assessment europe

Conceptualising and interpreting reliability

Sandra Johnson
Rod Johnson

Ofqual/10/4706
December 2009

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

Preface

This report is the principal outcome of a conceptual analysis project focusing on assessment reliability commissioned by the Office of the Qualifications and Examinations Regulator (Ofqual) in January 2009. As required, the analysis is contextualised with reference to the kinds of tests, examinations and qualifications that are common in the UK.

The specific requirements for the project were to:

- *identify different approaches to conceptualising 'truth' and 'error' within reliability theory – alternative models of reliability*
- *identify different approaches to estimating reliability, highlighting:*
  - *the assumptions that they make*
  - *their basis in models of reliability*
  - *their strengths and weaknesses*
- *consider how best to evaluate estimates of reliability, based on different approaches, given complications such as the following:*
  - *that not all sources of random error are likely to be accounted for*
  - *that other (systematic) sources of error will be unaccounted for*
  - *that results may be used for a variety of different purposes*

The following definition of 'reliability' was given:

> *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. Unreliability can be attributed to 'random', unsystematic causes of error in 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.*

A variety of information sources were consulted during production of this report. These include original theoretical expositions, seminal texts on reliability, academic journals, Ofqual itself, awarding bodies, and individual academics and practitioners. In the case of academic journals, our literature review has been selective rather than comprehensive, and we have appealed to a number of online library access tools to aid with this. Whenever possible we have accessed the original source material.

Acknowledgements

A number of individuals helped us with the production of this report. Firstly, we thank Qingping He and members of Ofqual's reliability programme Technical Advisory Group for extremely useful feedback on the first draft. We also gladly acknowledge the very supportive input that we received from Anton Béguin, Robert Brennan, Jean Cardinet, Dany Laveault, George Marcoulides and Richard Shavelson, who assisted us in identifying relevant publications and who were also willing to share their thoughts about reliability estimation with us. As far as the awarding bodies in the UK are concerned, we are very grateful to the following individuals and their organisations for responding so helpfully to our email questionnaire enquiry into reliability practice in the field: Michelle Meadows and Chris Wheadon of the Assessment and Qualifications Alliance (AQA), Kathryn Benzine of the Institute of Hospitality (IoH), Ray Burberry of the Waste Management Industry Training and Advisory Board (WAMITAB), Lesley Cook of the English Speaking Board (Intl) Limited (ESB), Peter Evans of the Association of Sports Qualifications (ASQ), Paul Johnson of the National Open College Network (NOCN), Carol Hulm of the British Computer Society (BCS), Allan Murray of the Glass Qualifications Authority (GQA) and Lorne Stather of the Gemmological Association (Gem-A). Finally, we offer our thanks to Ofqual for financially supporting this particular research endeavour, and to Joanna Taylor for her helpful and efficient project management.

Contents

1 Introduction

1.1 Tests, examinations and qualifications in the UK

There are currently over 120 awarding bodies operating in the UK's market-driven qualifications system, between them offering over 6000 nationally accredited academic and vocational/occupational qualifications (for details see the *National Database of Accredited Qualifications*: www.accreditedqualifications.org.uk). The system is regulated in England by the Office of the Qualifications and Examinations Regulator (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).

In the academic arena, the principal qualifications for upper secondary school students are the *General Certificate of Secondary Education* (GCSE), normally taken at the end of Year 11 (16 year olds) and the *General Certificate of Education Advanced Level* (A level), typically taken as a school leaving examination at the end of Year 13 (18 year olds). These qualifications are currently offered by three awarding bodies in England – the Assessment and Qualifications Alliance (AQA), Edexcel, and Oxford, Cambridge and RSA Examinations (OCR). There is a single awarding body in each of Wales and Northern Ireland, respectively the Welsh Joint Education Committee (WJEC) and the CCEA. In Scotland, the national qualifications system comprises *Standard Grade*, *Highers* and *Advanced Highers,* offered by the SQA. In all 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 typically modular, with individual units often assessed in different ways within a single examination. Tests might comprise multiple choice questions, delivered on paper or online, structured response questions or essays, or could take the form of a practical test or an oral interaction with an assessor. Course work might also be included in the final performance profile that eventually leads to a grade.

In the vocational arena, *General Vocational Qualifications* (GVQ) and their Scottish equivalents (SVQ), introduced just over 20 years ago, are available in a number of different employment areas, addressing a growing variety of occupations. These include, for example, hotel management, child care, engineering, music direction, equine transport, glass manufacture, shift management, law and legal work, dementia care, gemmology, to name but a handful. The salient feature of assessment here is that it is workplace based. Assessors observe candidates as they perform relevant tasks – tasks that would be, or might be, required to be performed in the course of carrying out the occupation concerned. The assessors' observations and judgements contribute to a portfolio of performance evidence, that is then subject to internal and external verification, by verifiers (moderators) whose standards of judgement are assumed to be equivalent after participation in standardisation exercises of one type or another.

The recently introduced Diploma merges elements of academic and vocational achievements in a single qualification, comprising three components: principal learning, generic learning and additional and/or specialist learning (ASL). Principal learning is specific to the 'Line of Learning' that a learner chooses, for example

creative and media, hair and beauty studies, science, construction and the built environment, hospitality, engineering. Generic learning is common across all Diploma Lines of Learning. It includes a core of skills that employers and higher education demand, and requires learners to develop their organisational, research and presentation skills through completion of a project – for example a performance or written report – in addition to undergoing a minimum of 10 days' work experience. Additional and/or specialist learning gives learners the opportunity to deepen or broaden their learning. The component outcomes – often stand-alone qualifications – are aggregated to form the full Diploma.

Underneath this vast and still growing qualifications system, the National Curriculum assessment programme continues in England. This was introduced into primary and early secondary education in the late 1980s, and is an annual census of pupil attainment at key stages in the school system. Managed by the Qualifications and Curriculum Development Authority (QCDA), the programme essentially 'certificates' every pupil as that pupil progresses through the key stages. The results of the certification serve multiple purposes, from reporting learning progress to parents and teachers to monitoring the effectiveness of schools, authorities and the national education system as a whole. The programme has undergone a number of evolutions since its introduction, and currently focuses on the assessment of English and mathematics at the end of Year 2 (Key Stage 1), and English, mathematics and science at the end of Year 6 (Key Stage 2). At Key Stage 1, and for science attainment at Key Stage 2, pupil attainment is currently assessed through teacher judgement, supported by in-class use of National Curriculum tasks. At Key Stage 2 mathematics attainment is assessed each year using two national pen and paper tests (both taken by each pupil), while English is assessed using a single reading test and two writing assignments, one short and one extended.

This report explores the issue of assessment reliability against this complex reality. The project requirement was to consider different conceptualisations of reliability, interpreting and evaluating these with particular reference to the kinds of tests, examinations and qualifications common in the UK. Questions of interest are:

- How has assessment reliability been conceptualised?
- What are the technical assumptions associated with one or other conceptualisation?
- Are there particular conceptualisations that serve particular assessment purposes more adequately than others?
- What are the consequences of adopting an inappropriate conceptualisation for investigating and reporting reliability?

Before addressing these questions, let us first remind ourselves why assessment reliability, and indeed assessment validity, is such an issue.

1.2 The nature of educational assessment

Educational assessment is essentially to do with gathering evidence about what individuals know, understand, think and can do, with some particular purpose in mind, typically placement and certification. We can assess knowledge, skills and attitudes informally, through interaction, observation and questioning over some more or less lengthy period of time, or formally, through use of interviews, tasks, tests,

questionnaires and product evaluations. And individuals can assess themselves, or be assessed by others, including peers, teachers and external agencies. But what evidence do we look for when we engage in educational assessment? Where do we look for it? How exactly should we gather it, and how will we know when we have enough? The answers to these questions depend very much on the nature of what is being assessed, what the context for assessment is, who is doing the assessment, and why it is being carried out. Why is this apparently simple process so fraught with difficulty?

The defensible assessment of intellectual skills and abilities has always been, and will continue to be, a challenging endeavour, no matter the form of assessment. This is because the skills and abilities that we are trying to measure are often difficult to define in any absolute sense, and cannot be directly observed. It is these properties that distinguish them from more readily measurable physical properties like height and weight. Because they are not directly observable we are constrained to employ a number of different strategies in efforts to elicit observable evidence of their existence. We pose questions to individuals – questions whose answers provide us with some relevant information about their subject knowledge and ability. Or we give instructions – instructions that require particular observable behaviours to be deployed, behaviours that tell us something about conceptual understanding or skills development.

The instruments of assessment might be informal task-based exercises conducted during normal class time. Examples would be producing a piece of fictional writing in English, sketching a portrait in art, setting up some laboratory equipment in physics, or crafting a soup ladle in woodwork. Alternatively, they might be conventional tests, timed or untimed, comprising a set of obligatory 'atomistic' test items (usually in objective format), a small number of structured or essay questions with choice options, or an oral interaction with peers or an assessor. The test might focus on a particular curriculum topic in depth, or thinly sample the curriculum as a whole. Alternatively, it might involve a lengthy practical demonstration of knowledge and skill. The test might be a stand-alone device, such as a standardised reading test used by teachers in their own classrooms, or a 'significant' task-based assessment carried out in the workplace under assessor observation. Or it could simply be one component in a multi-component external examination, an objective test perhaps. As noted earlier, other components might include an essay paper, a structured question paper, a practical demonstration or an oral test, the variety of component depending on the nature of the subject concerned.

The products of the questions and tasks – the answers and the behaviours – are marked or rated, and, where relevant, summarized in some way, typically as total or average test scores. The outcome might be an immediate decision for the individual concerned: an end-of-term 'reading age' based on the result of a standardised reading test, a classification into a 'performance percentile' for the class or entire school, based on the results of a French oral test, an attainment level decision based on performances in National Curriculum mathematics assessments, or a skills mastery decision based on the individual's performance in the practical workplace task. Alternatively, the test result could simply be recorded for use later, perhaps for contribution to a portfolio, or to add to the results of a series of similar standalone tests held over the year, whose combined results might be summarised to furnish the basis for an evaluative decision about achievement, progress and future placement. If

part of a multi-component examination, the test result might be adjusted in some way, and weighted, before contributing to a global examination mark, to which cut scores would be applied to produce grade classifications (see Robinson, 2007, for an overview of the complexity of results processing in the academic examinations system in England).

