8 presents the results of the analysis of variance for this nested design. The between-candidate variance is unchanged as the largest component, contributing 62% of the total variance. The candidate-question variance contributes another 17% to the total variance, with the confounded residual variance contributing another 10%. Other contributions are extremely small, particularly those involving marker teams, which are essentially non-existent.

The negative variance estimates in Table 3.8 are worthy of comment in this regard. Where there is no variation associated with a potential source of variance, i.e. when there is very little or no between-level variance for that factor or factor interaction, then the associated variance estimate will be extremely small and could be a short distance either side of zero. Alternatively, negative variance estimates can indicate that the analysis model is inappropriate and/or that there is insufficient data available to provide adequate variance estimates.

The negative variance estimates in this example are most likely attributable to zero components for marker teams and for interactions involving marker teams. Following common practice the small negative estimates are set to zero in follow-on computations (an argument could be made for doing the same with the very small positive components associated with the interactions between marker teams and questions and between marker teams and candidates).

Table 3.8
ANOVA results for 30 candidates attempting three questions, all scripts marked independently by five markers in each of eight marker teams

Source of variance	*SS*	*df*	*MS*	$\hat{\sigma}^2$	*% contribution**
Candidates	56726.3167	29	1956.0799	14.9126	62
Questions	2060.2217	2	1030.1108	0.7082	3
Marker teams	473.8478	7	67.6925	-0.1461	0
Candidate by questions	9555.7117	58	164.7536	4.0659	17
Candidates by marker teams	797.5856	203	3.9290	0.0531	<1
Questions by marker teams	246.3472	14	17.5962	0.0315	<1
Candidates by questions by teams	860.1194	406	2.1185	-0.0478	0
Markers within teams	4101.2133	32	128.1629	1.2671	5
Candidates by markers within teams	3129.4533	928	3.3723	0.3382	1
Questions by markers within teams	839.0533	64	13.1102	0.3584	1
Confounded residual	4375.8800	1856	2.3577	2.3577	10
Total	83165.7500	3599			

** Percentage contributions of the estimated variance components to the total variance, where the total variance is the sum of the components.*

If we consider not only questions and markers but also marker teams as simply samples representing larger groups, then in principle contributions to relative measurement error will come from interactions between candidates and one or more of the other factors, *viz*. questions, marker teams and markers. Thus, we have:

[3.12] $$\hat{\sigma}^2(\delta) = \hat{\sigma}^2_{cq} / n_q + \hat{\sigma}^2_{ct} / n_t + \hat{\sigma}^2_{cqt} / (n_q n_t) + \hat{\sigma}^2_{cm:t} / (n_m n_t) + \hat{\sigma}^2_r / (n_q n_m n_t)$$

For this example $\hat{\sigma}^2(\delta)$ = 1.394, giving a value for the relative generalizability coefficient, Γ, of 0.91, as before.

The absolute error variance is given by:

[3.13] $$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(\delta) + \hat{\sigma}^2_q / n_q + \hat{\sigma}^2_t / n_t + \hat{\sigma}^2_{qt} / (n_q n_t) + \hat{\sigma}^2_{m:t} / (n_m n_t) + \hat{\sigma}^2_{qm:t} / (n_q n_m n_t)$$

The value of $\hat{\sigma}^2(\Delta)$ is 1.8324, giving a value for the absolute generalizability coefficient of 0.89.

As usual, once we have the set of estimated variance components we can not only calculate the relative and absolute error variances, generalizability coefficients and SEMs, but we can plug different values for factor sample sizes into the relevant expressions and predict likely changes. Chapters 4 and 5 will illustrate this more fully, through a variety of applications to 2009 operational examination data.

4 Applications of the basic *post hoc* c × q design

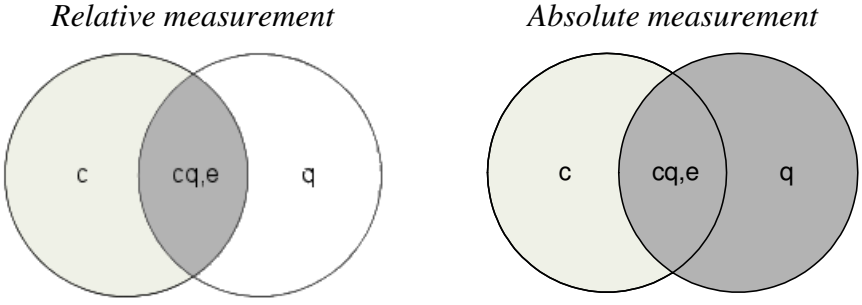
4.1 Introduction

The application examples described in this chapter and in Chapter 5 all feature written component papers set in 2009. The datasets underpinning the applications were deliberately selected to span a variety of component structures and of examination subjects, with the expectation that patterns of impact on reliability of test-based and candidate-based factors might differ from one example to another. Unfortunately, since single marking is the norm currently in live assessment, written or practical, none of the analyses described here offers the possibility of commenting on marker influences on measurement error – hence the chapter title. With the exception of objective tests, therefore, markers are a hidden factor whose importance cannot be quantified, but whose impact can be expected to be higher in some subject components than others.

The basic c × q design is ideal as a *post-hoc* screening device where single marking has been used – where multiple marking is employed then it should be replaced with the c × q × m design. Should analysis of a dataset result in high values for both the relative and the absolute generalizability coefficients, say around the conventionally agreed 0.8 or higher, then there would be no compelling need to carry out more complex mixed-model analyses to explore possible interaction effects.

Included here as an *aide mémoire*, Figure 4.1 illustrates the general partition of total observed candidate-question score variance into the three constituent variance components that can be quantified in this simple design: between-candidate variance, between-question variance, and the confounded residual variance.

Figure 4.1
Contributions to relative and to absolute measurement error in the c × q design*



** In the c × q design, c represents candidates and q questions. In the diagrams the letters represent the associated variance components. Light shading indicates valid variance (the focus of the measurement), dark shading identifies those components that contribute to measurement error, and no shading indicates sources of passive variance, i.e. score variation that contributes neither to valid nor to error variance.*

The majority of the analyses in this chapter, which are ANOVA-based and which were carried out using EduG (see Chapter 2, Section 2.7, for access details), involve question scores only, even though in some cases subquestion scores were electronically recorded and made available to us. The additional information supplied

by subquestion scores, along with the added analysis complexity associated with mixed-model designs involving nesting, is explored in Chapter 5, in which we extend and elaborate some of the analyses described in this current chapter.

Before reviewing the results of the analyses readers are reminded that each of the papers considered here was just one component, or even one section of one component, in multi-component examinations. The other examination components might be written or practical, or both. In some cases all components would have been attempted by candidates in the same period whereas in other cases, depending on the structure of the examination concerned and on the candidate's predilection, they might have been attempted at different times of the year or even in different years. Whatever the number of components that constituted the examination, and whenever each component paper might have been attempted, the fact is that there would have been other attainment evidence available for candidates that would have supplemented and complemented the attainment evidence from the single component papers analysed here. This should be borne in mind as the analysis results are reviewed.

Note that in all tables SEM and ME refer, respectively, to the standard error of measurement associated with a candidate's total mark on the component paper and the margin of error with which a 95% confidence interval around that total mark would be constructed.

4.2 GCSE Business Studies (equally weighted questions)

Our first example is a 2-hour higher tier GCSE Business Studies written paper that comprised five compulsory 20-mark structured questions. An additional five marks were available in the paper total for quality of written communication. The marks for quality of written communication were dropped for convenience, so that we have a notional test score scale of 0-100 marks for the 5-question paper. Subquestion marks were not included in the dataset for the paper, so that the analysis described here was based on question scores only.

The response dataset for the paper contained records for just under 2,000 candidates, 44% of whom were female. Almost all the candidates, 92%, were entered from secondary comprehensive schools, with another 5% from secondary independent schools. Figure 4.2 shows the relatively symmetric test score distribution for the paper. Note that the bottom fifth of the intended mark scale was in practice unused, the distribution centring on an average paper mark of 61.4 for an achieved mark scale of 20-100. There was no significant gender difference in average paper performance (males 61.5, females 61.3), and there was relatively little variation in question mean scores, which ranged from 11 (55%) to 13.7 (68.5%). In a principal components analysis the first principal component accounted for almost 65% of the total candidate-score variance.

The G-study results for the paper are given in Table 4.1. The first point to note in Table 4.1 is the large contribution of between-candidate variance (just over 50%) to total variance (i.e. to the sum of the three estimated variance components), the next largest contribution (over 40%) coming from the confounded residual variance, part of which will be candidate-question interaction variance. The low between-question variance contribution, at under 10%, which is good news for absolute measurement, is

in line with the details given above about the very small variation in question mean scores.

Figure 4.2
Test score distribution for GCSE Business Studies Higher Tier
(five 20-mark questions and 1,965 candidates)

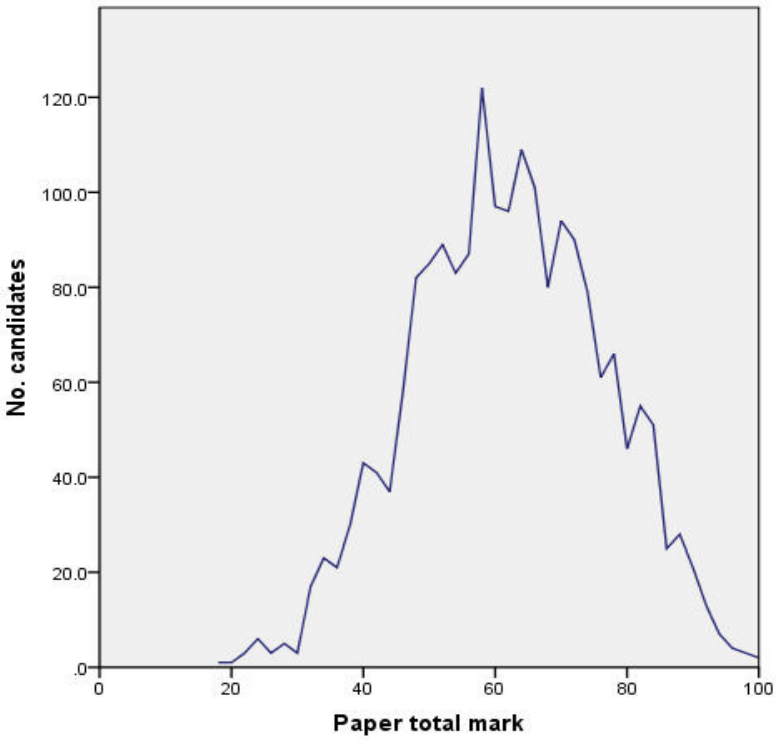


Table 4.1
G-study results for GCSE Business Studies Higher Tier
(five 20-mark questions and 1,965 candidates)

Variance component estimates and % contributions	
Candidates (7.0854)	51
Questions (1.0978)	8
Confounded residual (5.6540)	41
Generalizability coefficients	
Relative measurement (*Γ̃, equivalent to α for this design*)	0.86
Absolute measurement *(Φ)*	0.84
Standard errors of measurement	
SEM relative overall	5.3
SEM absolute overall	5.8
Margins of error	
ME relative	*10.4*
ME absolute	*11.4*

The relative reliability coefficient, *Γ*, which is equivalent to α for this simple crossed design, is a very acceptable 0.86, with the absolute coefficient, *Φ*, only slightly lower

at 0.84 (because the between-question variance contribution to measurement error is so low). The relative and absolute margins of error for candidate test scores are also close, at 10.4 for relative measurement and 11.4 for absolute measurement – this is just over 10% of the notional mark scale in each case, though higher in terms of the achieved 20-100 mark scale.

A *what if* analysis (Table 4.2) predicts the effect on reliability coefficients and on margins of error of increases in question numbers, on the assumption that the resulting samples of questions would be similar to the original in all respects save size and that the group of candidates taking the longer question paper would also resemble in all important respects those that sat the existing one.

Table 4.2
GCSE Business Studies Higher Tier: Estimated changes in coefficient values, SEMs and margins of error of changes in question numbers

Relative measurement						
No. questions	*No. markers*	*Mark scale**	*Γ*	*SEM*	*ME*	*ME as % of mark scale*
6	1	0-120	0.88	5.824	11.4	9.4
7	1	0-140	0.90	6.290	12.3	8.7
8	1	0-160	0.91	6.726	13.2	8.2
9	1	0-180	0.92	7.133	14.0	7.7
Absolute measurement						
No. questions	*No. markers*	*Mark scale**	*Φ*	*SEM*	*ME*	*ME as % of mark scale*
6	1	0-120	0.86	6.365	12.5	10.3
7	1	0-140	0.88	6.875	13.5	9.6
8	1	0-160	0.89	7.350	14.4	8.9
9	1	0-180	0.90	7.795	15.3	8.5

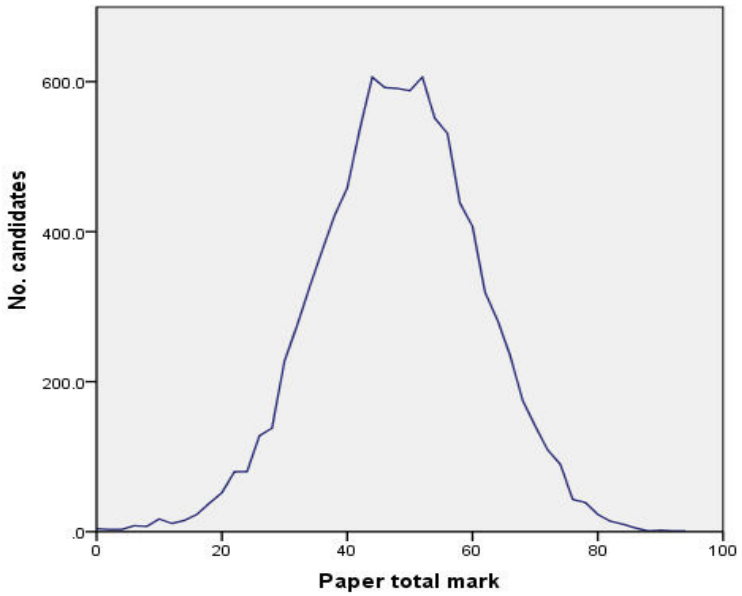
** This is the intended test score scale, not necessarily the achieved scale.*

As anticipated, increasing the numbers of questions would increase score reliability and reduce margins of error. However, the potential benefits would appear to be relatively small in this particular case, even for a question paper almost twice the length (in questions and time) of the original. Double marking in place of single marking might show greater promise, but this possibility cannot be explored with the given dataset, since scripts were not multiple-marked.

It might be interesting to compare the reliability results for this particular unit paper, from Board A, with those of an alternative, identically structured and identically timed, paper offered in the same year by a different examining board, Board B. The candidate entry for Board B's paper was larger in size than that for Board A's paper, at almost 10,000 candidates, but its composition was similar: just over 40% of the candidates were female, and almost 90% of all candidates had been entered from comprehensive secondary schools, 5% from secondary modern schools and 3% from independent schools. Despite the similarity in entry demographics, the paper itself showed some interesting differences in performance.

For example, unlike the mark distribution shown in Figure 4.2 for Board A's paper, the symmetric distribution for Board B's paper uses almost the full mark scale (Figure 4.3). Also, while there was relatively little variation in question mean scores for the first paper, question means in this alternative paper vary markedly, from 7.4 (37%) to 12.7 (64%). In a principal components analysis the first principal component accounted for just over 50% of the total candidate-score variance, slightly less than in the earlier case, where the proportion was over 60%.

Figure 4.3
Test score distribution for an alternative GCSE Business Studies Higher Tier paper
(five 20-mark questions and 9,627 candidates)



The G-study results for the two papers are shown in Table 4.3. Note that the residual variance accounted for similar, and quite high, proportions of the total variance in the two cases, at 41% and 45%, respectively. Note also, however, the marked difference in the relative contributions of between-question variance to total variance, at 8% and 27%, respectively. This difference is in line with the finding that the first paper showed much less variation in question mean scores than did the second paper. The between-candidate variance contributions also differ substantially, with a contribution of just over 50% for the first paper and under 30% for the second.

Because of the lower between-candidate variance contribution to total variance, and the higher between-question and residual variance contributions, both the relative and the absolute reliability coefficients are lower for Board B's paper than for the paper from Board A, at 0.76 and 0.66, respectively, compared with 0.86 and 0.84. SEMs and margins of error are also higher for the second paper compared with the first.

Clearly, neither an identical paper structure nor an almost identical candidate mix (in terms of gender and centre type) guarantees identical levels of reliability or score precision. Despite the similar make-up of the candidate entry there could be calibre differences between the entries that could in part explain the differences in mark

distributions. But it is evident that the particular questions that comprised the papers could be an alternative explanatory factor.