In principle, testing might appear to be a quite straightforward exercise. But numerous extraneous influences impact on the process, introducing variability that ultimately contributes to inconsistency and 'error' in the measurement results. The apparently straightforward task of assessing subject knowledge is already a difficult exercise. We can ask a student to tell us the date of some famous battle in English history. We can ask the name of the king of England at the time. We can ask how many soldiers died in that battle. And so on. Students might answer all three questions correctly or all three incorrectly. They might answer one or two correctly and the third not. We might ask another 20 similar questions, perhaps on the same general theme. But what would the outcome of the questioning then tell us about the individual's historical knowledge? If we added some questions that required the student to reason about events, perhaps to explain why this or that strategy was adopted by the commanding officer, we might change the picture again. If we had asked the same questions the day before, or the day after, or the following week, would the outcome have been the same in general, and for every individual student? What difference would it have made had the students' knowledge been explored through an examination comprising essay questions? Would students have been able to show more evidence of their relevant knowledge this way? Either way, what influence would individual markers have had on the assessment results?

Skills assessment can be more straightforward, or equally challenging, depending on the nature of the skills being assessed. For example, if we want to know whether a candidate in a chemistry examination can weigh a gram of copper sulphate crystals to the nearest milligram, we can simply ask the candidate to do that and judge accordingly. But what if this is just one small task within a longer laboratory experiment? We might devise a checklist, and have examiners note which of the various steps are completed adequately according to some given set of criteria. If not all the tasks are satisfactorily completed by the candidate how do we use the overall profiles of successes and failures to come to a decision about laboratory skills for this candidate on this particular experiment? And how far could we generalise the result to the broader domain of 'chemistry laboratory skills' at the level concerned? How confident could we be that the generalisation is defensible? What relative importance might we give lab skills as opposed to theoretical chemistry knowledge within a global chemistry examination? How might we judge how well we had measured each aspect? And what can we say about the meaning and value of the combined result? This, of course, is where we need to consider issues of validity and reliability.

1.3 Reliability and validity

In the context of academic testing and examining, as noted by Wood (1991, Chapter 12), assessment *validity* is essentially to do with how appropriately we operationalise our definitions of subject knowledge, ability and skill in practice, in the form of assessment (measurement) tools. The principal concerns here are content validity (how well the content of a test or examination relates to the curriculum being assessed

in terms both of relevance and coverage) and construct validity (the degree to which a test or examination elicits evidence of the particular ability or skill that is in principle being assessed). In vocational assessment predictive validity is clearly also important. This is because vocational qualifications serve not only as attestations of an individual's current levels of work-relevant knowledge and skills but also as direct indicators of future occupational competence. [See the seminal text by Messick, 1989, for further discussion of validity.]

Reliability has to do with how well we are able to measure what we set out to measure using the given tools, and how well we might measure the same thing if our tools or procedures were to be changed in some way. It is important to remember here that reliability is itself a contributor to validity. Together the two aspects contribute to the 'dependability', or quality, of measurements.

Any health professional measuring an individual's height would use a wall-mounted stadiometer for this purpose, not a flexible tape measure. The height of the individual, easily defined, is what is being measured. The choice of an appropriate tool for the purpose, along with correct use of this tool, should guarantee a high degree of validity. What about measurement reliability? The height will be measured well, but not necessarily perfectly, because human beings are not inanimate objects. How carefully the measurer sets out to measure the height will partly depend on the purpose for which the height measurement is required. For some purposes, general medical monitoring perhaps, a height measurement to the nearest centimetre could be more than sufficient. For other purposes this might not be an adequate degree of precision. For example, should the individual concerned be a participant in some pharmaceutical trial, perhaps being monitored for the effects of growth hormone administration, then the measurement might need to be more precise than this, perhaps accurately estimated to within a millimetre. In this case, special care would need to be taken that the individual being measured should stand with the same posture each time, feet flat on the floor, straight back, muscles relaxed, since changes in these variables could well lead to different height readings. Recognising this, several different readings might be taken, and averaged for greater security. So, even for something as apparently straightforward as height measurement there are factors that affect the measurement adversely, which if uncontrolled could lead to a non-valid measurement, that is a height measurement that is too imprecise for the purpose intended.

So it is with educational assessment and test results. Except that the challenges of quality measurement are greater in this context. This is partly because, as noted earlier, even when we have a workable definition the knowledge, ability or skill being assessed cannot always be directly measured, as height can. To find out what individuals can do in mathematics we have to ask them questions, and give them problems to solve. If the questions contained in the test are clearly mathematical, and if the mathematics in the test is appropriate to the age of the individuals being assessed (that is to the curriculum being taught), then we would consider the test to be 'valid', i.e. appropriate. Unless, that is, there is a lot of reading involved, as in word problems. For in this case the ability to read could interfere with demonstration of mathematics ability, especially for weak readers, and 'cloud' the assessment of mathematics. We then risk producing non-valid measurements of mathematics ability

for poor readers. The test would be said to be 'biased' in favour of good readers (see Ackerman, 1991, for further examples), and its validity therefore compromised.

Now what about reliability, the 'how well' of measurement? Can we assume that the test would give the same result for the same individual whenever, wherever and however it might be used? Common sense would suggest not. It is a fact of life that pupils and examination candidates rarely produce consistent performances in a test. The most able might do so, as might the least able, but most do not. Individuals might answer one question correctly and the next wrongly, and so on, providing a fluctuating profile of success from beginning to end. What one person finds easy another might find hard, and vice versa. This interaction between test takers and questions or tasks, this inconsistency in performance, is a potentially important contributor to unreliability in assessment.

We all know that children, like adults, make mistakes in calculation, even when they know how to solve a particular problem. And the younger and the less able the pupils the more likely they are to forget learned facts from one day to another. Some pupils might have covered a particular topic in school the day before, while others might have left it behind weeks earlier. If one class has covered fractions more recently than another, for example, then we might expect the pupils in that class to perform better than those in the other on fractions items. Even if we accept that once taught and mastered such skills are there for life, recent practice could provide an advantage in a testing context. Then again, while all classes might have covered the same material in the same period, perhaps one teacher has been particularly effective in one particular area, and this could show up in better item scores in that area for that teacher's pupils. These are just two possible examples of school/class effects that might lead to inconsistency in pupils' performances on different test questions.

More personal factors are also relevant in this sense. These include the particular subject interests of individuals that affect both their learning and their assessment motivation. Some test takers more than others become flustered with anxiety during formal testing sessions. Others simply refuse to make the effort to show what they can do. Extraneous factors can also have an unpredictable influence, like distractions outside the classroom window, a heat wave on the day of testing in June, or other examinees leaving the examination room early. Such factors potentially affect young adults undergoing academic or vocational assessment, in the school or in the workplace, as much as pupils in primary classrooms. They are all potential contributors to inconsistency in measurement, many of which are beyond our control.

These are just some of the factors that can affect an individual's performance on a single test. If we consider any test as simply a 'container' for test questions, we can imagine that if we replaced some of the questions in that container with others we might see different outcomes for the same individuals. Yet, compared with the efforts made by awarding bodies to minimise the potential influence of marker or assessor differences on test outcomes (for a recent comprehensive review see, for example, Meadows and Billington, 2005), this issue of question effects seems to have received rather little attention.

But what exactly is 'reliability'? How is it defined? How can it be measured? How do we know when we have enough of it? How can it be increased? In Chapter 2 we offer

a brief overview of the evolution in reliability conceptualisation, since first introduction of the concept over a hundred years ago to the present day, before moving on in later sections to consider how reliability estimation might most appropriately be approached in the many different kinds of testing context that operate in the 21st century in England.

2 The evolution in reliability conceptualisation

The material in this chapter is likely to be familiar to any reader with more than a passing acquaintance with measurement theory. Even so we feel that a brief overview of the evolution of the concepts of measurement, and particularly of reliability, over the course of the 20th century will be helpful in locating where we are now and, to some extent, why we are where we are. The following exposition will require that we resort from time to time to mathematical notation, which we nonetheless try to limit to a minimum.

2.1 Observed scores, true scores and measurement error

When we use a single test with a group of individuals the result is a set of scores, one score for every individual tested on each question in the test. There will be variation in these scores, some individuals producing high scores for most questions, others low scores for most questions, with typically many in-between, showing a mixed picture. These are scores that we can see. For this reason they are called 'observed scores', and the variation in them is 'observed score variation'. Genuine differences among the individuals in terms of what is being assessed will normally explain most of the variation in scores. This 'genuine', or 'valid', score variation is technically called 'true score' variation. But, as noted in Chapter 1, there are always other influences at play in testing situations that also contribute to score variation, such as the conditions of testing (temperature in the room, amount of noise disturbance, the number of students finishing early and distracting others), the nature of the test itself (for example, different students preferring some topics more than others and doing relatively better on related test questions, some doing better than others on multiple choice questions, and so on), and marker differences and inconsistency. These factors account for some of the observed variation in scores, and this part of the variation is unwanted – it is 'noise' in the assessment process. Technically, we say that factors such as these, and the score variation they create, contribute to 'measurement error' in assessment.

In fashioning tests, we try – or we should try – to reduce measurement error as much as possible, so that what remains is as close as possible to the 'true score'. To help us to quantify how much of the variation in a test score is due to differences in the true scores of the candidates and how much is due to error, or noise, measurement theorists have constructed a number of 'reliability coefficients', which are computed in different ways, but all of which have analogous interpretation. They all range from 0 to 1 – theoretically, at least (some of them can under certain circumstances be negative). If, then, the value of some reliability coefficient is, say, 0.7, the implication would be that 70% of the variation in candidates' scores is due to real differences in their 'true scores', and the remaining 30% is attributable to errors of measurement, or 'noise'. The theory is that the higher the value of the coefficient the less error, or uncertainty, there is in the test results. In testing situations we typically expect to have coefficients above 0.8 (80% of the variance being attributable to valid variance) and preferably above 0.9 (90% valid variance).

2.2 True Score Theory

The study of measurement error began in the United Kingdom about a century ago. Its origins can readily be traced back to a series of papers written by Spearman (1904a, 1904b, 1907, 1910, 1913) and Brown (1910, 1911) at the opening of the 20th century. The body of principles and ideas constructed upon the early work of Spearman and Brown has come to be known, sometimes disparagingly, as "Classical Test Theory", often abbreviated as CTT. We prefer the more descriptive label "True Score Theory", which we shall use henceforth. For a properly rigorous, formal treatment of the fundamentals of True Score Theory, we refer the interested reader to Chapters 2 and 3 of Lord and Novick (1968).

True Score Theory starts off from the fundamental premise that the observed score of an individual on a test, i.e. X, is equal to the sum of a true score, T, and a measurement error, E. Conventionally, in symbols, we write

[2.1] $X = T + E$

Of these three quantities, we only have access to one, the observed score, while the other two are not directly measurable. Yet we need to have some way of quantifying the size of measurement error in order to know to what extent we can rely on the result of the test.

Strictly, we should rewrite [2.1] as something like

[2.1a] $X_{ip} = T_p + E_{ip}$,

where the subscripts i and p stand for, respectively, test i and person p. Thus adorned, expression [2.1a] says that the observed score of person p on test i is made up of the true score attributable to person p plus some measurement error specific to that person's performance on that test. Where there is no likelihood of misunderstanding we shall use the simpler, unsubscripted forms X, T and E in place of the subscripted X_{ip}, T_p and E_{ip}.

Note that X, T and E should be considered as statistical quantities. They stand not for specific values obtained from a specific individual on a specific test on some specific occasion, but for some value that you could get from a sample individual on a sample test. Technically they are called 'random variables'.

An important property of a random variable is that if we sample from it enough times, observing its value each time, the sample values will eventually converge on a fixed quantity, called the 'expected value', or 'expectation', of the variable. The expectation of a random variable X is typically notated $E(X)$. The first major assumption of True Score Theory states, reasonably, that the expected value of the observed score, X, is the same as the expected value of the true score T, often written as the Greek letter *tau* (τ). In symbols

[2.2] $E(X) = E(T) = \tau$

From which it follows that

[2.3] $E(E) = 0$

the expected value of the error of measurement is zero: in other words, in the long run measurement errors will "cancel themselves out".

Another property of a random variable is that, in general, its values will vary from one observation to the next. This variation is typically summarised in the symbol σ^2, called the variable's 'variance'. When we refer to the variance of X, we may write σ^2_X or $\sigma^2(X)$, or even, occasionally Var(X). The parenthesised forms are useful when the variable has its own subscripts, $\sigma^2(X_{ip})$, for example.

Two random variables can vary together, or *covary*. Take height and weight, for instance. Taller people tend as a rule to be heavier, but the relationship is not deterministic. Similarly, we would expect pupils' test results in reading to vary more often than not in the same direction as their results in numeracy. In both cases there is a positive 'covariance' between the two variables. We symbolise the covariance between two random variables X and Y as σ_{XY}, $\sigma(XY)$, or $Cov(X,Y)$. Note the similarity with the notation for the variance, σ^2. Indeed, the covariance of a variable with itself is defined as being the same as its variance:

[2.4] $\sigma_{XX} = \sigma^2_X$

The variance of two random variables added together is the sum of their variances plus twice their covariance:

[2.5] $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$

Clearly, if the covariance of X and Y is zero (i.e. they do not covary) the variance of their sum is equal to just the sum of their variances. We shall need to use this important result later.

Although the covariance is very useful in giving us information about the relationship between two variables, its descriptive value is lessened by the fact that it is not easy to relate to the scale of either variable. For example, the covariance of weight (expressed in grams, say) and height (expressed, say, in centimetres) is expressed neither in grams nor in centimetres (in fact it is a function of the product of the two). Consequently another quantity derived from the covariance is more frequently used to describe the relationship between two variables. This is the 'correlation', usually notated by the Greek letter *rho,* i.e. ρ, which is essentially a covariance adjusted to be on a scale from -1 to +1. A correlation of -1 between X and Y means that X and Y vary in a perfect inverse relation to each other (when X goes up Y always goes down proportionally, and vice versa). Similarly, a correlation of +1 means that X and Y always vary together in the same direction. And a correlation of zero means that there is no linear relationship at all between X and Y. Values in between indicate a greater or lesser linear relationship between X and Y.

The formal definition of the correlation between two random variables X and Y is

$$[2.6] \quad \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

We can now state the second major assumption of True Score Theory, namely that the covariance (hence also the correlation) of true score and measurement error is zero:

$$[2.7] \quad \sigma_{TE} = \rho_{TE} = 0$$

Assumption [2.7], together with equation [2.5], permits us to formulate the important result

$$[2.8] \quad \sigma^2_X = \sigma^2_T + \sigma^2_E$$

that the observed score variance is equal to just the sum of the true score variance plus the error variance. This result is fundamental for much of True Score Theory, as well as its successor generalizability theory (G-theory), which we introduce later in this chapter and cover in some detail in Chapters 3 and 4.

In the light of [2.8], the quest for measurement reliability becomes essentially the attempt to reduce as much as possible the error variance, σ^2_E, as a proportion of the total variance, σ^2_X.

Now it can also be shown (for example by Lord & Novick, 1968, p.57) that

$$[2.9] \quad \sigma_{XT} = \sigma^2_T,$$

the covariance of the observed score and the true score is the same as the true score variance.

But we also know, by rearranging [2.6], that $\sigma_{XT} = \rho_{XT}\, \sigma_X\, \sigma_T$. Substituting for σ_{XT} in [2.9], and squaring both sides, we have

$$[2.10] \quad \rho^2_{XT} = \sigma^2_T / \sigma^2_X = 1 - \sigma^2_E / \sigma^2_X$$

Thus, formally, the squared correlation of the true score with the observed score, which is *defined* as reliability, is also equal to the true score variance divided by the observed score variance, or, equivalently, the proportion of observed score variation which is not attributable to measurement error.

Note that the value of any correlation coefficient is restricted to the range from -1 to +1. Hence the reliability coefficient, ρ^2_{XT}, which is the square of a correlation, must necessarily be limited to the range from 0 to 1. By definition, then, for any reliability coefficient, ρ^2, derived from [2.10] we can assert

$$[2.11] \quad 0 \leq \rho^2 \leq 1$$

The great value of [2.10] is that it permits us to switch between a view of reliability based on correlation, as predominated in early 20th century studies of reliability, and one based on variance ratios, which has now become the prevailing position.

2.3 Equivalent tests

At this point we have a conceptual basis for dealing with reliability, but only on an abstract level. Throughout the preceding discussion, the only observable quantity is the observed score, and the theory as thus far expressed gives us no means of unravelling the true score from the observed score.

Suppose, however, that we have available two measurements, $X = T + E$ and $X' = T' + E'$, such that

[2.12] $E(X) = E(T) = E(X') = E(T')$,
$\sigma^2_X = \sigma^2_{X'}$,
$\sigma_{EE'} = \rho_{EE'} = 0$

that is, the two measurements have the same expectation and the same variance, and their measurement errors are not correlated. Tests of this kind are what came to be called *parallel tests* (*cf* Lord & Novick, 1968, p.47).

We do not give the working here, but it is easy to show that, given [2.12],

[2.13] $\rho_{XX'} = \sigma^2_T / \sigma^2_X = \rho^2_{XT}$

Thus, by introducing a notion of parallel tests, it is possible to find a way of arriving at an expression for the reliability of a measure, $\rho_{XX'}$, the correlation between two parallel measures, which does not involve the unobservable true score T.

Spearman and Brown had already arrived, apparently independently, in 1910, at the idea of calculating reliability by correlating measurements taken from two halves of the same test, so-called *split halves*, making the more or less implicit assumption that the two components would effectively be parallel.

However, a reliability coefficient based on a pair of split halves only describes a test which is half the length of the original. A lasting contribution of both Spearman (1910) and Brown (1910) was the derivation of expression [2.14] which allows us to compute the reliability of the full length test, $\rho_{XX'}$, on the basis of the correlation ρ_{12} of the two component parts.

[2.14] $\rho_{XX'} = 2\rho_{12} / (1 + \rho_{12})$.

In effect [2.14] is just a special case for $k = 2$ of the ubiquitous Spearman-Brown *Prophecy Formula*

[2.15] $\rho_{KK'} = k\rho_{12} / (1 + (k - 1)\rho_{12})$

where ρ_{12} is the correlation of two parallel tests of equal length, k is a multiplier (which need not be an integer), and $\rho_{KK'}$ is the predicted reliability of a test k times the length of the original tests.

Some relaxation of the constraints expressed in [2.12] led subsequently to the introduction of the notion of *equivalent tests*, of which parallel tests are a special case. Debates continued through much of the century on the most appropriate way of defining equivalence in tests. Some of the better known attempts to come up with a suitable notion of test equivalence, apart from the 'split halves' of Spearman and Brown, include: comparable tests (Kelley,1923); repeated tests (test-retest); τ-equivalent tests (Lord & Novick, 1968, Chapter 10); parallel test forms (Stanley, 1971, pp 404-406); and many more. We do not enter here into discussion of the rationale of these many different ways of inducing replication of a measurement, which is well summarised, for example, in Brennan (2001a). The question of reliability indices corresponding to these many different types of equivalent test is revisited in Chapter 3.

2.4 Coefficient alpha

The insight supplied by [2.10] and [2.13], that reliability can be expressed as a ratio of variances, was available from the earliest times in the development of True Score Theory. But almost all of the reasoning about reliability coefficients was cast initially in terms of correlation rather than ratios of variances, using favoured methods of the time based on test-retest or split-half. This is not in fact so surprising, as the correlation coefficient had only recently been proposed, in the late 1880s, by Galton, who was very influential in Spearman's thinking (*cf* Stigler, 1989), and derived shortly after by Pearson (1896).

Even though Fisher (1925) had published his seminal text on the analysis of variance 12 years earlier, it was not until 1937 that Kuder and Richardson proposed a number of new reliability coefficients based on ratios of variance estimates drawn from a single test. The most famous of these coefficients, and for a time the most frequently applied, is the one reported as formula 20 in Kuder and Richardson (1937), now universally known as KR-20.

Apart from its explicit expression as a variance ratio rather than a correlation coefficient, the main innovation of KR-20 is that it uses the sum of individual item variances in a test to estimate the error variance for the test as a whole. We do not here offer further detail, since KR-20 turns out to be a special case, applicable only to tests containing exclusively dichotomously scored items, of Coefficient α, which we consider below. Virtually all introductory texts on measurement theory can be expected to offer the interested reader some treatment of KR-20 (for example Nunnally, 1967, pp 196-197; Mehrens & Lehmann, 1984, p. 276; Bachman, 1990, p. 176).

Kuder and Richardson's paper does not appear to have been a powerful catalyst for a new wave of variance-oriented approaches to reliability, though there are a few exceptions (Baker, 1939; Jackson, 1939; Hoyt, 1941; Burt, 1947). Indeed it is not clear that Kuder and Richardson were influenced in any way by the growing interest

in the analysis of variance, which did not really catch on among transatlantic measurement practitioners until after World War II.