Table 4.3
G-study results for two alternative GCSE Business Studies Higher Tier papers
(five 20-mark questions; 1,965 and 9,627 candidates, respectively)

Variance component estimates and % contributions	*Board A*	*Board B*
Candidates (7.0854, 4.9054)	51	28
Questions (1.0978, 4.7273)	8	27
Confounded residual (5.6540, 7.9064)	41	45
Generalizability coefficients		
Relative measurement (*Γ, equivalent to α for this design*)	0.86	0.76
Absolute measurement *(Φ)*	0.84	0.66
Standard errors of measurement		
SEM relative overall	5.3	6.3
SEM absolute overall	5.8	7.9
Margins of error		
ME relative	*10.4*	*12.3*
ME absolute	*11.4*	*15.5*

Also, given that sometimes quite extended written responses were expected from candidates to many of the subquestions, marker effects could play a role in the final picture, despite pre-marking standardisation. Unfortunately, in the context of single live marking we cannot know whether marker effects are relevant or not, and if so to what extent.

4.3 GCSE Biology (equally weighted questions)

This 30-minute foundation tier paper comprised nine questions, each marked on a 0-4 scale. Five questions were 'matching' questions, while the other four each comprised four dichotomously-scored multiple-choice items. The total paper mark was 36. Although item marks were available we have for ease of analysis and interpretation chosen here to work at the level of question scores (we focus on item-level marks in Chapter 5, Section 5.4).

The dataset contained records for almost 25,000 candidates. In contrast with the Business Studies papers described in Section 4.2, Figure 4.4 shows a rather peaked and left-skewed total score distribution: the median mark was 23 (64% of the total mark) and two-thirds of all candidates achieved paper marks in the range 20-30 inclusive (56% to 83%). Question mean scores ranged from 1.48 (or 37% of the maximum of four marks) to 3.23 (81%) – see Figure 4.5. Question intercorrelations were uniformly low, and a single principal component accounted for just over 30% of the total variance. This suggests the presence of an important candidate-question interaction effect, along with other interaction effects involving questions, which would have a negative impact on score precision.

Figure 4.4
Test score distribution for GCSE Biology Foundation Tier
(nine 4-mark questions and 24,666 candidates)

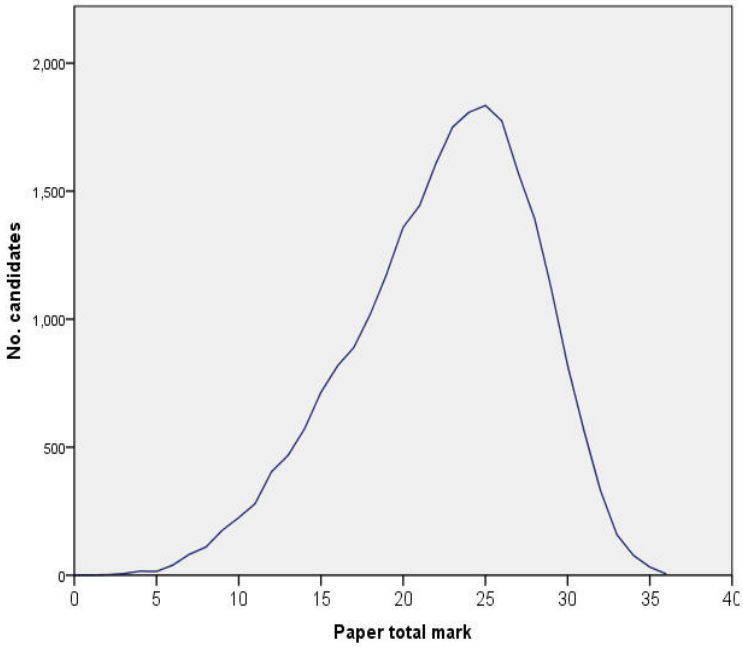
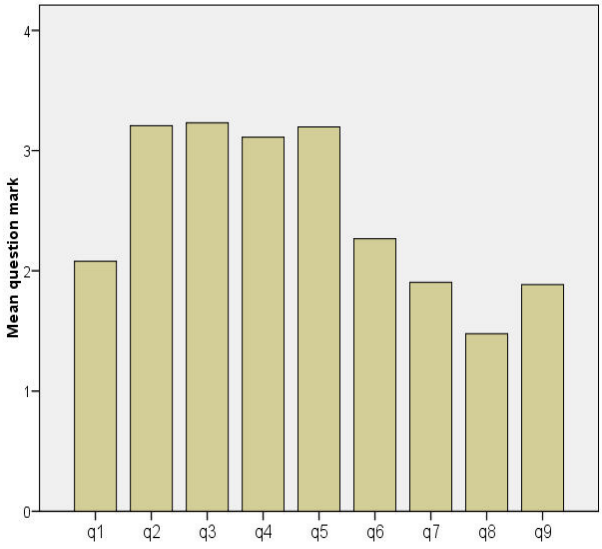


Figure 4.5
Variation in question mean scores
for GCSE Biology Foundation Tier
(nine 4-mark questions and 24,666 candidates)



The results of the generalizability analysis for this dataset are shown in Table 4.4, where the first feature to note is indeed the large variance component associated with the confounded residual – this accounts for over 55% of the total variance. This is unusually high, and it is this, combined with the low between-candidate variance, that explains the low reliability coefficients in this case: 0.71 for relative measurement and 0.62 for absolute measurement. It is the high between-question variance that further explains the relatively large drop in reliability between relative and absolute measurement. The bottom line is that the margins of error associated with candidates'

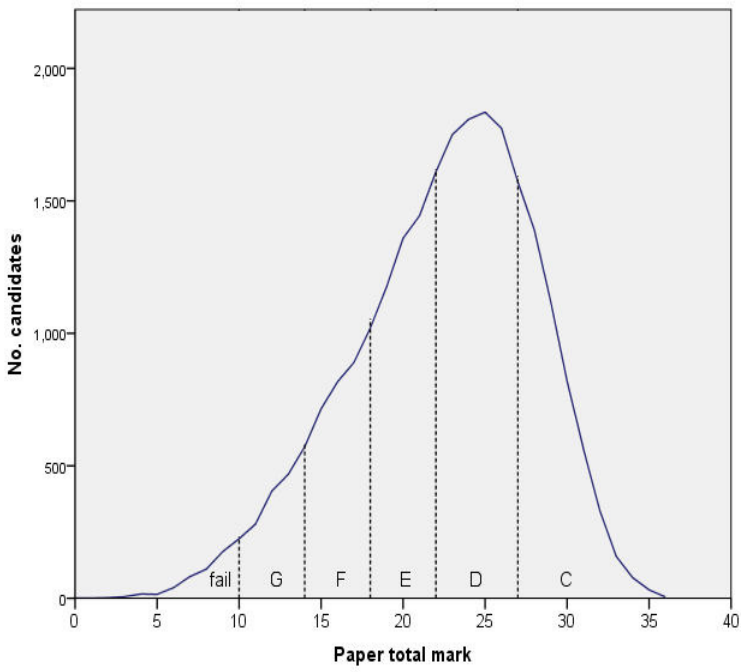
total marks are also relatively high, at just under six marks for relative measurement (over 15% of the 37-point total mark scale) and seven marks for absolute measurement (almost 20% of the total mark scale).

Table 4.4
G-study results for GCSE Biology Foundation Tier
(nine 4-mark questions and 24,666 candidates)

Variance component estimates and % contributions	
Candidates (0.2697)	15
Questions (0.4875)	28
Confounded residual (0.9867)	56
Generalizability coefficients	
Relative measurement (*Γ, equivalent to α for this design*)	0.71
Absolute measurement *(Φ)*	0.62
Standard errors of measurement	
SEM relative	3.0
SEM absolute	3.6
Margins of error	
ME relative	*5.9*
ME absolute	*7.1*

This is a paper for which boundary marks were available to us (Figure 4.6), allowing exploration of SEMs at these critical points.

Figure 4.6
Boundary marks for the GCSE Biology Foundation Tier paper
(nine 4-mark questions and 24,666 candidates)



The grading procedure for foundation and higher tier papers results in the division of the entire candidate entry for each tier into six subgroups: grades C, D, E, F, G and fail for foundation tier papers, and grades A*, A, B, C, D and E for higher tier papers (for details of the grading procedure for tiered papers see, for example, Wheadon & Béguin 2010). It is of interest to establish whether the generic SEMs shown in Table 4.4 (3 marks for relative measurement and 3.6 marks for absolute measurement) apply equally at each of the boundary marks.

To explore this possibility the dataset was sliced into five candidate subgroups, each subgroup containing all those candidates whose total paper marks were within a band of two marks either side of a boundary mark. Thus, for the G/fail boundary mark, which was 10, the marks obtained by the G/fail boundary subgroup varied from 8 to 12. Clearly, generalizability coefficients become irrelevant in this situation, having values at or close to zero, given that there will be little between-candidate mark variation. SEMs, in contrast, retain their relevance. The same generalizability analysis as before, i.e. the $c \times q$ design, was repeated for each candidate subgroup, and the two SEMs calculated.

It might be expected that SEMs will have lower values around the boundary marks at the extremes of the mark distribution, where the most able and the least able candidates in the paper entry are to be found, on the assumption that these candidates will show the most consistent performances from one question to another (mostly successes or mostly fails) thus reducing the importance of the candidate-question interaction variance. To some extent this is the case here. The relative SEMs for candidates at or close to the C/D, D/E, E/F, F/G, G/fail boundary marks are, respectively, 2.7, 3.1, 3.3, 3.3 and 3.0 marks. The corresponding absolute SEMs are, respectively, 3.4, 3.9, 4.0, 3.8 and 3.3 marks. Grade misclassification for individual candidates is therefore more likely to occur around the middle grade boundaries. [For details of the calculation of conditional SEMs for individual candidates see Brennan 2001a, pp.160-164].

Looking again at the full candidate group, optimisation studies (Table 4.5) show that it would require an increase in paper length to 14 questions for the reliability coefficient for relative measurement to reach the more comfortable level of 0.80, the minimum level generally considered to be acceptable in a high-stakes assessment context. To achieve this for absolute measurement the number of questions would need to be increased to 20. And even then measurement errors would be higher than might be hoped.

It is interesting to speculate about the reasons for the rather poor performance of this particular paper in reliability terms, which was shared by the partner 9-question higher tier Biology paper and corresponding foundation and higher tier papers in Physics, which all had similar structures and showed closely parallel reliability characteristics. Hidden marker effects are not responsible, since the subquestions in this paper were all of objective format. Gender effects might be relevant, as might centre type effects – unfortunately, relevant candidate demographic data were not included in the supplied dataset, with the consequence that these possibilities could not be explored.

Table 4.5
GCSE Biology Foundation Tier: Estimated changes in paper properties of changes in numbers of questions

Relative measurement					
No. questions	*Mark scale**	*Γ*	*SEM*	*ME*	*ME as % of mark scale*
14	0-56	0.79	3.7	7.3	12.8
16	0-64	0.81	4.0	7.8	12.0
18	0-72	0.83	4.2	8.2	11.2
20	0-80	0.85	4.4	8.6	10.6
Absolute measurement					
No. questions	*Mark scale**	*Φ*	*SEM*	*ME*	*ME as % of mark scale*
14	0-56	0.72	4.5	8.8	15.4
16	0-64	0.75	4.9	9.6	14.8
18	0-72	0.77	5.2	10.2	14.0
20	0-80	0.79	5.4	10.6	13.1

** That is the notional test score scale*

4.4 GCE General Studies (objective test section)

Our next example is a 90-minute 2-section AS paper from a unitised A/AS examination in General Studies. Section A comprised 30 multiple-choice items while Section B comprised three constructed response questions. We will here consider Section A only, leaving consideration of the reliability of the unit as a whole to Chapter 5 (Section 5.6).

Complete records were available for more than 22,000 candidates, with an even gender split. Figure 4.7 reveals a slightly left-skewed total score distribution for the section, with the majority of candidates in the mark range 10 to 30 and centring on around 20 marks (the mean mark was 18.7). There was a very slight, though statistically significant, gender difference in section performance, females achieving a slightly higher mark, at 18.8, than males, at 18.6. Question mean scores (item facilities in this case) ranged widely, from 0.37 to 0.84 (Figure 4.8).

Question intercorrelations were uniformly low, and a principal components analysis revealed a multidimensional structure on this occasion. The first principal component accounted for just under 12% of the total variance, the first seven together accounting for a modest 33%. Just as in the case of GCSE Biology, this suggests the presence of interactions between candidates and other factors, including questions, a possibility confirmed by the G-study analysis (Table 4.6).

Figure 4.7
Score distribution for AS/A level General Studies: Section A of a 2-section unit paper
(30 objective questions and 22,424 candidates)

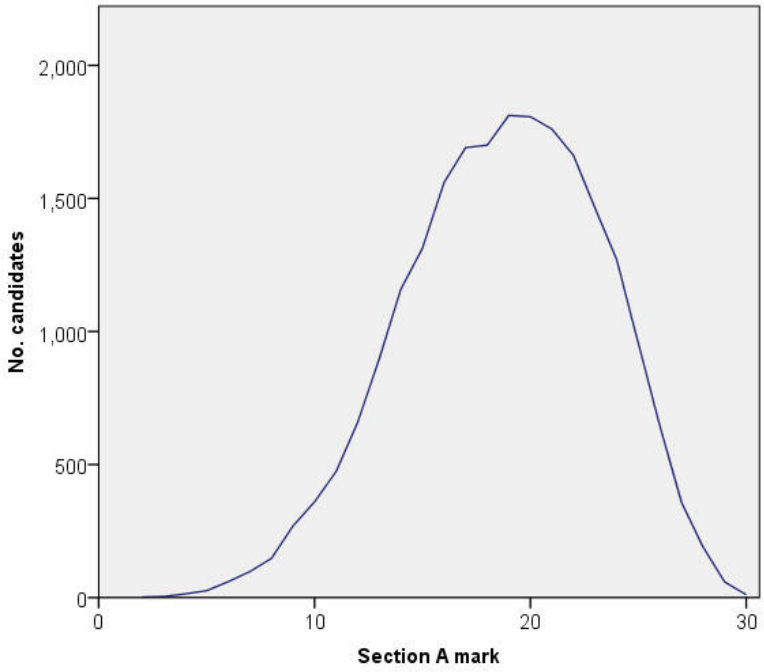


Figure 4.8
Variation in question mean scores for AS/A level General Studies unit Section A
(30 objective questions and 22,424 candidates)

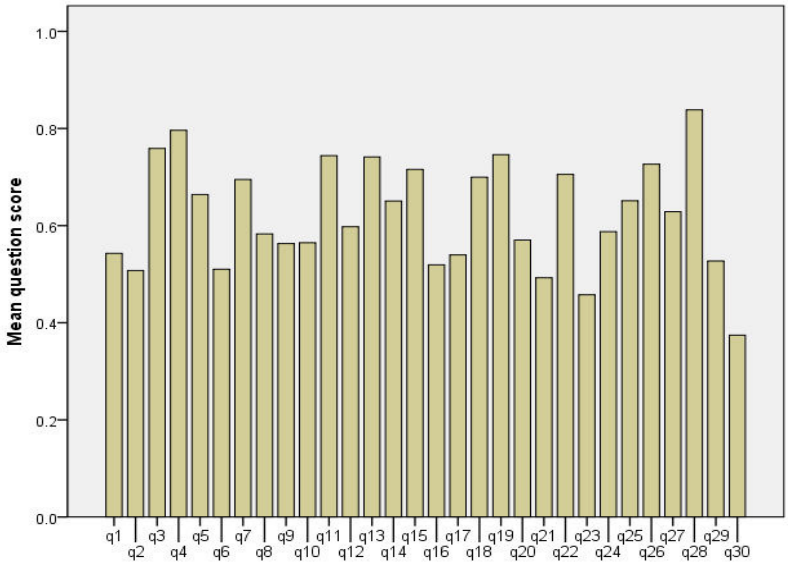


Table 4.6
G-study results for AS/A level General Studies unit paper section
(30 objective questions and 22,424 candidates)

Variance component estimates and % contributions	
Candidates (0.0161)	7
Questions (0.0123)	5
Confounded residual (0.2068)	88
Generalizability coefficients	
Relative measurement (*Γ, equivalent to α for this design*)	0.71
Absolute measurement *(Φ)*	0.69
Standard errors of measurement	
SEM relative	2.5
SEM absolute	2.6
Margins of error	
ME relative	*4.9*
ME absolute	*5.1*

Table 4.6 shows clearly that, despite the evidently high between-candidate variance shown in Figure 4.7 and the high between-question variance illustrated in Figure 4.8, of the three variance components that can be quantified in this simple design the confounded residual is by far the largest, accounting for fully 88% of the total variance. This compares with a 7% contribution from between-candidate variance and a 5% contribution from between-question variance. These results explain both the close similarity between the relative and the absolute generalizability coefficients, as well as their modest values (0.71 and 0.69, respectively). The margins of error for the two types of measurement are also virtually identical, at 4.9 marks for relative measurement and a slightly higher 5.1 marks for absolute measurement, both around 16% the length of the notional underlying section score scale.

In Table 4.7 we see that a 50-question section, i.e. a 50-item objective test, would bring the generalizability coefficients up to 0.80 or thereabouts. The margin of error would even then, however, still be equivalent to more than 12% of the section score scale.

For the results of an analysis of the entire 2-section paper, and comparison with an alternative General Studies paper offered by another examining board that same year, see Section 5.6 in Chapter 5.

Table 4.7
AS/A level General Studies Unit paper section: Estimated changes in reliability coefficients, SEMs and margins of error of changes in numbers of questions

Relative measurement					
No. questions	*Section mark scale*	*Γ*	*SEM*	*ME*	*ME as % of mark scale*
35	0-35	0.73	2.69	5.3	14.7
40	0-40	0.76	2.88	5.6	13.7
45	0-45	0.78	3.05	6.0	13.0
50	0-50	0.80	3.22	6.3	12.4
Absolute measurement					
No. questions	*Section mark scale*	*Φ*	*SEM*	*ME*	*ME as % of mark scale*
35	0-35	0.72	2.77	5.4	15.0
40	0-40	0.75	2.96	5.8	14.1
45	0-45	0.77	3.14	6.2	13.5
50	0-50	0.79	3.31	6.5	12.7

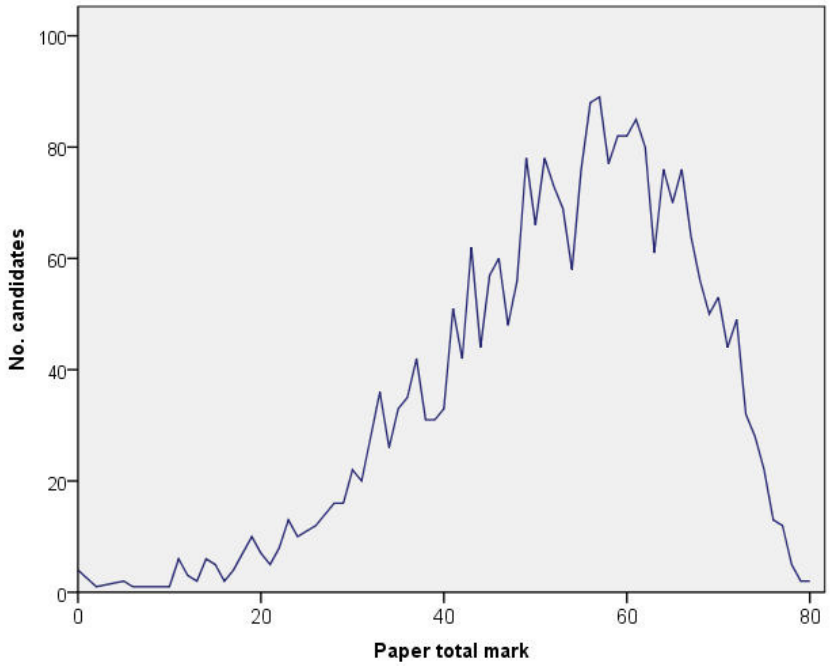
4.5 GCSE Drama (equally weighted questions, with choice)

This single 2-hour written paper actually contained a series of separate parallel component papers, all identical in length and style, but whose specific content differed to accommodate the assessment of candidates who had studied different combinations of set plays: one common play plus one other from a list of options. Each subsumed paper comprised four sections, two relating to the obligatory play and two to the optional choice. The first section contained a single question comprising four constructed response subquestions, with varying mark allocations. The second section contained two extended response questions, among which candidates were to choose one. Each section carried a 20-mark total. Thus each candidate would attempt four questions in total, two on each of their set plays, obligatory plus optional choice, for a maximum possible 80 marks (20 marks per question per play).

The resulting dataset contained valid records for just over 2,700 candidates. Over 60% of the candidates were female, and almost 85% were entered from comprehensive secondary or middle schools with almost all the rest entered from independent schools. The dataset revealed three candidate subgroups. One subgroup had studied plays 1 and 2 (hereafter referred to as option 2), a second had studied plays 1 and 3 (option 3) while the third had studied plays 1 and 4 (option 4).

The score distribution for the paper is shown in Figure 4.9. The total score distributions for the three subgroups on the mandatory play were very similar, but there were statistically significant differences in subgroup mean scores for that play: these were 26.8, 27.3 and 26.1, respectively, for candidate option groups 2, 3 and 4.

Figure 4.9
Score distribution for the GCSE Drama paper
(four 20-mark questions, two per play, and 2,720 candidates)



For each candidate subgroup question inter-correlations were high (between 0.60 and 0.75), and a single principal component that accounted for between 72% and 75% of the total variance. In other words, all three embedded papers were well-structured and essentially unidimensional. This would suggest high technical reliability and, in particular, high score precision. Table 4.8 confirms this to be the case.

Table 4.8
G-study results for GCSE Drama alternative papers (two plays)
(four 20-mark questions, and 994, 838 and 888 candidates per option)*

Variance component estimates and % contributions	*Option 2*	*Option 3*	*Option 4*
Candidates (10.5893, 9.9114, 10.8371)	56	54	57
Questions (2.5417, 2.6513, 2.7904)	13	14	15
Confounded residual (5.7500, 5.7518, 5.4750)	30	31	29
Generalizability coefficients			
Relative measurement (Γ, *equivalent to* α *for this design*)	0.88	0.87	0.89
Absolute measurement *(*Φ*)*	0.84	0.83	0.84
Standard errors of measurement			
SEM relative	4.8	4.8	4.7
SEM absolute	5.8	5.8	5.7
Margins of error			
ME relative	*9.4*	*9.4*	*9.2*
ME absolute	*11.4*	*11.4*	*11.2*

** Each option consisted of play 1 plus one choice from plays 2, 3 and 4.*

We see from Table 4.8 that for all three candidate subsets the contribution of between-candidate variation to total variance was high, at over 50% in each case, and that the confounded residual contributed another 30%. The generalizability coefficients, for both relative and absolute measurement, were over 0.80, approaching 0.9 for relative measurement. The SEMs for relative measurement are all under five marks, and the margins of error just over nine marks on the 80-mark scale. For absolute measurement the SEMs rise to almost six marks and the margins of error to over 11 marks.

4.6 GCE History (equally weighted questions, with choice)

Our next example is a 90-minute AS unit paper in History. The paper was one of several alternative Unit 1 papers, each focusing on a different period of history, but with identical structures: three 60-mark extended response questions, each question comprising two subquestions with mark allocations of 24 and 36, respectively. Candidates were to choose two questions for a paper total mark of 120. Centres entered candidates for whichever alternative paper matched their curriculum choice, while candidates (presumably) would themselves have had the responsibility for choosing the questions to respond to. We will call the paper to be considered here Version A.

Around 750 candidates produced answers to two of the three questions in paper Version A. Just under 60% of the candidates were female and almost 90% of all candidates had been entered for Version A by their comprehensive secondary schools. The numbers of candidates who chose different combinations of two from the three questions varied widely, from 436 for questions 1 and 2 (option q1q2) to 249 for questions 1 and 3 (option q1q3) and just 66 for questions 2 and 3 (option q2q3). The candidates who chose option 1 or option 2 came from a number of different centres, with anything from one candidate to 40 or so in each. Candidates who chose option 3, on the other hand, came from just two centres.

This particular examination was newly introduced in 2009, and its performance shows some intriguing features that will have been of great interest to the examiners who set the various unit papers. Look, for example, at the mark distribution for Version A as a whole (Figure 4.10). The mean mark for the paper was 40.8, or 34% of the full intended mark scale, and 93% of all candidates achieved a percentage mark lower than 50%.

The three questions had very similar mean marks, at between 19 and 21.5 (32%-36%) across all relevant candidates. Interestingly, however, the inter-question correlation was very low for option groups 1 (q1q2) and 2 (q1q3), at 0.4 or lower, but quite high, at 0.7, for the smaller group of candidates who went for option 3 (q2q3). From this fact alone one might predict that the reliability of the paper would be different from one option to another. Table 4.9 confirms this to be the case.

While between-candidate variance for the option 3 group contributed 69% of the total variance, for option groups 1 and 2 the corresponding contributions were just over 40% and 33%, respectively. And while the residual variance, containing the candidate-question interaction variance, contributed around 26% to total variance for the option 3 group, the contribution rose to 57% and 67%, respectively, for groups 1 and 2.

Figure 4.10
Score distribution for GCE History Unit 1 Version A
(two 60-mark questions per candidate and 751 candidates)

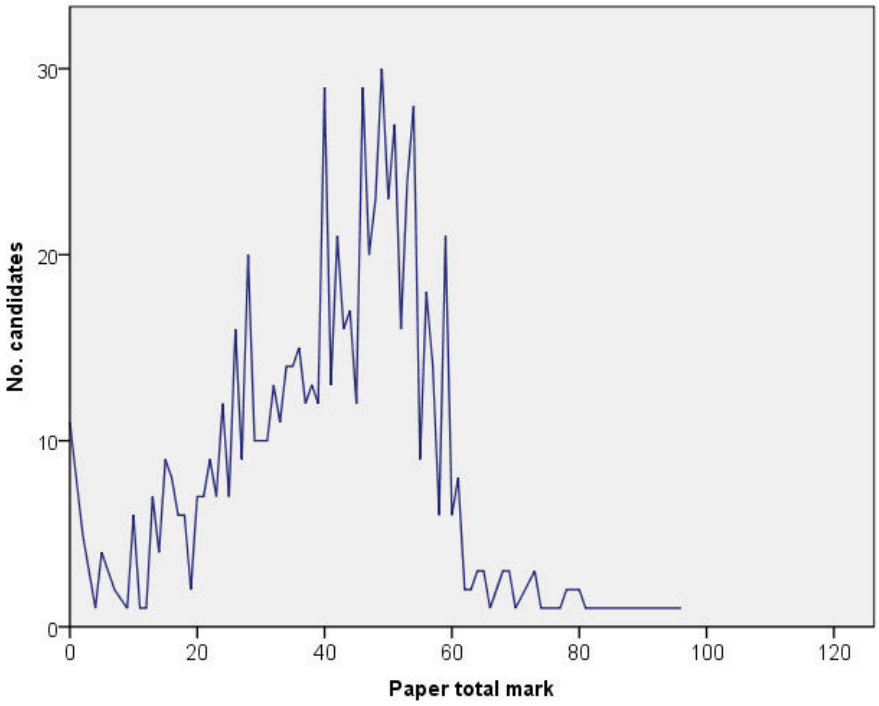


Table 4.9
G-study results for GCE History AS Unit 1 Version A
(two 60-mark questions; 436, 249 and 66 candidates per option)*

Variance component estimates and % contributions	*Option 1 (q1q2)*	*Option 2 (q1q3)*	*Option 3 (q2q3)*
Candidates (31.7822, 26.6967, 94.4592)	41	33	69
Questions (1.3305, 0.1528, 7.6424)	2	<1	5
Confounded residual (44.5021, 53.9195, 35.6682)	57	67	26
Generalizability coefficients			
Relative measurement (*Γ, equivalent to α for this design*)	0.59	0.50	0.84
Absolute measurement *(Φ)*	0.58	0.50	0.81
Standard errors of measurement			
SEM relative	9.4	10.4	8.5
SEM absolute	9.6	10.4	9.3
Margins of error			
ME relative	*18.4*	*20.4*	*16.7*
ME absolute	*18.8*	*20.4*	*18.2*

** Candidates attempted two questions from a choice of three.*

The between-question variation was minimal in all groups, making a tiny contribution to total variance. For this reason, as Table 4.9 shows, the relative and the absolute coefficients are virtually identical. For groups 1 and 2, however, we see reliability

coefficients of just under 0.6, and 0.5, respectively, while for group 3 the picture is very much more positive, with coefficients over 0.8 for both types of measurement.

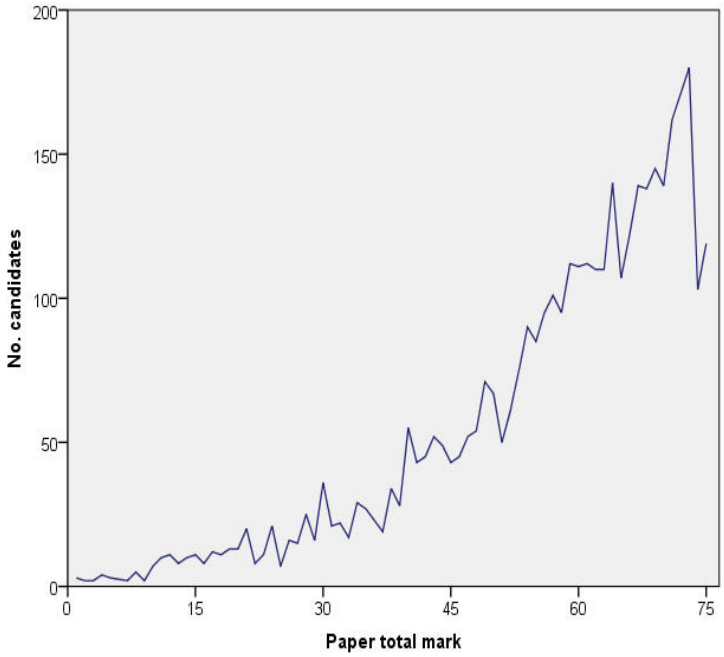
Finally, because each candidate script was marked by just one marker, it should be noted that, as in other cases, the analyses carried out on this unit paper have not been able to explore the further influence of markers on score reliability, overall or for each possible question combination.

4.7 GCE Statistics (unequally weighted questions)

This was a 90-minute paper comprising seven structured questions, with no question choice. One important difference between this paper and those described earlier in this chapter is that it was not only the subquestions that had different metrics, the questions did too – to be more specific, the total marks assigned to the questions were unequal, and varied from 8 to 17. The total mark for the paper was 75. Subquestion marks were not available electronically, so that the analyses described here are based on question totals. Full data records were available for just under 4,000 candidates, almost 60% of them male, from a total of over 500 centres. Among centres, 80% entered 10 or fewer candidates, with 25% of all centres entering a single candidate.

Figure 4.11 illustrates a rather interesting asymmetric score distribution for this paper, with a mean score of 56.8 (or almost 76%) on the 0-75 mark scale. There was no gender difference in test performance, and neither were there any for individual questions, with just one exception in which females candidates performed significantly better than male candidates (but not so much better as to produce a better overall test performance). Question intercorrelations were modest to relatively high, varying between 0.24 and 0.68. A single principal component accounted for 54% of the total variance.

Figure 4.11
Test score distribution for A/AS Statistics unit paper
(7 variable-mark questions and 3,980 candidates)



As far as the between-question variance is concerned we have an equally interesting picture. Six of the seven questions carried total marks between 8 and 11, and showed quite similar mean scores, at between 6.7 and 7.8. The remaining question carried 17 marks, and had a mean score of 12.8. The large discrepancy in maximum score and mean score between this one question and the other six is clear in Figure 4.12. The result of the consequent high between-question variance could be predicted to impact on score precision for absolute measurement. For interest, Figure 4.12 compares the raw mean scores for the seven questions with the mean scores after adjustment onto a common 0-8 scale. An important consequence of the adjustment is clearly seen in Figure 4.12 to be a reduction in the variation in question mean scores, i.e. a reduction in the between-question variance. This will impact positively on the absolute reliability.

Figure 4.12
Variation in question mean scores
for the GCE A/AS Statistics unit paper
(seven variable-mark questions and 3,980 candidates)

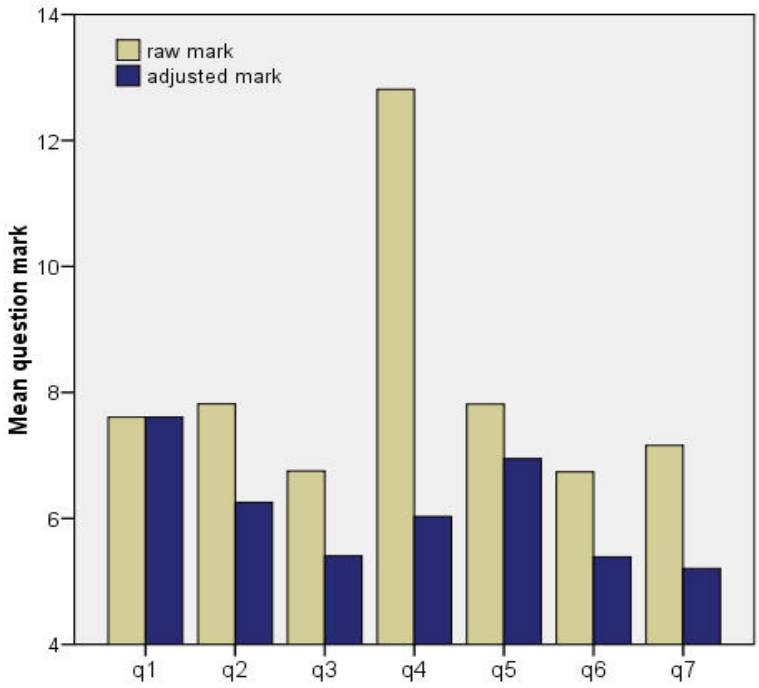


Table 4.10 presents the G-study results for analyses based separately on raw scores and on adjusted scores. The fact that the questions in the paper carried different, and in one case widely different, mark allocations raises interesting issues for G-theory, and for other alternative analysis techniques, since the theory was developed for application in cases where questions (items, tasks) carry the same weight within the whole paper or within a paper section. The complication of varying metrics, with just one single question associated with any one metric, poses a particular challenge when it comes to using a specific paper for *what if* analyses, where these explore the impact on reliability of smaller or larger samples of similar questions.