Then in 1951, Cronbach published the landmark paper *Coefficient Alpha and the internal structure of tests* (Cronbach, 1951). Coefficient alpha, sometimes spelled out, sometimes written as the Greek letter α, is undoubtedly the most used (and abused?) of all the reliability coefficients, and seems to be almost *de rigueur* in almost any reported test application.

Hogan, Benjamin & Brezinski (2000), for example, looked at a sample of 696 psychometric tests listed in the Directory of Unpublished Mental Measures (Goldman, Mitchel and Egelson, 1977), a frequently cited information resource for measurement professionals in the United States; of the tests sampled, 533, almost exactly two thirds, reported a value of alpha, as opposed to other measures, as their index of reliability.

Cortina (1993) also reports that

> A review of the Social Sciences Citations Index for the literature from 1966 to 1990 revealed that Cronbach's (1951) article had been cited approximately 60 times per year and in a total of 278 different journals. (Cortina, 1993, p.98)

Although Coefficient alpha is virtually synonymous with the name of Cronbach (indeed it is more often than not referred to as "Cronbach's alpha") Cronbach himself has readily conceded (Cronbach and Shavelson, 2004) that he was by no means the first to propose the formula. Equivalent formulations had been previously proposed by, at least, Hoyt (1941) and Guttman (1945).

The formula for α is

$$[2.16] \quad \alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right)$$

where k is the number of items in the test, σ^2_i is the variance of the ith item in the test, σ_X is the total variance over test scores (a person's test score being the sum of all the item scores for that person) and the summation is over all items. We recall that if all items in the test are dichotomously scored [2.16] reduces to the KR-20 formula which we mentioned above.

Cronbach (1951) makes the claim, repeated in Cronbach and Shavelson (2004), that alpha is the mean of all possible split-half reliabilities for a given test application. In fact the claim is true provided that the notional component half-tests conform to the Spearman-Brown requirements for parallel tests summarised in [2.12], in particular that the item variances are equal (Cortina, 1993). Otherwise alpha will be less than the mean of all split-half reliabilities. In any case, we can expect intuitively that alpha, as an average, is likely to be more stable than its component parts taken separately.

We can further recognise alpha as being a form of reliability index by comparing the formula of [2.16] with the second form of the equation for ρ^2 [2.10], restated below as [2.17]

[2.17] $\rho^2_{XT} = 1 - \sigma^2_E / \sigma^2_X$ (from [2.10])

Comparing [2.16] with [2.17], disregarding for the moment the constant term *k/(k-1)*, which approaches 1 as the number of items, *k*, increases in size, we can readily see the similarity between α and ρ^2, with the sum of individual item variances, $\Sigma\sigma^2_i$, taking the place of the error variance σ^2_E.

Coefficient alpha is sufficiently important to the practice of reporting reliability to merit looking a little more closely at [2.16], to seek further insight into the way alpha works. Technically averse readers may find the following discussion a little heavy, and might wish to skip lightly over the material in the next few paragraphs.

First of all, we recall that [2.16] involves both the test variance, σ^2_X, and *k* variances, σ^2_i, of individual item scores X_i. But the test score *X* is just the sum of the (possibly weighted) item scores

[2.18] $X = \sum_{i=1}^{k} w_i X_i$

We assume here for simplicity that all items in the test have unit weight in the calculation of the total score (all the w_i are equal to 1), so that [2.18] simplifies to

[2.19] $X = \sum_{i=1}^{k} X_i$

We already know from [2.5] that the variance of the sum of two random variables is equal to the sum of their variances plus twice their covariance. The generalisation of [2.5] from 2 to any number of variables is

[2.20] $Var(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} Var(X_i) + \sum\sum_{i \neq j} Cov(X_i, X_j)$

Equation [2.20] might look daunting, but it is in fact simpler than it first appears. It just says that the variance of the sum of a set of random variables is equal to the sum of their variances plus the sum of all their covariances. There are in all, for a set of *k* variables, *k(k-1)* covariances, of which, in general, *k(k-1)/2* are different (because *Cov(X,Y) = Cov(Y,X)*, always).

Now, given [2.20] and using the more succinct notation σ^2_i for $Var(X_i)$ and σ_{ij} for $Cov(X_i,X_j)$, summations as before being understood to be over *k* terms, we can rewrite [2.16] as

[2.21] $$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sum \sigma_i^2 + \sum\sum_{i \neq j} \sigma_{ij}}\right)$$

$$= \left(\frac{k}{k-1}\right)\left(\frac{\sum\sum_{i \neq j} \sigma_{ij}}{\sum \sigma_i^2 + \sum\sum_{i \neq j} \sigma_{ij}}\right)$$

If we now rewrite [2.21], putting $K = k/(k-1)$, $V = \Sigma\sigma^2_i$ and $C = \Sigma\Sigma\sigma_{ij}$, we can get a clearer view of the basic structure of α

$$[2.22] \quad \alpha = K\left(1 - \frac{V}{V+C}\right) = K\left(\frac{C}{V+C}\right)$$

Looking at [2.22] we can easily see that the most influential contributor to the value of an alpha coefficient is the sum of item covariances, C. Thus, if the items in the test are highly correlated, their covariances will tend to be higher relative to their variances, and will push up the value of alpha correspondingly. On the other hand, if items are not correlated with each other, their covariances will tend towards zero and α will be close to zero.

An unfortunate property of α is that, if C is negative, as it will be if negative covariances between items outweigh positive ones, α will itself be negative, as Cronbach & Hartmann (1954) pointed out quite early on. Sometimes this can occur just because of flawed mark schemes which invert the intended scoring of correct and incorrect answers for some items. At other times, negativity is an indicator of some serious breach of the assumptions of [2.12] which are supposed to hold between the items of the test. It is clear that negative-valued α cannot be interpreted as a reliability coefficient, which is defined to be necessarily greater than 0 by [2.10] and [2.11].

Given the considerable influence of interitem covariances on the behaviour of alpha, it is not surprising that it has frequently been interpreted, and used, as a measure of constructs like internal consistency, homogeneity, unidimensionality, rather than – or as well as – reliability. Such interpretations of alpha, along with various lower-bound claims, are comprehensively treated – it is fair to say with some degree of scepticism – by Cortina (1993), Schmitt (1996), Huysamen (2006) and Sijtsma (2009), far more competently than we can pretend here.

2.5 Beyond alpha

By the early 1950s, almost all the important results of True Score Theory as originally formulated by Spearman and Brown had been produced. Gulliksen (1950) had published a complete account of the state of the art as it then was, and Cronbach's article which definitively established alpha as the primary reliability index had appeared in 1951. But the technical development of the underlying theory had still not really moved on since its inception half a century before.

Observant readers will have noticed that the preceding discussion is framed entirely in terms of population parameters like τ, σ and ρ. This is not accidental. For 50 years leading practitioners of measurement theory had consistently confused sample and population quantities, observed and abstract quantities, estimates and estimators, in both conception and notation. Cronbach was himself, at the time, equally guilty of such confusion. He writes

> In thinking about reliability, one can distinguish between the coefficient generated from a single set of n persons and k items, or about the value that would be obtained using an exceedingly large sample and averaging coefficients over many random drawings of

> items. … In the history of psychometric theory, there was virtually no attention to this distinction prior to 1951, save in the writings of British-trained theorists. My 1951 article made no clear distinction between results for the sample and results for the population. (Cronbach & Shavelson, 2004, p.204)

The same confusion persists still in citations of the formula for alpha. Some commentators use the lower case Roman letter *s* in the notation for the variance s^2, suggesting that they have in mind an observed value with no particular sampling properties (e.g. Stanley, 1971). Others (e.g. Nunnally, 1967) use the form σ^2, which leads us to suppose that they are thinking of a population value.

It was, and unhappily still is in many circumstances, commonplace to take an index, of reliability, say, defined with respect to some imprecisely specified population of persons and test items, compute a value for the index by inserting values derived from a single set of observations, and then generalise about the computed value without any regard to the theoretical sampling properties of the index itself or to the relationship between the observations and the possible population(s) from which they might have been supposed to have been taken.

Lord & Novick (1968) eventually supplied the missing underlying mathematical and statistical framework in a landmark publication which effectively marked the high point and the beginning of the end for the development of True Score Theory. For once Lord and Novick had laid out what is generally agreed to be the definitive formalisation of the work begun by Spearman and Brown, there was really very little to add.

What we mean by this last assertion is twofold: firstly, (virtually) all reliability measures cited routinely in the applied measurement literature (as documented, for example, by Hogan, Benjamin, & Brezinski, 2000) had been invented by 1951; and, secondly, any new contributions to the theory of reliability after 1968 have been either new interpretations of alpha (as an index of homogeneity, unidimensionality, or whatever); critiques of alpha or of its more recent interpretations; proposals for new alpha-like coefficients offering better theoretical lower (or upper) bounds; or new indices looking exactly like alpha except that one of the two dimensions in the notional matrix of observations (persons and items) was replaced by some other variable, like raters (e.g. inter-rater reliability measured by alpha over raters and persons).

From 1968 onwards, it should have been time for new ideas and new approaches to take over. The circumstances were certainly favourable. Computers were available then, and becoming famously twice as powerful every 18 months (Moore, 1965). Massive advances had been made in statistical theory and practice, in the analysis of variance, maximum likelihood estimation and numerical methods. The old days, when extracting factors, inverting matrices, finding derivatives iteratively, were all done by hand, were about to be permanently consigned to history. Under these circumstances, it is astonishing that the old reliability theory introduced just after the turn of the last century should have held on for so long, holding so much appeal that measurement professionals have been reluctant to let it go.

However, a century on, there are finally signs that the old order is at last changing. The most conspicuous development in measurement theory and practice generally over the last half-century has been the spectacular rise of “Item Response Theory” (IRT). IRT differs both conceptually and operationally from True Score Theory, and for many educational and psychological testing applications has become the dominant theoretical and methodological instrument (computerised testing and international attainment surveys, for example). All IRT models require intensive calculation, and their use would not have been feasible without modern computers; the generalised availability of the simpler, mainstream IRT procedures to a wide audience of measurement practitioners is only now possible because of the ready availability of specialised IRT software on personal computers.

It is not possible in these times to address any issues in measurement without some reference to IRT. However, in the special case of the pursuit of a comprehensive treatment of reliability and measurement error, it is our impression that mainstream IRT has very little to say. We therefore devote the next section to a brief discussion of IRT, but only to set out our rationale for not offering in the report any further consideration of the contribution of IRT to current practical reliability issues in the UK.

On the other hand, the most important development which is of relevance to the pursuit of a comprehensive treatment of reliability and measurement error is due to Cronbach and his associates, who made the link between the definition of the reliability coefficient as a variance ratio, the role of variance components in refining the definition of true score variance, and the contribution of the analysis of variance in providing ready-made apparatus for manipulating variance components. The result was “Generalizability Theory” or *G-theory* (Cronbach, Rajaratnam & Gleser, 1963; Cronbach, Gleser, Nanda & Rajaratnam, 1972).

We recall that Spearman's original account of measurement error only allowed for two sources of variation, one in some sense ‘desirable’ (between true scores) and the other ‘undesirable’ (everything else), whereas in effect there can be many influences affecting a subject's performance on a test, some of which can be controlled, some of which can not. Fisher (1925) showed, with the analysis of variance, how certain sources of variation can be manipulated and their effects taken into account when analysing experimental data. Not the least of the benefits which can be derived from the incorporation of the analysis of variance into generalizability theory is the considerable know-how in sampling practice and experimental design which is part and parcel of the intellectual baggage that the analysis of variance carries with it.

Although Fisher’s analysis of variance dates from 1925, early applications of the technique tended to be in areas like agriculture, where experiments were primarily set up to differentiate between a small, fixed number of effects or treatments, like plant variety or type of fertilizer. Applications involving samples of effects drawn from large, perhaps infinite, domains like test subjects, test items or test raters were not the norm. It was not until Eisenhart's (1947) paper that a clear distinction was drawn between the fixed and random effects models. Now the way was clear for serious application of analysis of variance techniques to reliability studies. The possibility for development of complex G-theory designs has also benefited from the availability today of substantial computing power on the desktop.

We return to the contribution of generalizability theory to the study of measurement reliability in the final section of this chapter.

2.6 Item response models

IRT arose, at least in part, out of a desire for direct modelling of (1) items rather than tests and (2) individual abilities underlying test scores rather than the scores alone. The result is a new class of models, quite distinct from those based on the idea of true scores, in which the performance of an individual on a particular item is defined to be a joint function of both the level of ability of the individual and the level of difficulty of the item. Models of this kind are called 'item response models'.

IRT arrived on the scene quite late. One of the earliest exponents was the Danish mathematician Rasch (1960), whose work was to become quite influential, particularly in Europe. Rasch visited the University of Chicago, where he influenced Wright and his students (Wright & Stone, 1979; Wright & Masters, 1982). A parallel line of development in the United States is built on the chapters submitted by Allan Birnbaum to Lord and Novick (1968) and on the work of Lord himself (1980). There are now numerous popularisations of IRT, for example Van der Linden and Hambleton (1997), Embretson and Reise (2000).

IRT models items very rigidly. It makes very strong assumptions about the nature of items and their relationship to individual abilities and to each other. Any items whose behaviour does not fit the model based on these assumptions must be discarded. In order to establish the fit of items to a model they need to be extensively trialled. But once accepted, their properties are considered fixed (in other words attributes like item difficulty are not considered to be variable over different applications of the model), with the typical implication that the sample used for trialling will adequately represent the population of test takers in all subsequent applications of the item.

While this procedure may be typical of many IRT analyses, particularly using Rasch-based models, Anton Béguin has pointed out to us that in more flexible IRT models

> … it is possible to describe behaviour on existing tests instead of constructing tests according to a model. Evaluating fit of an item to a more flexible model can be seen in the same way as quality control procedures using classical indices in Test and Item Analysis (for example the correlation between the item and the score on the remaining part of the tests). (Béguin, 2009, personal communication)

This seems to conflict, however, with the assertion of Lee, Brennan and Kolen, cited in Brennan (2001a, p.304):

> … the IRT procedure assumes strictly parallel forms (or a fixed form), while the other procedures [i.e. classical indices] assume randomly parallel forms. … In effect, error attributable to content sampling is assumed not to exist in the IRT procedure, but is an integral part of the other procedures. (Lee, Brennan & Kolen, 2000, pp.14-16)

There is in effect no error term as such in the standard presentation of an IRT model. Sampling theory has been proposed for IRT, for example by Holland (1990) and Bechger, Béguin, Maris and Verstralen (2003). Invariably, though, the population on

which the sampling model is based is the population of test subjects; the test items are considered fixed. Estimates of model parameters are obtained from one of a number of possible maximum likelihood fitting procedures, depending on the software used. What could be considered as error terms in an IRT analysis are the residuals obtained from constructing the deviations between observed person scores and scores predicted by the maximum likelihood fitting procedure.

IRT is essentially a scaling model rather than a sampling model (Cardinet, Johnson & Pini, 2009): that is, it is designed more to place candidates on a scale than to facilitate the study of the variation from one application of a test to another. As such, reliability has not been a preoccupation of its devotees in the way that it has for true-score-based theories. The position which underlies this report is that reliability makes little practical or intellectual sense unless it is based on a notion of replication, as has been powerfully argued by Brennan (2001a).

Brennan remarks, referring to the concept of reliability found in IRT, that:

> It is certainly true that statistics that have a reliability-like form can be computed based on an IRT analysis, but it is equally true that almost all such analyses treat items as fixed. … IRT has no explicit role for error of measurement relative to investigator-specified replications. (Brennan, 2001a, pp.304-305)

IRT does indeed have a measure of precision, the *information function*, which can be used to compute a quantity which is in some ways analogous to the reliability of true-score-based theories, but has the added property of being sensitive to differences in ability of the test subjects. Determination of the information function is complex, and is dependent on the stability of the model fitting procedure. Embretson and Reise (2000, pp 183-186), for example, provide a concise discussion of the information function.

As has been pointed out by Doran (2005), IRT 'reliability', based on the information function, relates to

> the degree of certainty that a [given] test is an accurate measure of ability for any given value of [the ability parameter] θ. Clearly this does not describe the deviation of an observed score from a true score. … [CTT reliability] is a metric which relates to replication over measurement procedures (Doran, 2005, p.674).

Brennan (2001a), moreover, observes that the IRT concept of information requires fixed, predetermined item and test attributes, with the consequence that the reliability of a test cannot depend on the circumstances of the administration of the test or on variation in any aspect of the test taker other than ability.

In the emerging multidimensional IRT, the ability parameter θ can be a vector and hence reflect more than one 'ability'. But this does not change the fundamental point unless the interpretation of the ability parameter is extended to embrace any temporary or external influencing factor.

The other issue which arises is that conventional, mainstream IRT, whose models routinely admit only properties of test subjects and of highly constrained sets of items, with no explicit error term, has no obvious apparatus which we can use to investigate the contribution of multiple sources of error such as we typically find in educational assessment. This limits its potential usefulness in the context of the kinds of tests and examination that are prevalent in the UK.

Some work is ongoing to improve the treatment of reliability in IRT by grafting on techniques drawn from generalizability and elsewhere (*cf*, for example, Bock, Brennan & Muraki, 2002). The issue of extending IRT coverage beyond persons and items has also recently begun to be addressed, for example in multidimensional IRT (MIRT) (e.g. Béguin & Glas, 2001 and references therein), which is becoming ever more accessible through advances in Bayesian estimation techniques. Many-facet Rasch measurement (Linacre, 1994; Linacre & Wright, 2002) is a potentially interesting direction which could merit further investigation. It is to be hoped that IRT treatments of reliability may benefit from some of these advances.

2.7 Variance analysis and generalizability

Generalizability theory arises out of classical True Score Theory, but differs from it in flexibility and sophistication. In both cases the assumption is that an 'observed score' is equal to a 'true score' (called 'universe score' in G-theory) plus an error component. But in ordinary True Score Theory the error component derives from one *single* undifferentiated source of variance. In an alternative forms reliability study this source of error variance is the interaction between persons and tests, in a test-retest reliability study it is the interaction between persons and occasions (of testing), in a split half reliability study it is the interaction between persons and subtests, and in an internal consistency reliability study (KR-20 and alpha) it is the interaction between persons and test questions. These different possibilities for quantifying reliability lead to as many different reliability coefficients, each depending on the nature of the conceptualised measurement error – in G-theory terms we might say depending on assumptions made about the relevant 'universe of generalisation' (respectively, the universe of substitutable tests, the universe of testing occasions, the universe of half tests, or the universe of test questions).

In G-theory the error component can be broken down into several different subcomponents, the contributions to error of these separated components quantified, and their effects combined in a single comprehensive reliability coefficient, or 'generalizability coefficient'. This is one of G-theory's principal strengths. It also, importantly, offers a "what if?" facility, which allows us to use information about contributions to measurement error to see how we might improve assessment procedures to reduce this error in future applications. This is its second principal strength.

Also in contrast with conventional True Score Theory, which addresses only one type of measurement aim, that of ranking individuals as consistently as possible on the measuring scale, G-theory identifies two types of measurement: 'relative measurement' (the ranking application) and 'absolute measurement'. Absolute measurement is concerned with the precision with which individuals are located on the scale, irrespective of where others might be. In criterion-referenced measurement,

the aim is to make mastery or grading decisions with maximum confidence when applying criterion cut scores to test results, such as in National Curriculum testing, school leaving qualifications and workplace assessment.

G-theory as originally formulated is underpinned by the conceptual and computational apparatus of the analysis of variance (ANOVA), which provided the machinery needed for extracting whatever variance components were called for by the investigator. The analysis of variance freed measurement theorists from having to work with only two variables at a time. With conventional reliability analyses, the experimenter must choose, for example, between test reliability and marker reliability. Using analysis of variance techniques, item, person and marker effects could all be included in the same analysis, leaving it to the investigator to decide which of these contributed to the 'true score' variance and hence to determine the numerator of the reliability coefficient.

Statistical and computational theory and practice have moved on, of course, since the theory of generalizability first appeared in the 1960s and '70s. In particular, the analysis of variance, if viewed merely as a collection of procedures for manipulating relatively simple linear models and extracting their variance components, is now effectively subsumed into more sophisticated, encompassing constructs like Generalized Linear Models (Dobson & Barnett, 2008; McCulloch, Searle & Neuhaus, 2008) and Multilevel Models (Snijders & Bosker, 1999; Goldstein, 2003).

As a conceptual tool, however, G-theory continues to offer valuable insights into the subtleties of determining reliability through the partitioning and unravelling of often complicated variance structures. It also makes accessible the extensive experience in experimentation and survey design that form an inseparable part of its ANOVA heritage.