Looking at Table 4.10, we see that for the raw score analysis the variance component estimates for candidates, questions and the confounded residual are of relatively similar size, each contributing between 30% and 40% of the total score variance. The

relative measurement coefficient is adequately high at 0.84, but the absolute coefficient is lower at 0.74. Margins of error are 12.0 and 16.3 in each case, being equivalent to almost 16% and over 21%, respectively, of the paper's total score scale (0-75).

Table 4.10
G-study results for the A/AS Statistics unit paper
(seven variable-mark questions for the 'raw marks' analysis and seven 8-mark questions for the 'adjusted marks' analysis; 3,980 candidates)

Variance component estimates and % contributions	*Raw marks*	*Adjusted marks*
Candidates (4.0104, 2.0954)	29	38
Questions (4.5275, 0.8003)	33	15
Confounded residual (5.2495, 2.5520)	38	47
Generalizability coefficients		
Relative measurement (*Γ, equivalent to α for this design*)	0.84	0.85
Absolute measurement *(Φ)*	0.74	0.81
Standard errors of measurement		
SEM relative	6.1	4.2
SEM absolute	8.3	4.8
Margins of error		
ME relative	*12.0*	*8.2*
ME absolute	*16.3*	*9.4*

If, however, we turn attention to the results for the adjusted question marks, in which all questions are essentially given equal weight in the total paper mark, then the picture improves. In particular, the between-question variance reduces considerably, which explains the improvement in the absolute coefficient (from 0.74 to 0.81) when compared with the barely changed relative coefficient. Margins of error are reduced for both types of measurement, down to 8.2 for relative measurement and 9.4 for absolute measurement, equivalent to under 15% and under 17% of the paper total score scale (which has become 0-56 in place of the original 0-75).

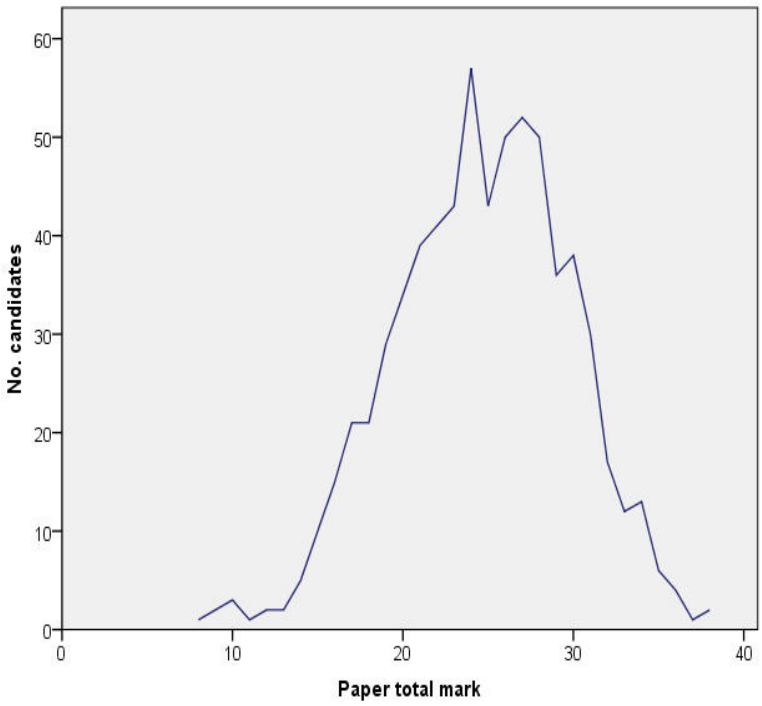
Thus, for this paper at least, equalising question contributions to the paper total score would improve 'absolute' reliability. The question is would equalising question contributions result in a less valid paper in the eyes of the examiner(s) who originally set the paper?

4.8 GCSE Music (aural test, unequally weighted questions)

This 45-minute written paper comprised seven multi-part questions based on several different recorded excerpts of pieces of music. The number of parts in each question varied from one to six, some being multiple choice and others short answer. Most parts carried one mark, and question totals varied between four and eight marks. The total mark for the paper as a whole was 40. Data records were available for just under 700 candidates. Of these, just under 55% were female. All but three candidates were entered for the examination from comprehensive secondary schools.

The mark distribution for the paper is shown in Figure 4.13. As we see, the distribution is fairly symmetric, but note that the lowest quarter of the intended score scale is virtually unused. The overall mean score was 24.7 (or just under 62%), and there was a statistically significant gender difference in mean scores (25.2 males, 24.3 females). Question mean scores varied from 2.2 to 5.4, or from 55% to 68% of their total mark allocations. Figure 4.14 illustrates the variation in raw mean marks and the variation in mean marks after adjustment where necessary onto a common 0-4 mark scale.

Figure 4.13
Test score distribution for the GCSE music paper
(7 variable-mark questions and 678 candidates)



Question intercorrelations were modest to low, at between 0.33 and 0.15, and a single principal component accounted for just 34% of total variance. This suggests multidimensionality in the response data, which would be borne out by a review of the wide range of different types of music knowledge being tapped in the paper.

Once again, generalizability analyses were carried out on raw marks and on adjusted marks. The results are given in Table 4.11. For the raw mark analysis the table shows that the between-candidate component is lower in value than both the between-question variance and the confounded residual variance. This does not bode well for reliability as far as generalizability coefficients are concerned. And indeed we do see rather low coefficient values for these, at 0.67 for relative measurement and 0.51 for absolute measurement. The margins of error are 5.7 and 8.0 respectively, equivalent to just over 14% and 20% of the paper's total mark scale (0-40).

Figure 4.14
Variation in question mean scores for the GCSE Music paper
(seven variable-mark questions and 678 candidates)

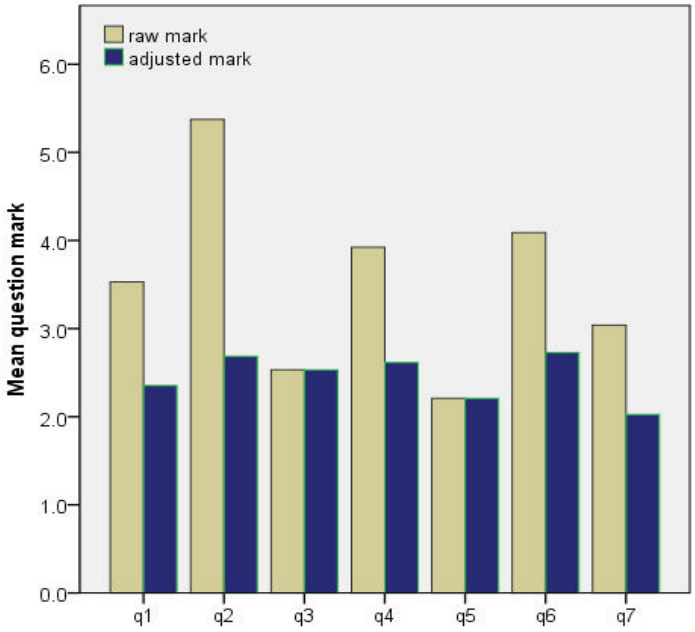


Table 4.11
G-study results for the GCSE Music paper
(seven variable-mark questions for the 'raw marks' analysis and seven 4-mark questions for the 'adjusted marks' analysis; 678 candidates)

Variance component estimates and % contributions	*Raw marks*	*Adjusted marks*
Candidates (0.3548, 0.1777)	13	21
Questions (1.1392, 0.0681)	42	8
Confounded residual (1.2234, 0.6120)	45	71
Generalizability coefficients		
Relative measurement (*Γ , equivalent to α for this design*)	0.67	0.67
Absolute measurement *(Φ)*	0.51	0.65
Standard errors of measurement		
SEM relative	2.9	2.1
SEM absolute	4.1	2.2
Margins of error		
ME relative	*5.7*	*4.1*
ME absolute	*8.0*	*4.3*

Turning attention now to the corresponding results for the analysis of adjusted question scores, we see a dramatic reduction in the size of the between-question variance component relative to the other two components. An important and predictable consequence of this is a marked improvement in the reliability for

absolute measurement, though this is still low, with a coefficient value of 0.65 compared with the previous 0.51; since the between-question variance component makes no contribution to the relative measurement coefficient this is unchanged at 0.67. For each type of measurement the margin of error is now roughly four marks on the new 0-28 mark scale, i.e. just under 14% of scale length.

With varying question metrics the paper would need to be doubled in length, to 14 questions, to bring the relative reliability coefficient up to 0.80, and quadrupled in length to do the same for the absolute coefficient. That is assuming that it makes sense to talk about sampling questions from a virtual pool of questions with varying total score metrics.

With questions carrying equal weight in the paper total mark, doubling the length of the paper would bring both reliability coefficients up to 0.80, and margins of error, at around six marks, down to roughly 10% of the new 0-56 score scale (for questions marked on a common 0-4 scale). It would be for examiners to decide whether giving equal weight to every question would jeopardise the validity of such a paper. If it would not threaten validity then it would be for the awarding body and its centres to decide whether the gain in reliability would be worth the additional testing time and, presumably, cost. But this assumes that the same properties would be shown by the unit papers used before and after this one, and without analysis results we cannot know whether that assumption holds or not.

4.9 GCSE French (aural test, unequally weighted questions)

Our final example in this chapter is a 30-minute GCSE French foundation tier listening test, which formed one section in a unit whose other sections focused on reading and writing, respectively. The listening test comprised six questions, with maximum question marks ranging from four to seven for a section total of 30 marks. Each question contained a set of binary scored objective format subquestions. Candidates were advised to read through the questions before listening to a recording. After listening to the recording they were given a short time to read through the questions again, answering if they chose to at that point. They then heard the recording for a second time before finalising their responses to the questions.

Data records were available for just over 7,000 candidates, over half of them female. The candidates were drawn from almost 400 centres, with half the centres entering 10 or fewer candidates each, the maximum number entered from any one centre exceeding 150. The symmetric mark distribution, whose average mark is 21.7 (or 72%), is shown in Figure 4.15: interestingly, note the virtual absence of candidates in the bottom third of the mark scale, with the consequence that the intended 31-point scale has in practice become much shorter. There was a small but statistically significant difference in the mean scores of male and female students, with females averaging 21.8 marks and males 21.5; female candidates performed slightly, but statistically significantly, better than male candidates on two of the seven questions.

Question mean scores varied between 2.6 and 4.7, or, as percentage marks, between 52% and 93%: the profile of question mean scores when all questions are adjusted onto a common 0-4 scale is shown in Figure 4.16 alongside the profile for raw mean scores.

Figure 4.15
Test score distribution for the GCSE French listening test
(6 variable-mark questions and 7,109 candidates)

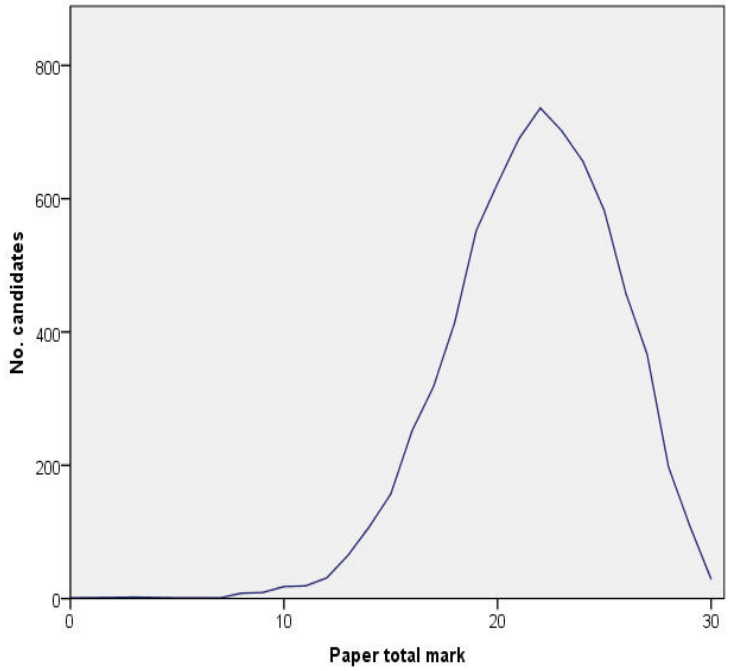


Figure 4.16
Variation in question mean scores
for the GCSE French listening test
(six variable-mark questions and 7,109 candidates)

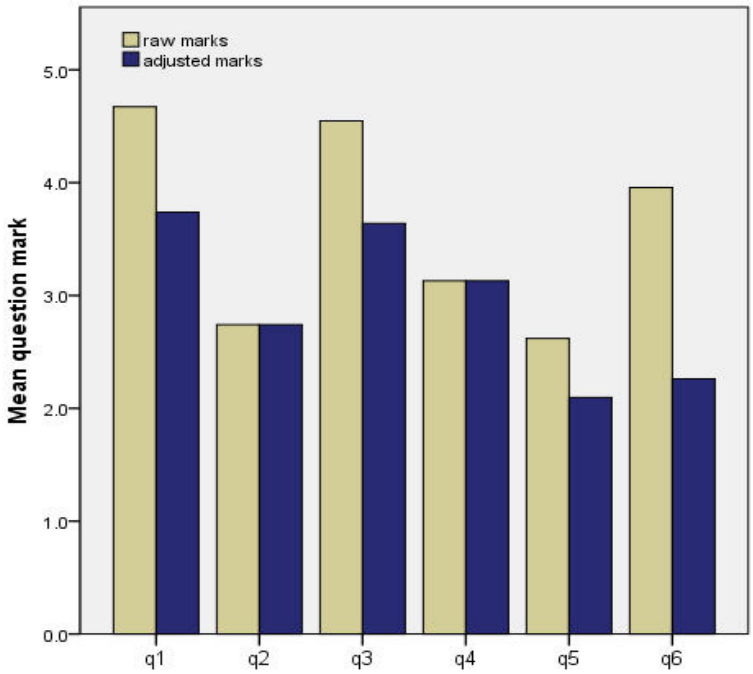


Table 4.12 presents the G-study results for the test, based first on raw question marks and then on adjusted question marks. Interestingly, we see that in this case equalising the contributions of the questions to the total section mark, through adjustment onto a

common scale, has had little effect on the section's performance. Between-question variance has remained high, contributing around 40% to total variance, with the confounded residual contributing just under another 50%. For these reasons the reliability coefficients are low, at around 0.6 for relative measurement and 0.45 for absolute measurement.

Table 4.12
G-study results for the GCSE French listening test
(six variable-mark questions for the 'raw marks' analysis and six 4-mark questions for the 'adjusted marks' analysis; 7,109 candidates)

Variance component estimates and % contributions	*Raw marks*	*Adjusted marks*
Candidates (0.2386, 0.1558)	12	13
Questions (0.8182, 0.4744)	40	38
Confounded residual (0.9920, 0.6047)	48	49
Generalizability coefficients		
Relative measurement (*Γ, equivalent to α for this design*)	0.59	0.61
Absolute measurement *(Φ)*	0.44	0.46
Standard errors of measurement		
SEM relative	2.4	1.9
SEM absolute	3.3	2.5
Margins of error		
ME relative	*4.7*	*3.7*
ME absolute	*6.5*	*4.9*

The margins of error as percentages of the mark scale remain at around 17% for both the raw 0-30 mark scale and for the adjusted 0-24 mark scale in the case of relative measurement, and at around 25% of each scale for absolute measurement.

Increasing the test length to 10 questions could be predicted to increase the relative reliability coefficient to 0.70, but it would require an increase to at least 16 questions to approach this value for absolute measurement. Margins of error for relative and for absolute measurement are 4.7 and 6.5, respectively, or just over 15% and 20% of the notional 31-point section mark scale (22% and 31% of the achieved 21-point mark scale). Equalising the contributions of the questions to the section total mark would in this case barely change the picture: the reliability coefficients are essentially unchanged, as are the margins of error as proportions of the new (0-20) total mark scale for six questions carrying four marks each.