The variance structure approach to reliability is discussed further in Chapters 3 and 4. In Chapter 3 we offer a reworking of the standard True Score Theory reliability indices, described earlier in this chapter, using unified G-theory tools and techniques. Chapter 4 gives a flavour of the kinds of analyses made possible by G-theory which would be inconceivable using only the conceptual apparatus of classical True Score Theory.

3 The variance components model

3.1 Reliability coefficients and standard errors of measurement

We have outlined in Chapter 2 the evolution in reliability conceptualisation from the correlational approach of the early pioneers to the variance analysis approach of G-theory. The consolidating role of G-theory in this evolution is very clear. G-theory has replaced the traditional reliability indicators of True Score Theory (alternate forms, test-retest, split half and also Cronbach's α), by subsuming them as special cases in a more all-embracing conceptualisation. G-theory is a random sampling model, which, under the usual assumptions of linear modelling, provides a means of estimating the precision of measurements in situations where these measurements are subject to multiple sources of error (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Cardinet & Tourneur, 1985; Shavelson & Webb, 1991, 2006; Brennan, 1992, 2000, 2001b; Cardinet, Johnson & Pini, 2009; Raykov & Marcoulides, 2010, Chapter 9). It is therefore an approach with natural potential for exploitation in the context of educational testing and examining, whether academic, vocational or professional.

For each type of measurement there is an appropriate form of reliability, quantified in a G coefficient. The 'coefficient of relative measurement', $E\rho^2$, was the coefficient that Lee Cronbach originally defined as *the* generalizability coefficient. This coefficient enables us to estimate how precisely an assessment procedure can locate individuals, whether pupils, examination candidates or company employees, relative to one another on a measurement scale. The 'coefficient of absolute measurement', Φ, that Brennan and Kane (1977a, 1977b; see also Brennan, 2001b, p.35) defined as the 'dependability coefficient', evaluates the ability of a procedure to locate individuals reliably on an *absolute* scale.

The formula for Φ is identical with that for ρ^2, any difference in computed value being due to the fact that in absolute measurement there are typically more contributors to error than there are in relative measurement. Developed at the same time as Φ by Brennan and Kane (1977a, 1977b), $\Phi(\lambda)$, a 'coefficient of criterion-referenced measurement', addresses cut score applications (Brennan, 2001b, p.48). This coefficient indicates how reliably an instrument can locate individuals with respect to a cut score, λ, on the measurement scale, with clear implications for the possibility of misclassification.

Each G coefficient is a ratio of true (or universe) score variance to the combination of true score variance and measurement error variance. Informally, we could consider this as a kind of signal-to-noise ratio, except that the denominator includes both signal and noise (signal-to-noise ratios also exist – see Brennan & Kane 1977b and Brennan 2003, pp.14-15). G coefficient values are in the range 0 to 1, as for α, with 1 indicating perfectly reliable measurement, 0 indicating totally unreliable measurement, and 0.7-0.8 generally agreed as the range of lowest acceptable values for scores to be considered 'reliable'. Criterion values are relatively arbitrary choices, and an 'acceptable' coefficient value could be different for different kinds of application, depending on the purposes of the assessment. For example, the Dutch Standards for testing (Evers et al., 2002) advise that for important decisions about individuals coefficients of 0.8 and above would be sufficient, 0.9 and above being very good.

The square root of the measurement error variance is the standard error of measurement, or SEM, which is a measure of score *precision* (not accuracy, which has to do with how valid, as in 'on target', the measurement is). In G-theory the SEM relates to an average score, which in an examining context would typically be an examinee's average test score (over several tests) or average question score (over the questions in a single test). As such it is based on the same metric as the average score that it refers to. In other words, if we have a test comprising questions each carrying three marks then the SEM for an individual's average question score will also be estimated on a 0-3 scale. To produce an SEM for a total test score the average score SEM is simply multiplied by the number of questions that comprise the test. Under Normal distribution assumptions we can use the SEM to produce confidence intervals around the appropriate score estimate in the usual way.

The SEM is a critical piece of information to consider, and is at least as important as calculation of reliability coefficients. It was already being proposed as the most appropriate measure of score reliability 40 years ago, by Skurnick and Nuttall (1968), in the context of classical True Score Theory. Most recently, reflecting on his lifetime's work, Cronbach himself identified the facility to calculate the SEM as *the* essential contribution of G-theory (Cronbach and Shavelson, 2004). Use of the SEM is also recommended in the US *Standards for educational and psychological testing* (AERA/NCME/APA, 1999). In many modern-day applications, particularly in workplace assessment, where variance ratios are often meaningless by default, the SEM is the only useful indicator of reliability.

As noted by Raykov and Marcoulides (2010, Chapter 9), "the notion of a universe lies at the heart of generalizability theory". In the G-theory approach, therefore, the first requirement is to identify the intended universe of generalisation. This means identifying all observable factors that can be assumed or suspected to affect the dependent variable, which in this context is an individual's test score or other form of assessment outcome such as a rater judgement, to decide over which of the factors the assessment outcome is to be generalised, and to note which factors are being implicitly or explicitly sampled in the assessment procedure. Those factors that are sampled will potentially contribute to measurement error. Test questions or assessment tasks, along with markers, raters, workplace assessors or verifiers, are examples. So also are essay topics, item formats, and so on.

It is rarely the case that the particular questions used in a test are the only ones of interest for subject assessment. They are therefore by default a sample of all the questions that might have been used in their place, and they will in consequence contribute to measurement error. Similarly, the markers employed to evaluate test and examination performances are seldom of interest in their own right. They, too, are essentially sampled from some marker population, comprising actual or potential markers, and they too will contribute to measurement error. Ignoring the sampling status of questions, or modifying the characteristics of the marker samples, for example through the usual procedure of standardisation, will typically result in higher apparent assessment reliability, but this will be at the cost of the validity of the reliability outcome (see Kane, 1982, for an interesting discussion on this 'reliability-validity paradox').

An appropriate generalizability study, or 'G study', will allow investigation of those contributions to score variation *whose effects can in practice be observed*. Respective influences on scores are quantified in the form of estimated variance components, using classical analysis of variance (ANOVA) or some other appropriate methodology (see Searle, Casella & McCulloch, 2006 for a comprehensive discussion of component estimation). The component information is then used to calculate SEMs and, if appropriate, G coefficients.

A follow-on decision study ('D study') – the "what if?" analysis – permits predictions of reliability and measurement error should features of the current assessment procedure be changed in a future application. Possible changes might include increasing or reducing the numbers of questions in the test or the number of workplace tasks to be assessed, independently or simultaneously increasing or reducing the number of markers marking each script or of assessors judging each workplace task, introducing a set of focused units in place of a lengthy one-off examination, and so on. What we might decide to change will depend on what the outcomes of the G study are. Those factors that are found to make the largest contributions to measurement error will be the prime candidates for increased sampling in the future (typically test questions and markers) while those that contribute least to measurement error will be potential candidates for decreased sampling. Using this facility allows us to optimize measurement quality by maximally increasing precision within any given financial, logistic or other constraints.

3.2 Subsuming classical reliability indicators as special cases

Subsuming classical test-based indicators

As we have noted more than once in this report, classical reliability indicators are subsumed as special cases in the more comprehensive framework of G-theory. Consider, for example, the alternate (or parallel) forms approach to reliability estimation. Two or more tests can be strictly parallel in the sense that they are designed to have exactly the same mean score and standard deviation. Alternatively, they can be randomly parallel, with no empirical constraints imposed; an example would be where tests are drawn by random sampling from within a large question bank, representing a subject domain. Or they can be tests constructed manually by an examiner following some given test specification, and designed on the basis of expert judgement to be interchangeable for the purpose of use in a particular assessment programme – the examiner's assumption would be that whichever test is used the candidate outcomes would be the same. The outcome of administration of two such tests, typically administered within a very short time interval to minimise maturation and relevant learning effects, will be two sets of test scores, one set for the first test and one for the second. Every person tested, whether pupil, school-leaving examination candidate or workplace employee, will have two scores in the set. For any individual the two scores are considered to be independent, in the sense that the score achieved on one of the tests is assumed not to have any direct influence on the score achieved on the other. The classical reliability indicator is simply the correlation between the two sets of test scores.

In experimental design terminology this is an example of a two-factor repeated measures design. Persons and tests are the two factors, or independent variables, that

impact on test scores, the dependent variable. Since every person attempts both tests, persons and tests are described as 'crossed factors'. At this point it might be useful to note that G-theory terminology differs from general experimental design terminology in some respects. In particular, G-theory typically speaks of 'facets' rather than 'factors'. A facet in G-theory is synonymous with a factor in ANOVA (but see below). Cronbach and his associates followed Guttman in opting for the new term 'facet', to avoid confusion in psychometric circles with the factors of factor analysis (Cronbach et al. 1972, p.2 footnote). Another example of terminological difference is that this particular testing pattern – a 2-factor crossed design – is in G-theory called a one-facet design. This is because in the original development of G-theory, 'persons' were the unique object of measurement, and the variables that could influence the testing outcome for individual persons were 'facets' of the assessment procedure. Here, 'tests' is the only identified facet. The shorthand symbolic representation of this design in G-theory is p × t, or more simply pt, where p and t represent, respectively, persons and tests, and '×' indicates that persons and tests are crossed, i.e. every person attempts all the tests.

We can express the observed score, X_{ij}, of person i on test j in terms of the following linear model:

[3.1] $X_{ij} = \mu + (\mu_i - \mu) + (\mu_j - \mu) + (X_{ij} - \mu_i - \mu_j + \mu)$

where

- μ is the overall mean score of all persons (in the relevant population) on all tests (in the universe of interchangeable tests)
- $(\mu_i - \mu)$ is the difference between the mean test score of person i (i.e. the mean test score of a randomly selected person over all the interchangeable tests in the test universe) and the overall mean score, i.e. person i's deviation score, also known as the 'person effect'
- $(\mu_j - \mu)$ is the difference between the mean score of test j and the overall mean score, i.e. test j's deviation score, or the 'test effect'
- $(X_{ij} - \mu_i - \mu_j + \mu)$ is the residual (what is left when everything else is accounted for).