Subquestion scores were available for this particular paper. These clearly contribute more measurement information than question total scores can. However, since subquestions are by definition nested within questions, any analysis at the level of subquestion scores should take this hierarchical structure into account (see Chapter 5 for examples).

5 Nested designs and composite scores

5.1 Introduction

The previous chapter has offered several illustrations of the basic c × q design, for papers of increasingly complex structures. In this chapter we move further to explore the impact of other variables on assessment reliability and in particular on score precision. The factors we consider separately or in combination include gender, centre, subquestions and paper sections. Once again, however, it should be noted that one factor, and one very important factor, that is not explored here is markers, since relevant data were not available for analysis.

Before reviewing the results of the analyses readers are again reminded that each of the papers considered in this chapter was typically just one component or one unit in a unitised examination offered in 2009 for GCSE or GCE qualifications. The reliability of that one component will have implications for the reliability of the examination as a whole, but identifying the implications for whole-examination reliability is beyond the scope of this report.

As in the previous chapter, in all tables SEM and ME refer, respectively, to the standard error of measurement associated with a candidate's total paper mark and the margin of error with which a 95% confidence interval around that total mark might be constructed.

5.2 GCSE Chemistry (gender and centre)

This foundation tier paper comprised eight multi-part questions, with up to five marked parts per question. Question marks varied from 4 to 8, and subquestion marks varied from 1 to 3; question marks, but not subquestion marks, were electronically recorded: The paper total mark was 50.

The paper was attempted by more than 10,000 candidates, just under 50% of whom were male and almost all of whom were entered by comprehensive secondary schools. Figure 5.1 presents the mark distribution for the paper, whose mean mark was 28.1. The profiles of question mean scores, raw and adjusted, are illustrated in Figure 5.2. The first principal component accounted for 45% of the total candidate-question score variance.

There was no gender difference in performance on the paper as a whole, but there were statistically significant differences for some of the questions in one direction or the other. It would therefore be interesting to evaluate the impact of the gender-question interaction on measurement error. To do this we adopt the analysis design (c:g) × q, indicating that we have candidates nested within gender and crossed with questions (all candidates attempt all eight questions). While the candidates and questions in the dataset are still considered to be random samples representing larger populations, i.e. they are technically termed 'random factors', gender is clearly not sampled. Since both genders are included in the dataset gender is technically a 'fixed factor': we are not implicitly or explicitly attempting to generalise results beyond the two genders that feature in the analysis. The appropriate variance partition diagram, with shading indicating valid and error variance contributions, is shown as Figure 5.3.

Figure 5.1
Score distribution for GCSE Chemistry foundation tier
(8 variable-metric questions and 10,134 candidates)

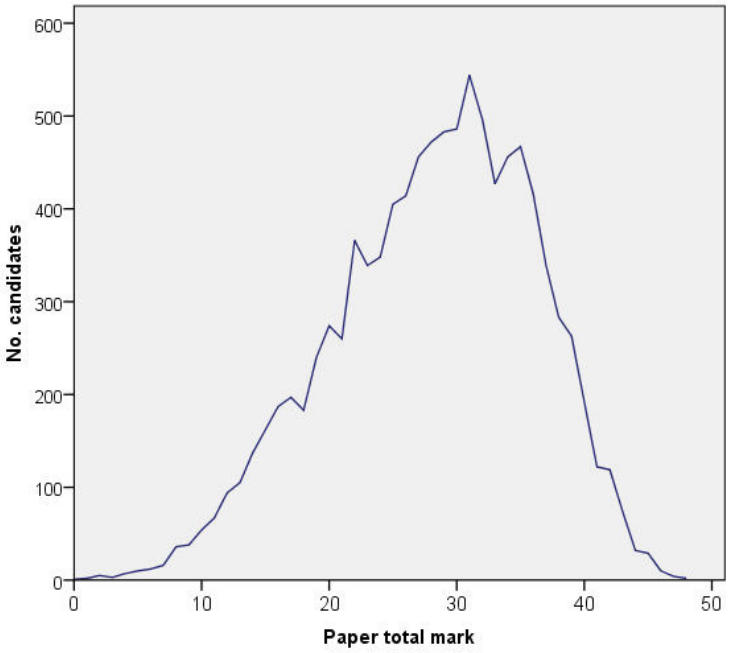


Figure 5.2
Variation in question mean scores
for GCSE Chemistry foundation tier
(8 variable-metric questions and 10,134 candidates)

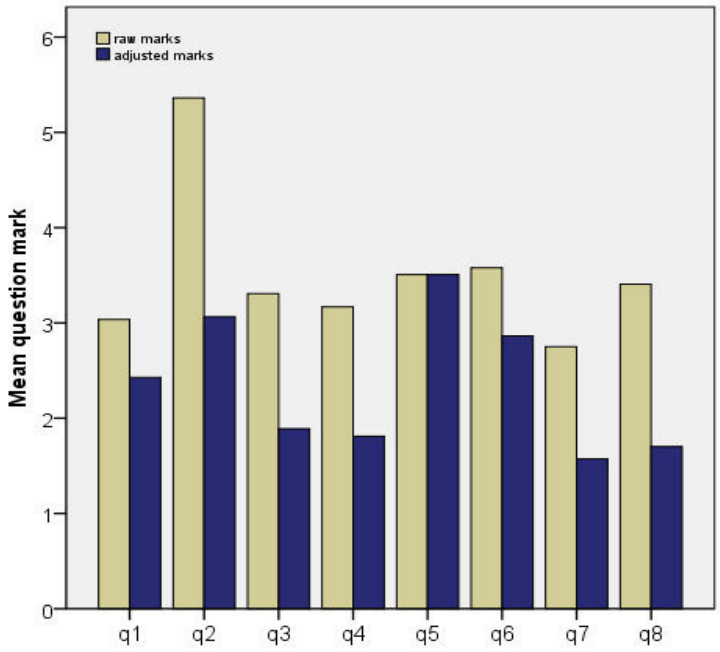
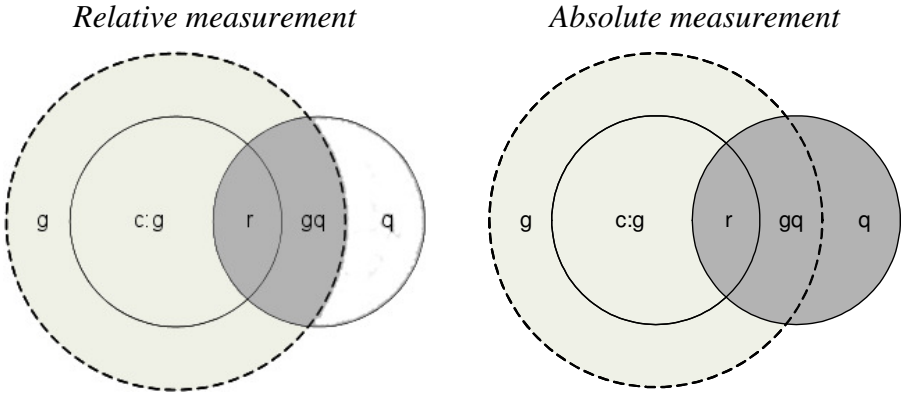


Figure 5.3
Contributions to relative and to absolute measurement error in the (c:g) × q design*



** In the (c:g) × q design, c represents candidates, g gender (a fixed factor) and q questions; r represents the confounded residual variance.*

The G-study results are shown in Table 5.1. Note first the very small negative gender variance component, reflecting the fact that there was no overall gender difference on the paper. Note also, however, the presence of a modest gender-question interaction variance – this reflects the fact that there were some questions among the eight that showed gender differences one way or the other, as mentioned earlier. This variance contributes to both types of measurement, but its contribution is negligible. Between-candidate variance and between-question variance each account for 20-30% of the total variance. But it is the confounded residual variance that contributes the most to total variance, at just under 50%: this component contributes to both relative and absolute measurement error, whereas the between-question variance contributes only to absolute measurement error.

The two generalizability coefficients have acceptable values, at 0.82 for the relative coefficient and 0.76 for the absolute coefficient. Test score SEMs are between three and four marks in each case, with margins of error again similar at between seven and eight marks.

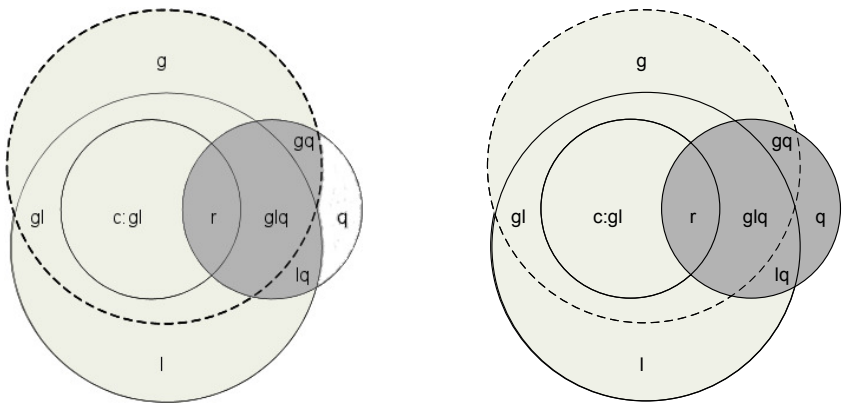
Interactions between gender and questions exist, but they are not serious enough to affect measurement reliability in this case – at least on the strength of the evidence from the eight questions used in 2009. Candidates, though, are also nested within centres, so that it might be interesting to explore the presence or otherwise of interactions between centres and questions, since these would be another potential contributor to measurement error. The number of candidates entered by individual centres varied enormously, from just one candidate to almost 200, and the gender mix within centres varied also. Purely in the interests of design exemplification, and partly for analysis convenience, a balanced data subset was created for analysis through the random exclusion of records: the new dataset consists of 118 comprehensive secondary schools, from within which 10 female candidates and 10 male candidates were selected at random (10 of the 128 schools that entered candidates were unable to provide these numbers for each gender).

Table 5.1
G-study results for GCSE Chemistry foundation tier
(8 variable-metric questions and 10,134 candidates)

Variance component estimates and % contributions	
Gender (-0.0006)	0
Candidates within gender (0.8086)	28
Questions (0.6271)	22
Gender by questions (0.0063)	<1
Confounded residual (1.4015)	49
Generalizability coefficients	
Relative measurement (*Γ*)	0.82
Absolute measurement *(Φ)*	0.76
Standard errors of measurement	
SEM relative	3.4
SEM absolute	4.0
Margins of error	
ME relative	*6.7*
ME absolute	*7.8*

Figure 5.4 illustrates the variance partition for this extended design. Centres are represented by the letter l for locations, retaining c for candidates.

Figure 5.4
Contributions to measurement error in the (c:gl) × q design*



** c represents candidates, g gender (a fixed factor), l locations (i.e. centres) and q questions; r represents the confounded residual variance.*

The G-study results for this extended candidate nesting design are given in Table 5.2. The largest variance components remain those associated with candidates, questions and the confounded residual. These, respectively, account for roughly the same proportions of the total variance as in the analysis that took only gender into account as a candidate nesting variable (Table 5.1). But perhaps the most interesting feature in Table 5.2 is the evidence of small interaction effects between gender and questions and between centres and questions, both of which contribute to error variance for both types of measurement. The reduction in the contribution of the between-candidate

variance and of the confounded residual variance to total variance is attributable to the fact that variance contributions associated with gender and centre have been isolated. The generalizability coefficients themselves, and the SEMs and MEs are virtually unchanged.

Table 5.2
G-study results for GCSE Chemistry foundation level
(8 variable-metric questions, 10 candidates in each gender in each of 118 centres)

Variance component estimates and % contributions	
Gender (0.0011)	<1
Centres (0.1153)	4
Questions (0.6077)	22
Gender by centres (-0.0029)	0
Gender by questions (0.0033)	<1
Centres by questions (0.0775)	3
Gender by centre by questions (0.0033)	<1
Candidates within gender by centre (0.6844)	24
Confounded residual (1.3300)	47
Generalizability coefficients	
Relative measurement (Γ)	0.82
Absolute measurement (Φ)	0.76
Standard errors of measurement	
SEM relative	3.4
SEM absolute	4.0
Margins of error	
ME relative	*6.7*
ME absolute	*7.8*

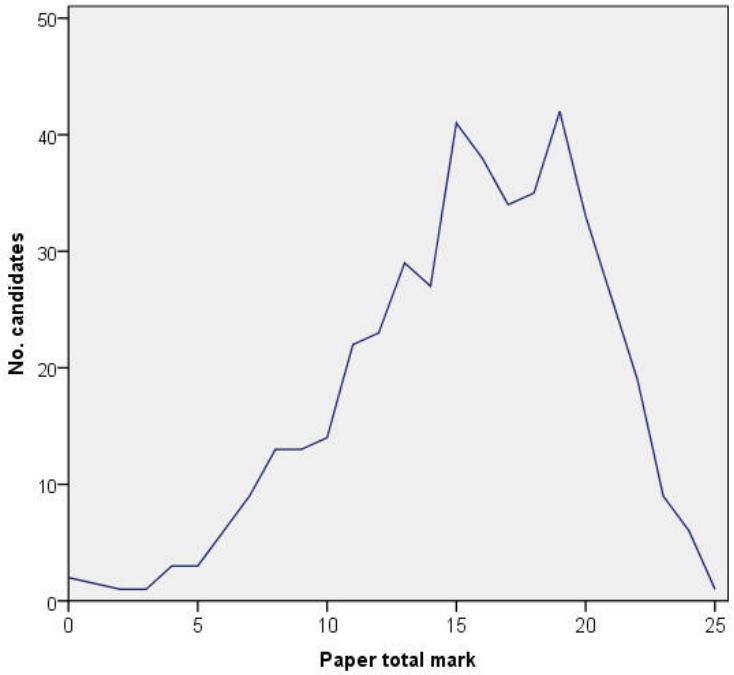
5.3 GCSE ESOL (composite scores)

This second example is a 30-minute listening test, which formed one element in an examination of English for speakers of other languages (ESOL) that also assessed reading and writing skills. The 16-question test was based on three recordings, two taking the form of dialogue and the third an informational lecture. There were five or six questions for each recording. Candidates were given one minute to read the questions relating to each recording before they listened to the recording itself. After a 10-second pause they listened to the recording for a second time, and one minute later started answering the relevant questions. All but three of the 16 questions were dichotomously-scored multiple choice items – the remaining three questions, which comprised four dichotomously-scored short answer subquestions, carried four marks each (subquestion marks were not electronically recorded). The total paper mark was 25.

Data records were available for 450 candidates. Of these, just under 70% were female and 75% were entered for the examination from FE colleges with the rest coming from comprehensive secondary schools. There were 49 centres in total, with between

1 and 86 candidates in each. The mark distribution for the paper is shown in Figure 5.5: the mean mark was 15.5, with male candidates achieving higher on average than female candidates, with 16.0 and 15.3 marks, respectively. This gender difference was fairly consistent across the questions – in other words there was little evidence of any gender-question interaction. Candidates from secondary schools fared better on average than those from FE colleges (means of 17 versus 15, respectively – a statistically significant difference).

Figure 5.5
Mark distribution for the GCSE ESOL listening test
(16 questions and 450 candidates)

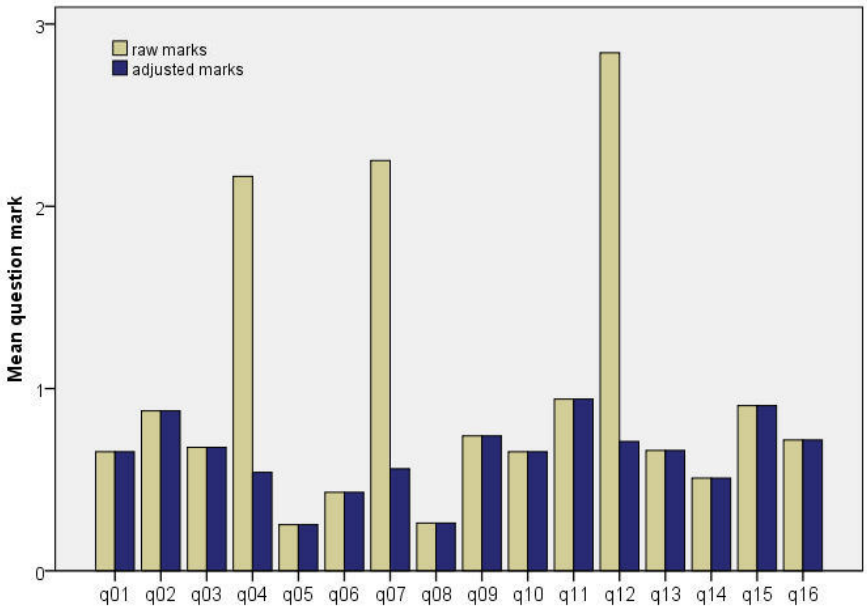


There was a quite high variation in question mean marks, facilities ranging from 25% to 94% for the 13 multiple choice questions, and mean marks from 2.2 to 2.8 for the three short-answer constructed response questions (see Figure 5.6 for profiles).

A principal components analysis with varimax rotation revealed a 5-factor structure for the test, the five factors jointly accounting for just 50% of the total variance (the detail is not offered here, but will be of interest to the examiners who set and marked the paper). This degree of multidimensionality suggests that reliability will not be high. This expectation is confirmed in Table 5.3, which presents the G-study results for the paper, first based on question raw marks and then on adjusted question marks, the 4-mark questions having been adjusted onto the 0-1 scale of the majority.

As Table 5.3 shows, for the raw mark analysis the greatest contribution to score variance (at almost 60%) arises from between-question variation, with the confounded residual contributing another 36%. The relative coefficient is a modest 0.73, while the absolute coefficient falls to 0.51, this latter due to the very high between-question variation, much of which is an artefact of the fact that three of the 16 questions were each marked on 0-4 mark scale *and* had mean marks of 50% or higher (50%, 56%, 71%).

Figure 5.6
Variation in question mean scores for GCSE ESOL listening test
(16 questions and 450 candidates)



When the marks for the three constructed response questions are adjusted onto the binary scale that applies to the other 13 questions we see that the between-question variance is immediately reduced, contributing just under 20% to total variance in place of the previous 60%. The confounded residual, however, has now increased from a 36% contribution to one of 70%. The relative generalizability coefficient has reduced slightly from the previous 0.73 to 0.70. Though still low, the coefficient for absolute measurement has improved from 0.51 to 0.65.

What these analyses have not done, of course, is acknowledge the fact that the questions are implicitly grouped, in the sense that they relate to three different recordings that candidates listened to one after the other before responding. In other words, we have questions nested within stimulus recordings. If we consider the stimuli as well as the questions based on them to be random factors, then Figure 5.7 illustrates the appropriate sources of error variance for relative and absolute measurement.

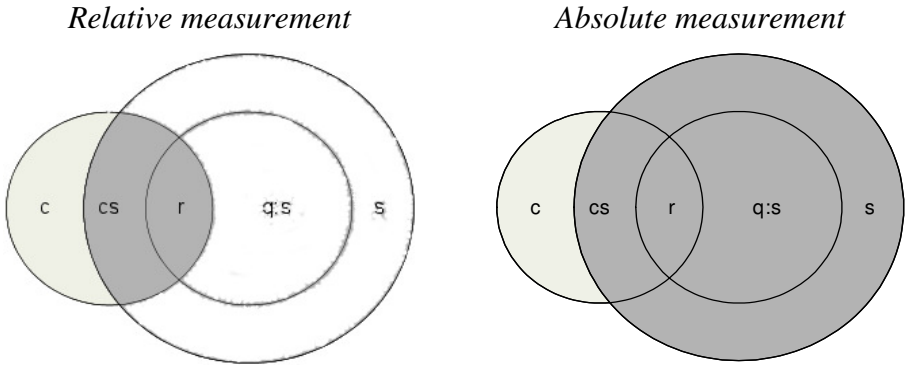
For ease of analysis, one of the multiple-choice questions relating to the first recording was eliminated so that each of the three recordings would be represented by four binary-scored multiple choice questions and one 4-mark multi-part constructed response question. The G-study results are given in Table 5.4.

Interestingly, we see from the results in Table 5.4 that candidates performed similarly across the three different recordings, so similarly in fact that any candidate-recording interaction is too small to make any contribution to total variance, and hence to measurement error.

Table 5.3
G-study results for the GCSE ESOL listening test
(13 binary-scored questions and three 4-mark questions for the 'raw marks' analysis and 16 binary-scored questions for the 'adjusted marks' analysis; 450 candidates)

Variance component estimates and % contributions	*Raw marks*	*Adjusted marks*
Candidates (0.0605, 0.0216)	6	10
Questions (0.5727, 0.0403)	58	19
Confounded residual (0.3586, 0.1464)	36	70
Generalizability coefficients		
Relative measurement (Γ)	0.73	0.70
Absolute measurement (Φ)	0.51	0.65
Standard errors of measurement		
SEM relative	2.4	1.5
SEM absolute	3.9	1.7
Margins of error		
ME relative	*4.7*	*2.9*
ME absolute	*7.6*	*3.3*

Figure 5.7
Contributions to relative and to absolute measurement error in the c × (q:s) design*



** In this design, c represents candidates, s stimulus recordings and q questions within stimulus recordings; all factors are considered random; r represents the confounded residual variance.*

We have so far addressed the reliability of the listening test by analysing raw question marks and adjusted marks. But it would clearly be more correct to recognise that there are two types of question involved here, 13 multiple choice questions worth one mark each and three constructed response questions worth four marks each. In reality there are actually two implicit sections of questions, the grouping being defined by question format. We should take this particular type of question nesting into account in the analysis. This suggests a multivariate G-study (see Brennan, 2001, Chapter 10, for full details, He, 2009, for a summary and simulated example, and Powers & Brennan, 2009, for an example application).

Table 5.4
G-study results for the GCSE ESOL listening test
(four binary-scored questions and one 4-mark question per recording; three recordings; 450 candidates)

Variance component estimates and % contributions	*Raw marks*	*Adjusted marks*
Candidates (0.0651, 0.0198)	6	10
Recordings (-0.1250, -0.0064)	0	0
Candidates by recordings (-0.0120, 0.0021)	0	1
Questions within recordings (0.6807, 0.0448)	61	22
Confounded residual (0.3776, 0.1409)	34	68
Generalizability coefficients		
Relative measurement (Γ)	0.72	0.66
Absolute measurement (Φ)	0.48	0.60
Standard errors of measurement		
SEM relative	2.4	1.5
SEM absolute	4.0	1.7
Margins of error		
ME relative	*4.7*	*2.9*
ME absolute	*7.8*	*3.3*

Paper total scores are the simple sum of section total scores, i.e. sections are equally weighted in the composite score. Table 5.5 provides the G-study results, for each implicit section of questions and for the paper as a whole. For this analysis, and one or two that follow, both GENOVA and EduG were separately used to process the data (in the case of EduG using SPSS to provide the section covariance) – see Chapter 2, Section 2.7, for software access details. In the case of GENOVA sections were given nominal weights that corresponded with the number of questions in each section, to obtain component and coefficient estimates for the total score metric rather than the default mean score metric (see Brennan, 2009, for details). EduG involved some manual calculation to achieve the same result.

As we see from Table 5.5, the set of three constructed response questions actually produced a more reliable subtest than did the set of 13 multiple choice questions, with relative reliability coefficients of 0.68 and 0.60, respectively, and absolute coefficients of 0.64 and 0.53. The margins of error around candidates' subtest scores, however, were almost identical. The reliability of the listening test as a whole was 0.78 for relative measurement and 0.74 for absolute measurement, with margins of error of four to five marks on the 25-mark scale (16% to 19%).

It was unfortunately not possible to explore the reliability of the entire ESOL paper, which also included reading and writing tests, given that the writing assessment was essentially based on a single extended written production from each candidate, permitting no analysis of potential measurement error contributions involving questions (in this case writing stimuli).

Table 5.5
G-study results for the GCSE ESOL listening test
(13 multiple choice questions and three 4-mark questions; unit section weights; 450 candidates)

Variance component estimates and % contributions	*MC questions*	*SA questions*	*Whole paper*
Candidates (0.0190, 0.5879)	8	38	
Questions (0.0488, 0.1342)	21	9	
Confounded residual (0.1672, 0.8406)	71	54	
Generalizability coefficients			
Relative measurement (Γ)	0.60	0.68	0.78
Absolute measurement (Φ)	0.53	0.64	0.74
Standard errors of measurement			
SEM relative	1.5	1.6	2.2
SEM absolute	1.7	1.7	2.4
Margins of error			
ME relative	*2.9*	*3.1*	*4.2*
ME absolute	*3.3*	*3.3*	*4.7*

5.4 GCSE Biology revisited (composite scores)

In Chapter 4 (Section 4.3) we described a 30-minute 9-question foundation tier GCSE Biology paper that was made up of five matching questions worth four marks each and five other questions each comprising four binary-scored multiple choice items (subquestions). In that chapter, for the sake of simplicity, we ignored the subquestion marks, which were in fact made available to us, and analysed the paper as a 9-question paper with nine equal-metric questions (0-4 mark scale).

Here, we use all of the performance information that is contained in the question and subquestion marks, and explore the reliability of the composite scores for the paper as a whole, with the two implicit sections carrying equal weight in the 36-mark paper total, i.e. the whole-paper mark is the simple sum of the section totals. The results are shown in Table 5.6.

The analysis for the matching questions is based on the design c × q whereas that for the 4-item multiple choice questions is based on the design c × (i:q), the latter reflecting the fact that each question nests four separate items. A first interesting feature in Table 5.6 is the absence of any indication of a between-question effect. Although the binary-scored items were presented four to a question this nesting property could have been ignored in the analysis, so that the analysis could have been based on the simpler c × q design used for the matching questions, this time q representing binary-scored items and not 4-mark questions.

At around 0.7, the reliability coefficients for the paper as a whole are higher than those of the two sections, but still modest. SEMs for the whole paper are around three marks for both types of measurement, while margins of error, at around six marks, are around 17% of the mark scale compared with 20-25% for the separate sections.

Table 5.6
G-study results for the GCSE Biology paper
(Five 4-mark matching questions for 20 marks and four questions each containing four multiple choice items for 16 marks; unit section weights; 24,666 candidates)

	Matching questions	*4-item questions*	*Whole paper*
Analysis design	c × q	c × (i:q)	
Variance component estimates and % contributions			
Candidates (0.2861, 0.0177)	17	7	
Questions(0.2470, -0.0027)	15	0	
Candidates by questions (, -0.0001)		0	
Items within questions (, 0.0369)		15	
Confounded residual (1.1139, 0.1990)	68	78	
Generalizability coefficients			
Relative measurement (Γ)	0.56	0.59	0.72
Absolute measurement (Φ)	0.51	0.55	0.68
Standard errors of measurement			
SEM relative	2.4	1.8	3.0
SEM absolute	2.6	1.9	3.2
Margins of error			
ME relative	*4.7*	*3.5*	*5.9*
ME absolute	*5.1*	*3.7*	*6.3*

It would be possible to calculate the SEMs and margins of error for candidates within two percentage points of a boundary mark, as illustrated for this paper in Section 4.3 of Chapter 4. It would also be possible through *what if* analyses to explore the effect on reliability and score precision of changing the mix of matching questions and multiple choice items in the paper. Such an analysis would simply be an extension of the *what if* analyses described in Chapter 4, the number of questions in each section being modified, and the impact on section reliability and then on whole-paper reliability estimated in the usual way. An example is given below in Section 5.6.

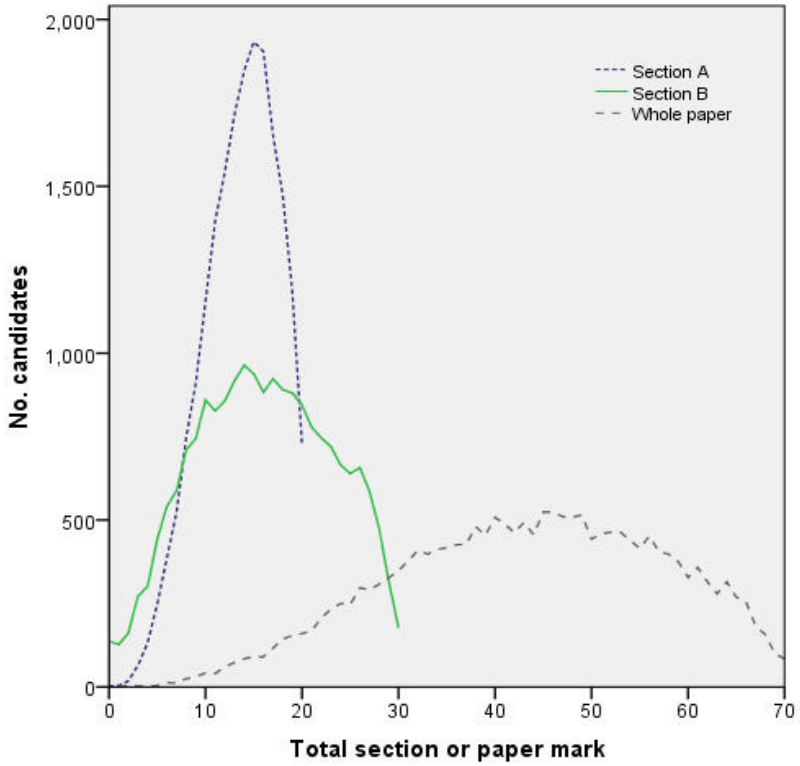
5.5 GCE Mathematics (sectioned and non-sectioned papers)

Our first mathematics example is a 2-hour 2-section GCE paper. Section A was a 20-item objective test comprising 4-option dichotomously-scored multiple choice items. Section B comprised four structured questions, with two or three parts each and with variable part and whole-question marks (the latter being 8, 8, 5 and 9). Section A was double weighted to contribute 40 marks of the paper maximum mark, with Section B contributing 30 marks. The maximum paper mark was 70, and there was no question choice.

Full data records were available for just under 20,000 candidates. Of these, almost 50% were female and almost 90% were entered for the examination from

comprehensive secondary schools with most of the rest coming from independent secondary schools. The whole-paper mark distribution is shown in Figure 5.8.

Figure 5.8
Mark distributions for the GCE Mathematics paper and its two sections
(20 binary items in Section A for 20 marks and four variable mark questions in Section B for 30 marks; 19,566 candidates; section weights 2 and 1, respectively)



The mean mark for the paper was 43.3 (62%), with no significant difference in the overall performances of male and female candidates (43.4 and 43.1 marks, respectively). Item facilities in Section A varied from 0.31 to 0.98, and question means in Section B varied from 38% (for the lowest-weighted question) to 70%. There was also evidence of gender-question interaction, with statistically significant gender differences for some items in one direction or the other (though with such high candidate numbers some differences are likely to reach significance). Across the paper as a whole three principal components accounted for just 30% of the variance, while each section had a 2-factor structure, with the first two principal components accounting for less of the variance in Section A than in Section B, at just over 25% and over 45%, respectively. The principal examiner and team leaders will find the detail behind these overall results particularly interesting.

The G-study results for the paper are given in Table 5.7, from which we see that of the two sections, Section A, the 20-mark objective test, is in principle the more reliable subtest, despite a higher proportional contribution to total variance of the confounded residual, at almost 75% compared with just over 40% for Section B. It is Section A that contributes most to the whole-paper mark. The outcome is very acceptable reliability coefficients for the paper as a whole, with 0.86 for relative

measurement and 0.81 for absolute measurement. Relative and absolute SEMs for whole-paper marks remain quite small relative to the total mark scale, at around five and six marks, respectively, for relative and absolute measurement, giving margins of error of around 14% and 17% of the mark scale.

.

Table 5.7
G-study results for the 2-section GCE Mathematics paper
(Section A, 20 binary-scored multiple choice questions for 20 marks, and Section B, four variable-mark constructed response questions for 30 marks; 19,566 candidates; section covariance 0.255; section weights 2 and 1, respectively)

Variance component estimates and % contributions	*Section A (MC)*	*Section B (CR)*	*Whole paper*
Candidates (0.0284, 2.3466)	13	29	
Questions (0.0294, 2.3987)	13	30	
Confounded residual (0.1606, 3.3261)	74	41	
Generalizability coefficients			
Relative measurement (Γ)	0.78	0.74	0.86
Absolute measurement (Φ)	0.75	0.62	0.81
Standard errors of measurement			
SEM relative	1.8	3.6	5.1
SEM absolute	2.0	4.8	6.2
Margins of error			
ME relative	*3.5*	*7.1*	*10.0*
ME absolute	*3.8*	*9.4*	*12.2*

It will be interesting to compare the results for this structured paper with those for a series of non-structured AS and A2 unit papers offered by a different board in the same year. Each paper contained nine or 10 structured questions, with variable total marks and variable numbers of parts and part marks. The maximum paper mark was 75 in each case, with no question choice.

The size of the candidate entry varied from one paper to another: between around 1,700 and 3,500. This variation reflects the flexibility candidates now have in unitised examinations in terms of when they elect to take particular units, and the fact that AS papers would typically be taken before A2 papers – many candidates who were entered for one or both of the AS papers in 2009 would not yet have had the opportunity to move on to study for and to take the A2 units required for a full A level. Just over 40% of the candidates in each case were female, and in all four cases the majority of candidates, 70-80%, were entered from comprehensive secondary schools.