[3.1] is simplified in [3.2], in which the person effect and the test effect are symbolised by α_i and β_j, respectively, with γ_{ij},e_{ij} representing the residual confounded with measurement error:

[3.2] $X_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},e_{ij}$

If we square each person's deviation score, producing $(\mu_i - \mu)^2$, then sum the squared deviations over all persons and average, we will have the between-persons variance, σ^2_p. In other words we will have a quantification of the amount of variation that exists between the average test scores (averaged over the two tests) of the tested individuals. We can do the same with test score deviations, to provide a quantification of the between-tests variance, $(\mu_j - \mu)^2 = \sigma^2_t$. Repeating the process for the final term in [3.2], this time averaging the deviations over both persons and tests, provides a quantification of the confounded person-test/residual variance, $\sigma^2_{pt,e}$. Finally, if we do the same with the squared deviations of the observed person by test scores from the

overall mean score, i.e. $(X_{ij} - \mu)^2$, we will have a quantification of the total variation in observed test scores, i.e. σ^2_X.

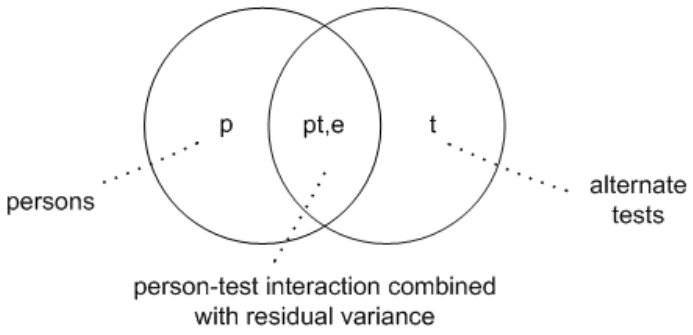
Why is all of this relevant? It is relevant in the sense that if we accept the usual linear modelling assumptions about the relationships among α_i , β_j and γ_{ij},e_{ij}, i.e. that these effects are uncorrelated and that the expected value of the errors is zero, then we will be able to express the total observed score variance as a linear combination of component variances: in this case the between-persons variance σ^2_p, the between-tests variance σ^2_t, and the residual-cum-error variance $\sigma^2_{pt,e}$. If we subtract μ from each side of [3.1] and then square the two sides, it follows that:

[3.3] $\sigma^2_X = \sigma^2_p + \sigma^2_t + \sigma^2_{pt,e}$

In other words the total observed score variation can be partitioned into the variation due to differences between persons plus the variation due to differences between tests, plus any remaining variation not otherwise accounted for.

At this point, and throughout this chapter, we offer a variance partition diagram for illustration (Figure 3.1). Variance partition diagrams were first introduced by Cronbach and his associates in their seminal book on G-theory (Cronbach et al., 1972, p.37), as a potentially useful graphical device to show how the total variance in a set of scores can be attributed to the various different identified contributors. Variance partition diagrams have the appearance of Venn diagrams, but they have a different interpretation. In a two-circle Venn diagram each circle represents a set of some kind, the members of the set sharing some particular quality, such as colour of hair, being mammals, being made of glass, round objects, being a member of the professional assessment community, or whatever. The interaction between the circles represents those individuals or objects that belong to both sets, in other words those that have both relevant qualities: round glass objects, dark haired assessment professionals, and so on.

Figure 3.1 Variance partition diagram for the crossed design p × t (alternate forms administration), with p and t representing, respectively, person and test variance and pt,e representing pupil-test interaction variance confounded with residual variance



In a variance partition diagram, on the other hand, circles represent factors (here, persons or tests), and the sectors created when circles intersect represent the contributions of these factors and their interactions to total observed score variance (but note that relative sector sizes do not reflect the relative importance of the different factors as contributors to total variance). The intersection in a two-circle

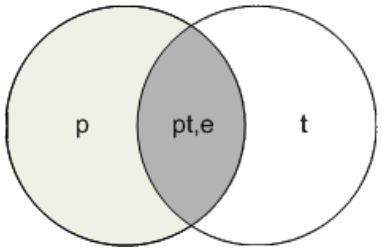
diagram represents a confounded interaction effect, since it will also represent all residual variance.

Usually, though not always, interaction variances that involve the 'object of measurement' contribute to measurement error, of both types, relative and absolute. The only time that such interaction variances do not contribute to measurement error is when the factors interacting with the object of measurement are not randomly sampled but are 'fixed'– this is explained more fully in Chapter 4. Main effects, and interactions among these, contribute only to absolute measurement error, and only then if the factors concerned are considered randomly sampled.

In Figure 3.1 the three sectors in the diagram represent three variances: between-persons variance (p), between-tests variance (t) and pupil-test interaction variance confounded with residual variance (pt,e). The residual variance will subsume any unidentified 'hidden' influences on the test scores – hidden factors – as well as random fluctuations.

In this context of relative measurement the estimated universe score variance is the between-person variance, σ^2_p. This will be the numerator in the *relative* G coefficient. The only contribution to error variance here will be the confounded person-test interaction effect, $\sigma^2_{pt,e}$. This is because if the aim is simply to place individuals relative to one another on the measurement scale then any between-tests variance, σ^2_t, will be irrelevant. If one test has a higher mean score than the other this simply moves all candidates up the scale without changing their standing relative to one another. Figure 3.2 illustrates the attribution of the three variances to 'valid' variance and to 'error' variance for the relative measurement of persons.

Figure 3.2 Variance attribution diagram for an 'alternate forms' administration, with p and t representing, respectively, persons and test forms



The error variance in this case is $\sigma^2_{pt,e}/n_t$, i.e. the person-test variance component divided by the number of tests that each person attempted (n_t is equal to two in this classical alternate forms example). This is the sampling variance for a generic person's average test score, and is denoted in G-theory for this type of relative measurement application by σ^2_δ. Since we are dealing with a sample-based estimator, we can consider the positive square root of the sampling error variance, i.e. σ_δ, as a standard error, in this case the standard error of measurement (SEM) for a person's average test score. [Note that if we are interested in a person's combined test score rather than the average test score then the appropriate SEM would be $n_t\sigma_\delta$]. Henceforth we shall assume that square roots of variance components used as standard errors are always positive, and will not labour the point each time in the text.

The denominator in the associated G coefficient is the sum of the between-persons variance and the adjusted person-test/residual variance (the 'noise'). The relative G coefficient, ρ^2, is, then:

[3.4] $$\frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pt,e}^2 / n_t}$$

In the traditional alternate forms application any between-test variance, i.e. any difference in the overall difficulty of the two tests, is of no interest. All that matters is how similarly the tests locate individuals relative to one another on the common score scale. Should the tests produce exactly the same relative locations for individuals on this scale then we would have perfect 'reliability'. The less well the relative positioning of individuals coincides the less reliable the assessment will be considered to be.

The G coefficient for relative measurement usefully replaces the classical alternate forms inter-test correlation. The more interesting point to note is that the alternate forms approach can very easily be extended under G-theory, by being applied to situations in which more than two tests are used. Thus candidates could be asked to sit three tests, or four or five, or however many we think they might tolerate, and that the budget can afford, and still remain motivated to apply themselves seriously. In this way we would have more observations of person by test interaction, thus reducing the measurement error arising from this when differentiating individuals, whether school pupils, external examination candidates or workplace employees. In expression [3.4] we would simply divide the interaction variance component by the appropriate number of tests used, which will now be greater than two.

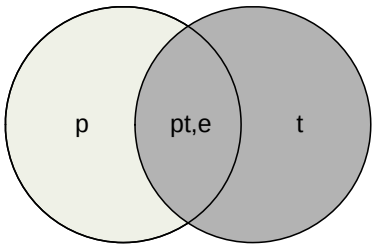
The higher the number of interchangeable tests that we might be able to administer to the same persons the more precise our variance component estimates will be. Once we have the estimates, we can plug them into [3.4], along with the current value of n_t, to estimate current score reliability, as a variance ratio or SEM. But more than this, we can substitute n_t with any other value to estimate reliability with any feasible alternative number of test administrations, and in this way see to what degree we might minimise the measurement error, or optimise the measurement, within any given constraints (the "what if?" analysis). If we assume that the various interchangeable tests that we use are drawn at random from the (real or virtual) 'universe' of such tests then we can generalise the empirical results to that universe.

The traditional test-retest and split half approaches to reliability estimation can be treated similarly, simply by replacing test forms (t) in the two-factor crossed design with test occasions (o) or split half subtests (s). And again we are not limited to just two occasions of testing or to two subtests, but can use as many as are practicable.

But suppose the primary aim of the assessment had not been person differentiation for the purpose of norm referenced decisions. Suppose that we were rather attempting to locate individuals *absolutely* on the measurement scale, irrespective of where their fellow test takers might be. What difference would this make to the way we conceptualise and quantify the measurement error? The answer is that we would now need to recognise a second contributor to measurement error, that is between-test

variation, since any differences in the general difficulty of the tests used would become newly important. This is especially the case where criterion-referenced cut scores are to be applied. Even if individuals are similarly located relative to one another on the measurement scale by each of two or more tests, if the tests differ in difficulty then the decision outcomes for each individual could be different, too, depending which test they take in the 'live' assessment. In this context the person-test interaction would be joined by the between-tests effect as a contributor to error variance. Figure 3.3 illustrates this.

Figure 3.3 Variance attribution diagram for absolute measurement of persons (p), using several different tests (t)



The error variance becomes $[\sigma^2_t/n_t + \sigma^2_{pt,e}/n_t]$, symbolised as σ^2_Δ. [We could have written the error variance expression as $(\sigma^2_t + \sigma^2_{pt,e})/n_t$, but have not done so in order to preserve the clarity of distinction between the two error contributions].

The *absolute* G coefficient, Φ, is given by $\sigma^2_p/[\sigma^2_p + \sigma^2_\Delta]$, or, in detail:

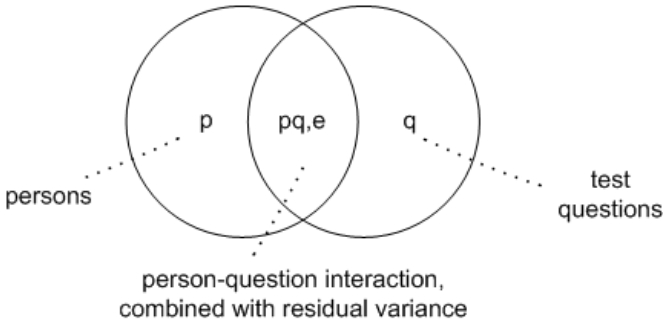
$$[3.5] \quad \frac{\sigma^2_p}{\sigma^2_p + \left(\sigma^2_t / n_t + \sigma^2_{pt,e} / n_t\right)}$$

The square root of the error variance, i.e. $\sqrt{(\sigma^2_t/n_t + \sigma^2_{pt,e}/n_t)}$ or σ_Δ, is this time the SEM for *absolute* measurement. As before, the SEM metric concerns an individual's average test score. If we need the SEM associated with a generic individual's total score over the different tests then we multiply σ_Δ by n_t.