The mark distributions for the four papers are shown in Figure 5.9: an immediate feature to note is the close similarity in shape shown by the four mark distributions. The distributions are similarly left-skewed with almost identical mean marks (between 43 and 46 on the 75-mark papers), presumably reflecting the question setting experience and style of the principal examiner.

Female candidates produced the better average performances on the two AS papers, both overall and consistently across questions. There were no overall gender differences on either of the A2 papers, although female candidates outperformed male candidates on one or two of the questions. Principal component analyses revealed that in each case the first principal component accounted for 50-65% of the variance.

Figure 5.9
Mark distributions for four GCE Mathematics papers
(9-10 variable mark questions per paper; candidate numbers varying from 1,377 to 3,521)

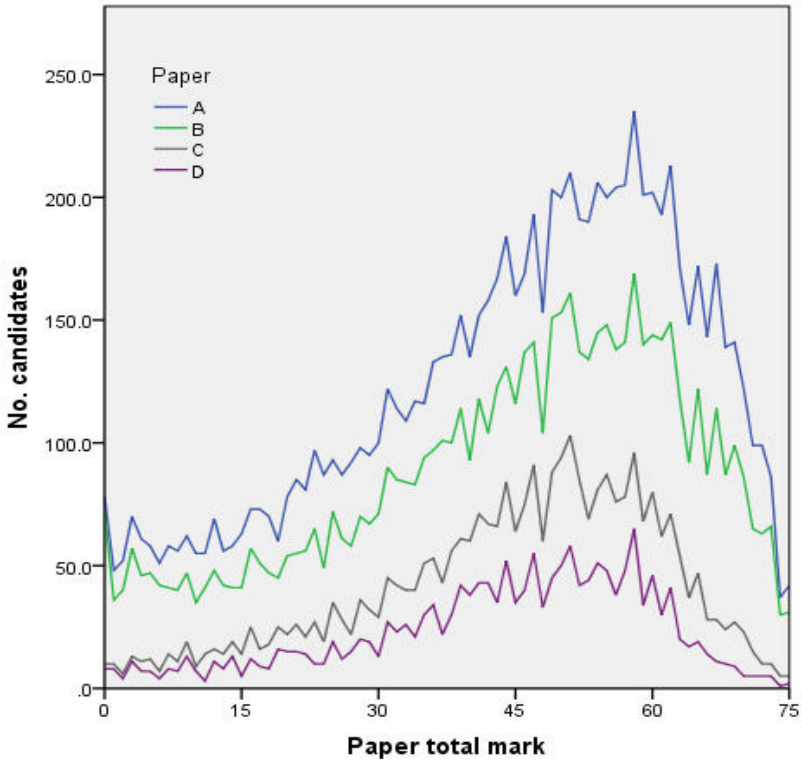


Table 5.8 presents the G-study results for the four papers. The contributions to total variance of the three factors, *viz*. candidates, questions and the confounded residual, which embraces candidate-question interaction, are roughly equal within and across the papers, each contributing 30-45% depending on the paper. This reflects the very similar pattern of variance contribution shown by Section B (structured-questions) in the previous 2-section mathematics paper. In comparison with the previous composite-score paper, we see in Table 5.8 slightly higher reliability coefficients for relative measurement for the AS unit papers, at just over 0.9 in each case, and closely similar values for the A2 papers. Absolute coefficients for the AS papers are slightly higher than for the structured paper, and lower for the A2 papers. SEMs are similar across the four papers, giving margins of error of 10-11 marks for relative measurement and 15-17 marks for absolute measurement on the common 75-mark scale (roughly 13-15% and 20-23%, respectively, of the full mark scale).

Table 5.8
G-study results for four AS/A2 Mathematics Unit papers
(9-10 variable mark questions per paper; candidate numbers 2,628, 3,521, 1,377 and 1,723 respectively)

Variance component estimates and % contributions	*Paper A*	*Paper B*	*Paper C*	*Paper D*
Candidates (3.1617, 5.096, 2.2382, 2.2012)	35	40	27	23
Questions (3.0357, 3.6322, 2.9431, 4.2848)	33	29	36	44
Confounded residual (2.8616, 3.9882, 3.1009, 3.1674)	32	31	37	33
Generalizability coefficients				
Relative measurement (Γ)	0.92	0.92	0.88	0.87
Absolute measurement (Φ)	0.84	0.86	0.79	0.75
Standard errors of measurement				
SEM relative	5.3	6.0	5.6	5.6
SEM absolute	7.7	8.3	7.8	8.6
Margins of error				
ME relative	*10.4*	*11.8*	*11.0*	*11.0*
ME absolute	*15.1*	*16.3*	*15.3*	*16.9*

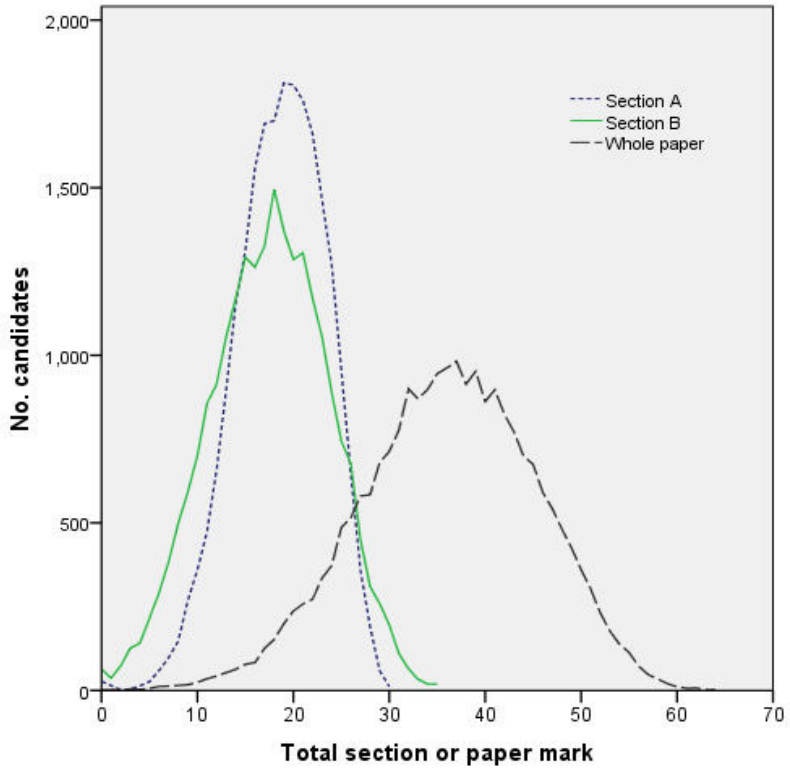
5.6 GCE General Studies (comparative composite scores)

In Chapter 4 (Section 4.4) we looked at one section of a 2-section GCE General Studies AS Unit paper. That section (Section A) was an objective test comprising 30 multiple choice items for 30 marks. The second section, Section B, was composed of three constructed response questions for a total of 35 marks: one question, which comprised two short answer questions carrying four marks each, was awarded an 8-mark total, the other two, each of which required extended written responses, carried 13 and 14 marks, respectively. Here we consider the paper as a whole, whose total mark – the simple sum of the section marks – was 65. The mark distributions for each section and for the paper are shown in Figure 5.10.

Table 5.9 presents the G-study results for each section, and for the entire paper. It should be noted before reviewing the analysis results that the analysis could not take account of any possible marker influence on the results for Section B, since the paper was single marked, i.e. each candidate script was marked by one marker only, as is current practice with all GCSE and GCE examination papers (after marker standardisation). That said, one of the most interesting features of the data in Table 5.9 is the way that the total score variance is constituted within each section.

As we saw in Table 4.5 in Chapter 4, in Section A (the objective test) the largest estimated variance component is that associated with the confounded residual, i.e. candidate-question interaction combined with random fluctuations and any other unidentified systematic variance. This accounted for fully 88% of the total score variance, compared with well under 10% for the between-candidate variance and for the between-question variance. Indeed question interactions with other factors were so strong that seven principal components jointly accounted for just 34% of the total variance. One of these interacting factors was gender.

Figure 5.10
Mark distributions for the GCE General Studies paper and its two sections
(Section A, 30 multiple choice questions for 30 marks; Section B, three variable-mark constructed response questions for 35 marks; paper total mark 65; 22,424 candidates)



For Section B we see a different picture, with between-candidate variance contributing over 40% of the total variance, and the confounded residual contributing just over 50%. The low between-question variance in both sections means that the reliability of absolute candidate measurement will be little different from the reliability of relative candidate measurement – as indeed is confirmed in Table 5.9, where both coefficients for both sections are around 0.7.

Score reliability for the whole paper is higher than that for each section. For both types of measurement the whole-paper reliability coefficient is at or around 0.8, compared with section reliabilities of around 0.7. Composite score precision is also better, with an SEM of just over four marks for both types of measurement, and a margin of error of eight marks on the 0-65 mark scale for relative measurement (around 12% of the scale) and just under 8.5 (13%) for absolute measurement.

Table 5.9
G-study results for the GCE General Studies Unit
(Section A, 30 multiple choice questions for 30 marks, and Section B, three variable-mark constructed response questions for 35 marks; unit section weights; 22,424 candidates)

Variance component estimates and % contributions	*Section A (MC)*	*Section B (CR)*	*Whole paper*
Candidates (0.0165, 3.0177)	7	42	
Questions (0.0123, 0.4482)	5	6	
Confounded residual (0.2065, 3.6477)	88	51	
Generalizability coefficients			
Relative measurement (Γ)	0.71	0.71	0.80
Absolute measurement (Φ)	0.69	0.69	0.78
Standard errors of measurement			
SEM relative	2.5	3.3	4.1
SEM absolute	2.6	3.5	4.3
Margins of error			
ME relative	*4.9*	*6.5*	*8.0*
ME absolute	*5.1*	*6.9*	*8.4*

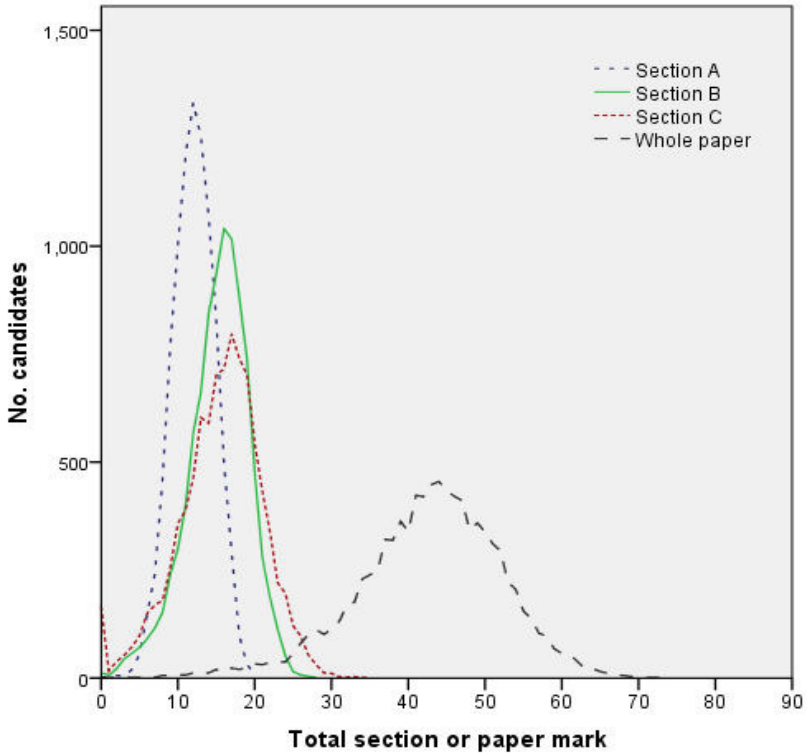
It will be interesting to explore the corresponding characteristics of an alternative GCE General Studies AS Unit paper, also offered in 2009 but by a different examining board. The total candidate entry for this unit paper was just over 9,000, of which over 55% were female: just over 60% of candidates were entered from comprehensive secondary or middle schools, just over 20% from selective secondary schools and over 10% from sixth form colleges.

Like our first example, this paper contained an objective test as Section A, this time comprising 20 items. But now we have two further sections. Section B comprised seven short answer questions, one with two subquestions, based on an informational text. The questions had different maximum marks: from 1 to 4 for the first five questions, and 8 marks each for the final two extended response questions, which invited candidates to evaluate arguments contained in the text. Section C comprised two extended response questions, worth 20 marks each, which focused on major issues in society: candidates were invited in one question to critically examine an assertion and in the other to evaluate opposing arguments. The three sections contributed, respectively, 20 marks, 30 marks and 40 marks to the paper total of 90 marks. The mark distributions for the whole paper and for its three sections are shown in Figure 5.11. The whole-paper composite score was again the simple sum of the three section total scores.

We see from the results in Table 5.10 that all three sections showed different patterns of variance contribution. For Section A and Section B the contribution of candidate variation to total variance was almost negligible, at 5% or less; Section C, on the other hand, shows a 30% contribution for between-candidate variance. In Sections A and C the highest variance contribution is the confounded residual, at around 70%. For Section B, which comprised variable-mark questions, between-question variation

accounted for the largest part of the score variation, with a contribution of 60%, followed by the confounded residual variance, at 35%. In Section C we see no contribution to total variance at all coming from differences in the mean scores of the two essay questions, and indeed the question means were almost identical. However, while the two questions might have produced closely similar mean scores candidates did not in general show consistent performances on both, hence the high residual variance, which contains candidate-question interaction.

Figure 5.11
Mark distributions for the alternative GCE General Studies paper and its three sections
(Section A - 20 multiple choice questions for 20 marks, Section B - seven variable-mark constructed response questions for 30 marks, Section C – two extended response questions for 40 marks; total paper mark 90; 9,324 candidates)



While reliability coefficients were uniformly low for each section, at the whole-paper level score reliability is good for relative measurement, at around 0.7, if rather low for absolute measurement, at just under 0.6. The precision of individual candidate scores was quite acceptable: SEMs for the paper as a whole are quite low, at around five marks for relative measurement and six and a half marks for absolute measurement. In consequence the margins of error are also comfortingly small, at 10 marks for relative measurement and under 13 marks for absolute measurement, translating to just around 11% and 14%, respectively, of the 90-mark scale. Thus, despite the lower reliability coefficients for this paper compared with the one described empirically in Table 5.9, the most essential indicator of reliability – score precision – is virtually identical in the two cases.

Table 5.10
G-study results for an alternative GCE General Studies Unit paper
(Section A - 20 multiple choice questions for 20 marks, Section B - seven variable-mark constructed response questions for 30 marks, Section C – two extended response questions for 40 marks; unit section weights; 9,324 candidates)

Variance component estimates and % contributions	*Section A (MC)*	*Section B (CR)*	*Section C (ER)*	*Whole paper*
Candidates (0.0099, 0.1706, 3.4142)	4	5	30	
Questions (0.0577, 1.9318, 0.0000)	24	60	0	
Confounded residual (0.1752, 1.1141, 7.9881)	72	35	70	
Generalizability coefficients				
Relative measurement (*Γ*)	0.53	0.52	0.46	0.69
Absolute measurement *(Φ)*	0.46	0.28	0.46	0.59
Standard errors of measurement				
SEM relative	1.9	2.8	4.0	5.2
SEM absolute	2.2	4.6	4.0	6.5
Margins of error				
ME relative	*3.7*	*5.5*	*7.8*	*10.2*
ME absolute	*4.3*	*9.0*	*7.8*	*12.7*

Table 5.11 provides an illustration of how *what if* analysis might be used to explore different paper structures, in this case in terms of different section sizes.

Table 5.11
***What if* results for different alternative General Studies paper structures, each with a 90-mark total**

	*Original paper**	*Different section lengths***	*Alternative section lengths****
Generalizability coefficients			
Relative measurement (*Γ*)	0.69	0.69	0.70
Absolute measurement *(Φ)*	0.59	0.61	0.68
Standard errors of measurement			
SEM relative	5.2	5.2	5.0
SEM absolute	6.5	6.2	5.3
Margins of error			
ME relative	*10.2*	*10.2*	*9.8*
ME absolute	*12.7*	*12.2*	*10.4*

** Section A - 20 multiple choice questions for 20 marks; Section B – seven variable-mark constructed response questions for 30 marks, Section C – two extended response questions for 40 marks.*
** Section A increased to 30 items for 30 marks, Section B decreased to 5 questions for a total of 20 marks; Section C unchanged – two extended response questions at 20 marks each for a total of 40 marks.
*** Section A – 50 multiple choice questions; Section B eliminated; Section C unchanged – two extended response questions at 20 marks each for a total of 40 marks.

As Table 5.11 shows, increasing the length of Section A, the objective question section, by 50% to 30 items, whilst decreasing the length of Section B, the structured-question section, from seven to five questions would have negligible effect on reliability outcomes, in terms either of the reliability coefficients or of SEMs. Some improvement, albeit modest, would result from increasing Section A further to 50 items and eliminating Section B altogether. The predicted reliability coefficients increase slightly to around 0.7 for both types of measurement, while the predicted SEMs and margins of error for the 90-mark scale decrease slightly, particularly for absolute measurement.

It would be up to principal examiners to decide what alternative section sizes and weights might be acceptable in validity terms. G-theory would allow the impact on reliability and score precision of the different possibilities to be explored.

6 Summary and reflections

6.1 Introduction

Throughout Chapters 4 and 5 we have presented the results of univariate and multivariate generalizability analyses for a range of GCE and GCSE examination subjects, exhibiting a variety of different paper structures. The findings are illuminating, and sometimes surprising. But what are the salient features, and what might be the implications for the work of the examining boards in the future? We offer reflections on these questions here.

As far as the analysis results are concerned, the mark distributions for the various sections and papers show an interesting variation in shape and scale coverage. This is the first issue that we discuss below, given the contribution that mark variation among candidates has on assessment reliability and also the impact that the underlying mark distribution must have on rates of candidate misclassification during the grading process. We then reflect on the pattern of reliability results, and consider the meaning and relative value of reliability coefficients and of standard errors of measurement in this context.

Finally, we consider some of the implications for further validity and reliability research.

6.2 Mark distributions

The mark distributions associated with the various component papers merit comment. A primary outcome of the work of the examining boards is the classification of examination candidates into achievement grades. Grade awarding is a complex process (see Robinson, 2007, and Crisp, 2010, for descriptions) that continues to embody a high degree of norm referencing practice tempered by criterion-referenced examiner judgement, as noted by Christie & Forrest, 1981, and Orr and Nuttall, 1983, almost 30 years ago in pre-GCSE days. Senior examiners review a wide range of qualitative and quantitative information about the current paper and past papers, and about candidate performances, to come to decisions about appropriate cut scores for critical grade boundaries, for example A/B at GCE and C/D at GCSE. Once the two relevant critical grade boundary marks have been determined, partly judgmentally on the basis of the quality of candidates' work and partly statistically on the basis of proportions of candidates achieving different grades in previous years, intervening grade boundaries are determined entirely empirically: the range of marks between the critical boundary marks is proportionately divided to produce the boundary marks for intermediate grades.

The ideal mark distribution for this kind of grade awarding practice would be a rectangular distribution, in which candidates are spread evenly across the entire mark scale for the paper concerned. While the unavoidable presence of measurement error in test scores makes a degree of misclassification inevitable in any grading procedure, a rectangular distribution would at least offer the possibility equalising to some extent the likelihood of misclassification around the different boundary marks. But rectangular distributions are rare in practice, if indeed they occur at all in educational assessment.

The next best option is a relatively flat bell-shaped distribution spanning the full mark range. Such distributions can be achieved when test papers are put together using questions whose empirical properties are known beforehand through pretesting with a representative sample of future examination candidates. But, for understandable reasons to do with test security, there is little pretesting or test piloting for GCSE or GCE examinations, and none for structured questions, essay papers or practical tasks. Principal examiners, who work in that role for many years, compile test papers following a set paper specification, which generally controls the degree of curriculum coverage within the paper and the relative importance to be given to different aspects both within the paper and within the mark scheme. But there is no opportunity to trial the resulting paper in the field prior to live use. The test distributions associated with component papers are therefore known only after the papers have been used operationally. This might explain some of the variety of distributional patterns seen in Chapters 4 and 5.

Some distributions span the whole mark scale but are to a greater or lesser degree skewed one way or the other. Others are symmetric but peaked, leaving portions of the mark scale unused. When mark distributions are truncated, the likelihood of candidates being misgraded increases, especially when measurement error is high and distances between grade boundaries are measured in single-digit marks (for an interesting historical discussion of issues concerning GCE grading, issues which remain to this day, see Whittaker & Forrest, 1983).

For foundation tier GCSE papers the grade range is C to G, for higher tier GCSE papers and for GCE papers it is A^* to E; grade classification is now being applied to performances on individual unit papers within the parent examinations and is therefore dependent on candidates' total marks on that paper. Unit papers inevitably have shorter mark scales than multi-paper examinations had. This means that the shape of the mark distribution produced when a particular unit paper is used operationally is in consequence an even more critical factor than before in determining the quality of candidate grading.

We were unable within the scope of this project to consider corresponding papers in the same GCSE and GCE subjects from years prior to 2009, and so we cannot say whether the distributions shown in this report are typical or not for those subjects. If they are typical then the responsible principal examiners will be aware of that fact, and could take steps to change the shape of distributions in future examinations where there would be no obvious threat to assessment validity.

6.3 Reliability coefficients, SEMs and confidence intervals

Generalizability theory is based on a sampling model through which the impact of sampled factors on assessment (un)reliability can be assessed and, at least as importantly, predicted. Important sampled factors include markers and test questions. Their impact can be evaluated through a generalizability study, or G-study, analysis, which ideally incorporates as many as possible of the factors that are hypothesised to influence candidates' test scores, including factor interactions, and quantifies their relative contributions to measurement error. This quantified information is then used in *what if* analyses, or D-studies, to predict the effect on reliability of changes in sample sizes, which might be marker numbers, question numbers, question numbers within different paper sections, and so on.

Generalizability coefficients, like most reliability coefficients, are ratios of valid variance to the combination of valid and error variance. The flatter a total mark distribution for a unit paper, and the more of the notional mark scale that is used, the greater will be the variation in candidates' test scores. In principle, therefore, a flat full-scale distribution is a positive ingredient for assessment reliability in this sense. But high candidate score variation is not itself sufficient to guarantee reliability. Measurement error variance is also relevant, and this can be more or less complex in composition, depending on the type, length and structure of the test paper, and the assumptions made by the researcher about which factors and factor interactions contribute to error variance and which do not.

Test questions are a potentially important source of measurement error, as are markers. Just as in any sampling application, the higher the between-question variation, and the less performance consistency individual candidates show across questions, the more questions will be needed in a test to counter the effect on measurement error when classifying candidates using cut scores. For the same reason, the greater the variation in markers' overall standards and marking consistency the more markers should be required to independently mark the work of candidates, so that the impact of this source of score variance is minimised through averaging across markers.

In norm-referencing applications, such as current examining board grading procedures, interaction effects involving candidates are in principle the only potential sources of measurement error, since it is only these that determine the rank order of individual candidates in any particular examination sitting. But in practice this is not necessarily true. For when different candidate scripts are marked by different markers, then unless those markers are genuinely interchangeable any between-marker differences in standards remain a relevant source of error variance. This is because the mark distributions produced for different groups of candidates by different markers are merged to produce the overall distribution for the candidate entry as a whole. Unfortunately, the impact of this source cannot be explored and quantified when the candidate groups processed by different markers might themselves differ in important ways: for example, when whole centres are assigned to particular markers, so that marker effects are confounded with centre effects.

We were unfortunately unable to explore marker effects alongside question effects in Chapters 4 and 5 precisely because operational examining is based on single marking of scripts, after marker standardisation. We were fortunate, on the other hand, to be given access to the one set of relevant data that we have explored in Chapter 3, which emanated from a marker standardisation exercise. The $c \times q \times m$ model applied to that set of data is the one that should routinely be used in pre-operational marker studies, to screen papers in order to identify which would most benefit from multiple-marking in an operational context of budgetary and other constraints.

But what can we say about the relationships between mark distributions, generalizability coefficients and SEMs on the basis of the analysis results offered in Chapters 4 and 5? Relationships are complex, for the reasons given above. But papers showing similarly-shaped distributions spanning similar mark ranges showed similar patterns of variance partition, and produced similar reliability coefficient values and

SEMs: look for example at the results for the GCSE French listening test (Section 4.9) and the GCSE Music paper (Section 4.8), and, alternatively, at those for the GCSE Drama paper (Section 4.5) and the GCSE Business Studies papers (Section 4.2), or those for the Mathematics papers described in Section 5.5.

A general finding from the analysis results is that margins of error for relative measurement tended to have values in the range 5% to 18% of the underlying mark scale, and most typically 15% to 17%. This means that, under Normal distribution assumptions, 95% confidence intervals around candidates' total paper marks will be of the order of 10% to 40% of the mark scale, clearly spanning several grade boundaries in some cases.

6.4 Implications for further assessment research

In this research report we have offered examples of generalizability analysis for a variety of written component papers. We encountered some anticipated problems. The first problem had to do with papers in which the different test questions merit different maximum marks, and where there were too few questions sharing the same maximum mark to permit analysis as an informally sectioned paper with composite scores as the outcome: at least two questions of any particular type, whether it be format or weighting, are required for variance analysis to be feasible. The consequence in such cases is that while reliability coefficients can nevertheless be calculated for the paper, they cannot be meaningfully extrapolated to an alternative paper, past or present. In other words, when there is a particular distribution of question mark allocations within a paper, when individual questions are differentially weighted, then it is difficult to appeal to the notion of domain sampling for generalisation of findings.

Generalizability theory assumes equal metrics for what are essentially sampled questions, at least within multi-question sections. We experimented by adjusting questions onto a common metric for a repeat analysis. Sometimes reliability was improved and sometimes not. But even when reliability improves it does not necessarily follow that mark adjustment is an acceptable strategy. Principal examiners set papers to match given specifications, and if those specifications expressly demand that some question types, or some topics or skills, should be awarded higher weight than others in an examination then modifying the weight distribution would risk jeopardising assessment validity. The question is how intentional are the different mark allocations in validity terms? This is a question that merits debate and perhaps research. It is frequently claimed that reliability can only be increased at the expense of validity, and that since validity is paramount we must sometimes accept inadequate levels of reliability in examinations. This is not true. Reliability can be increased without jeopardising validity, either by scrutinising and potentially redefining 'validity' or by increasing assessment costs and resolving logistic challenges in order to provide more assessment evidence about candidates and so improve reliability. But first we need to be able to evaluate reliability itself, and to be able to manipulate it.

The second problem that we encountered was that for some component papers it did indeed prove impossible even to quantify reliability, least of all to explore ways of improving it. Typical examples included English essay papers in which candidates were required to produce a single essay in response to a choice of themes, and foreign language papers in which a single-task writing assessment featured as one section.

This phenomenon extends to most practical examinations and to units based on portfolios, which, for similar reasons, we were not able to explore. Pre-operational research would be needed to investigate impacts on reliability for such examination papers, in which 'tasks' is a hidden factor that potentially threatens the validity of generalisation of candidate performances beyond the single tasks actually used and evaluated in the examination.

Ideally, G-studies, and follow-on *what if* analyses, should be set up and analysed during an ongoing research and development programme, in which marker effects and paper structures are simultaneously investigated, along with gender and other candidate-related effects. This would allow investigation of different alternative paper formats and lengths before component papers, whether written or practical, are used operationally. It would also help to focus attention on those component papers that would most benefit from, indeed would essentially demand, double marking in place of the single marking that currently prevails. Where pretesting is impossible to implement then the analyses could simply be based on existing operational data where appropriate.

Examining boards have conducted extensive marker reliability studies over the past 20 years or so (see Meadows & Billington, 2005, for a recent comprehensive review). Investigations into test-related influences on reliability, however, have received much less attention in that time period. The two aspects could now usefully be brought together, with the $c \times q \times m$ design becoming the default in place of the $c \times q$ design that we were constrained to appeal to in this report, or the $c \times m$ design that has typically underpinned marker studies in the past. The fact that question-level data are now being routinely electronically recorded by all examining boards might be expected to facilitate this move to a more comprehensive approach to reliability investigation.

But this assumes that the examining boards have in-house capacity for this kind of research, in terms of technical expertise, or have the flexibility to access external expert research assistance. It also assumes efficient archiving of cumulating operational data, access to appropriate user-friendly software, which remains a universal problem, and the time to carry out the volume of analyses that are required. On the latter point, the examination system in the UK has been subject to continuing evolution over the past two decades, partly in response to growing public demand for more varied and more flexible qualification possibilities, but partly also because of government-driven pressure to be constantly innovating. This has led to major changes in the GCSE and the GCE systems every five years or so, often leaving the examining boards with frighteningly short timescales in which to implement demanded innovations - see Baird & Lee-Kelley, 2009, for a fascinating account of the situation and ensuing pressures. A consequence has been that research resources within the boards have been virtually monopolised by new qualification development, leaving little time for research of the kind described in this report, that could suggest ways of improving existing examinations as well as offering pointers for the optimal design of new offerings. This is an issue meriting attention.

References

Baird, J-A. & Lee-Kelley, L. (2009). The dearth of managerialism in implementation of national examinations policy. *Journal of Educational Policy, 24*, 55-81.

Brennan, R.L. (1992). *Elements of Generalizability Theory* (Second edition). Iowa City: ACT Publications (First edition: 1983).

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

Brennan, R.L. (2001b). *Manual for urGENOVA.* Version 2.1. Iowa: Iowa Testing Programs. Occasional Papers, 49

Brennan, R.L. (2001c). *Manual for mGENOVA.* Version 2.1. Iowa: Iowa Testing Programs. Occasional Papers, 50

Brennan, R.L. (2009). *Notes about Nominal Weights in Multivariate Generalizability Theory.* CASMA Technical Note: No.4. Iowa City: University of Iowa Center for Advanced Studies in Measurement and Assessment. (Available on http://www.education.uiowa.edu/casma)

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

Christie, T. & Forrest, G.M. (1981). *Defining Public Examination Standards.* London : Macmillan.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, 78(1), 98-104.

Crick, J.E. & Brennan, R.L. (1983). *Manual for GENOVA: a GENeralized Analysis Of VAriance System.* Version 2.1. Iowa: American College Testing Program.

Crisp, V. (2010). Judging the grade: exploring the judgement processes involved in examination grading decisions. *Evaluation and Research in Education, 23*, 19-35.

Cronbach, L.J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L.J. & Shavelson, R. (2004). My current thoughts on Coefficient Alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391-418.

Dobson, A.J & Barnett, A.G. (2008). *An introduction to Generalized Linear Models*. (Third edition). Boca Raton: Chapman & Hall/CRC.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

He, Q. (2009). *Estimating the reliability of composite scores*. Coventry: Office of the Examinations and Qualifications Regulator (Ofqual).

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics, 9*, 226-252.

Hogan, T.P, Benjamin, A. & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Johnson, S. & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Coventry: Office of the Qualifications and Examinations Regulator (Ofqual).

Kane, M.T. & Case, S.M. (2000). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221-240.

Lord, F.M. & Novick, M.R (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

McCulloch, C.E., Searle, S.E. & Neuhaus, J.M. (2008). *Generalized, Linear and Mixed Models*. (Second edition). Hoboken: Wiley.

Meadows, M. & Billington, L. (2005). *A review of the literature on marking reliability.* London: National Assessment Agency.

Orr, L. & Nuttall, D. L. (1983). *Determining standards in the proposed single system of examining at 16+*. London: Schools Council.

Powers, S. & Brennan, R. L. (2009). *Multivariate generalizability analyses of mixed-format exams*. CASMA Research Report: No.29. Iowa City: University of Iowa Center for Advanced Studies in Measurement and Assessment. (Available on http://www.education.uiowa.edu/casma)

R Development Core Team, (2005). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raju, N. S., Price, L. R., Oshima, T. C. & Nering, M. L. (2007). Standardised conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31*, 169-180.

Rashbash, J., Steele, F., Browne, W.J. & Goldstein, H. (2009). *A user's guide to MLwiN, v2.10.* Bristol: Centre for Multilevel Modelling, University of Bristol.

Robinson, C. (2007). Awarding examination grades: Current processes and their evolution. In Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (eds), *Techniques for monitoring the comparability of examination standards,* 97-123. London: Qualifications and Curriculum Authority.

Searle, S.R., Casella. G. & McCulloch, C.E. (2006). *Variance Components*. (Second edition). Hoboken: Wiley.

Shavelson, R. & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

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

Snedecor, G.W. & Cochran, W.G. (1989). *Statistical Methods* (Eighth edition). Ames, Iowa, Iowa State University Press.

Verhelst, N.D. (2000). *Estimating variance components in unbalanced designs*. R&D Notities, 2000-1. Arnhem. Cito.

Wheadon, C. & Béguin, A. (2010). Fears for tiers: are candidates being appropriately rewarded for their performance in tiered examinations? *Assessment in Education*, 17(3), 287-300.

Whittaker, R. J. & Forrest, G.M. (1983). *Problems of the GCE Advanced level grading scheme*. Manchester: Joint Matriculation Board.

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 2011.

Office of Qualifications and Examinations Regulation
Spring Place
Herald Avenue
Coventry Business Park
Coventry
CV5 6UB