Subsuming Cronbach's α

The three classical approaches to reliability estimation considered thus far all focus on relative measurement only, and they are all based on test scores. The alpha coefficient, as we have seen in Chapter 2, is also uniquely focused on relative measurement, but this time looking at item or question scores within a single test rather than at whole-test scores. Again, a two-factor crossed design represents the practical situation; Figure 3.4 illustrates variance partition in this case. As before, the total score variance can be attributed to three sources: persons, whose average question scores will vary to some degree, test questions, whose overall scores (averaged over persons) will also typically vary, and the person by question interaction combined with any other unidentified influencing variables and residual random fluctuations.

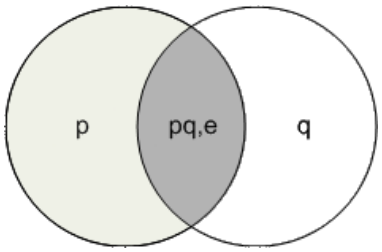
Figure 3.4 Variance partition diagram for the crossed design p × q (single test administration)



In this testing application persons are again the focus of interest, the collection of test questions – the test – being merely a device for person measurement in some given sense (numeracy ability, historical knowledge, height measurement skills, or whatever). Both persons and questions are considered conceptually as having been randomly sampled from some larger group, the population of persons or the universe of questions (this latter being virtual to a greater or lesser degree).

Once again, we can consider two application possibilities: relative measurement and absolute measurement. The variance of true scores is in both cases the between-persons variance, σ^2_p. A measurement error contributor for both types of measurement will be the confounded person-question interaction variance, $\sigma^2_{pq,e}$ (see Figure 3.5). A second contributor to measurement error variance for *absolute* measurement applications is the between-questions variance, σ^2_q (see Figure 3.6).

Figure 3.5 Variance attribution diagram for the relative measurement of persons using a set of test questions

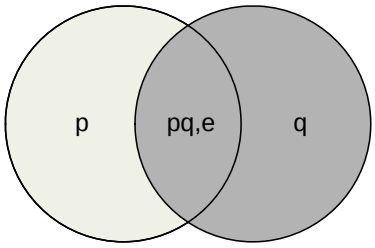


The G coefficient for relative measurement, ρ^2, is:

$$[3.6] \quad \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pq,e}/n_q}$$

As we see, this expression is exactly analogous to that given earlier [3.4] in the context of alternate forms reliability investigation, *t* for tests simply being replaced here by *q* for questions. In this particular case, however, what we have is in essence Cronbach's α coefficient: the values of ρ^2 and α will be identical. Again, the SEM is the square root of the error variance, i.e. $\sqrt{(\sigma^2_{pq,e}/n_q)}$, or σ_δ. This will be the SEM for a person's average question score. We multiply this SEM by n_q to find the SEM for a person's total test score, the usual focus of interest in cut score applications. In other words, the SEM for an individual's total score on a test is $n_q\sigma_\delta$.

Figure 3.6 Variance attribution diagram for the absolute measurement of persons using a set of test questions



If we are interested in locating a person as precisely as possible on a measurement scale, then mere rank ordering is not sufficient. Unless the questions comprising the test are considered to be the only questions of interest then variation in question difficulty will now become relevant, between-questions variance becoming a second contributor to measurement error. In other words, we need now also to take into account the additional measurement error that will have arisen from the question sampling. So the absolute error variance will again be a composite term, comprising both the question variance and the confounded person-question interaction variance, each term divided by the number of sampled questions, i.e. $(\sigma^2_q/n_q + \sigma^2_{pq,e}/n_q)$. As usual, the standard error of measurement for a person's average question score is the square root of this quantity. Multiply by n_q and we will have the SEM for a generic person's total test score.

The expression for Φ, the absolute G coefficient is:

$$[3.7] \quad \frac{\sigma^2_p}{\sigma^2_p + \left(\sigma^2_q / n_q + \sigma^2_{pq,e} / n_q\right)}$$

When cut scores are to be applied, Φ is replaced by $\Phi(\lambda)$, a criterion-referenced variant (Brennan 2001b, p.48), in which λ denotes the cut score:

$$[3.8] \quad \frac{\sigma^2_p + (\mu - \lambda)^2}{\sigma^2_p + (\mu - \lambda)^2 + \left(\sigma^2_q / n_q + \sigma^2_{pq,e} / n_q\right)}$$

Once we have the expressions for the reliability coefficients and SEMs, we need only to substitute the relevant values of the estimated variance components into them, along with n_q, to find their values estimated for a given testing situation. But again we can do more than this. We can substitute different question sample sizes, i.e. different values of n_q, into the expressions to predict the effect that changes in test length might have on score reliability and precision. As described earlier for the alternate forms approach, this would be a simple example of assessment optimization (always recognising that what might be optimal in theory might not be achievable in practice because of resource constraints – see, for example, Marcoulides 1993, 1995, 1997 on the issue of optimisation within budget constraints).

These G coefficients, like α itself and the three test-based correlation coefficients, are sensitive to the size of the pupil variance, and not only to the size of the measurement

error. The higher the pupil variance the higher will be the values of the coefficients, and conversely the lower the pupil variance – the more homogeneous the group of persons taking the test(s) – the lower will be the coefficient values. When the focus of the assessment is relative measurement, when we are attempting to differentiate among individuals as well as possible by spreading them on a score scale, then one strategy to achieve this is to select questions for the test that serve this purpose best. Hence the classical pretesting strategy of rejecting items that have very high or very low facilities, along with those whose discriminations (correlation between item and the rest of the collection of items) are low (traditionally below 0.3 or thereabouts). If the aim of the assessment is not to spread individuals, but is rather to locate them with an absolute value on a score scale, then this pretest strategy is inappropriate.

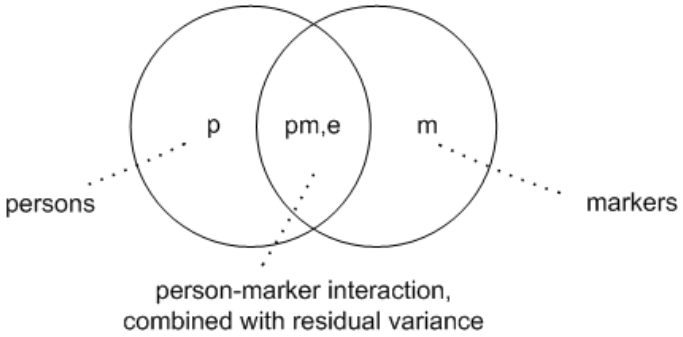
Other important influences on the magnitude of the coefficients are clearly the size of the person-question interaction variance (confounded as it is with residual variance), the size of the between-questions variance (in the case of absolute reliability), and the number of test questions in the test (since the higher the number of questions the more the two question-related variances are reduced). In the case of the relative coefficient (and α) the number of questions in the test has a greater influence on the measurement error, and hence on the value of the coefficient, than the degree of average inter-correlation among the test questions (for a demonstration see Cortina, 1993).

For $\Phi(\lambda)$, the location of the cut score relative to the mean of the score distribution is also influential. The further the cut score is from the mean test score the higher will be the value of $\Phi(\lambda)$. This makes intuitive sense, since applying a cut score to a high point in the score distribution could result in the maximum number of individuals being misclassified, whereas applying a cut score to the tail of a distribution would minimize classification error. Note in this sense that $\Phi(\lambda)$ tells us nothing about the *validity* of a cut score choice, nor about the practical meaning of the resulting classifications.

3.3 Marker reliability studies

As a final example of a two-factor design, consider a typical marker reliability study in the context of academic external examinations (for details, see, for example, Meadows and Billington, 2005, p.48; Sykes et al. 2009). Here, markers have traditionally been brought together for ‘standardisation’ before marking assigned examination scripts. Before and during the standardisation meeting the markers (assistant examiners), are expected to familiarise themselves with the examination paper that they will be marking, and with the mark scheme they will use. There will be discussion about the paper and the mark scheme, and a few exemplar scripts might be evaluated in a plenary session. Markers are then given a batch of scripts to mark, a sample of which is later also marked by a senior examiner. An alternative strategy is to have a senior marker pre-mark certain scripts or questions and to seed these among markers’ batches. The resulting mark distributions are compared across the two markers. The same procedure is used by awarding bodies in the vocational sector, when assessors are standardised for evaluating performances on ‘significant tasks’, through comparison with the judgements of internal verifiers. For these kinds of standardisation strategies, we have again a two-factor crossed design, as shown in Figure 3.7.

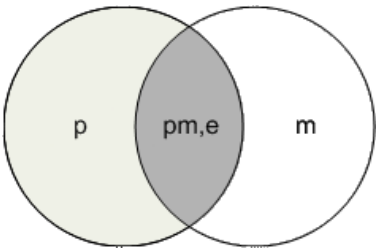
Figure 3.7 Variance partition diagram for the crossed design p × m (marker reliability study)



Should the assistant examiner's marks be consistently more severe or more lenient than those of the senior examiner then the marks for all the scripts marked by that examiner will be adjusted upwards or downwards by some appropriate amount, thus reducing the between-markers variance. Should the senior examiner feel uncomfortable with the degree of consistency evidenced in the assistant examiner's marks, in other words should the assistant examiner have produced a 'jagged profile', then a complete re-mark might be requested and the process repeated. A jagged profile, of course, is evidence of person-marker interaction, senior examiner and assistant examiner appreciating different scripts to different extents. Since the implicit assumption in such marker standardisation exercises is that the senior examiner is always 'correct', a jagged profile is considered as evidence that the assistant examiner cannot mark consistently. So what we have going on here is an attempted reduction in both between-marker variance and person-marker (i.e. script-marker) interaction variance for the live examination.

If the assumption that senior examiners 'carry' standards, an assumption which is questionable, and which has indeed been shown in at least one research study to be untenable (Johnson & Cohen, 1983, 1984), were to be abandoned, then both examiners could be considered as interchangeable in principle, and G coefficients and SEMs could usefully be calculated. And if more than two markers could be required to mark a batch of the same scripts, then the results would be more dependable and more readily generalised. Figure 3.8 illustrates the appropriate variance attribution for such a marker reliability study.

Figure 3.8 Variance attribution diagram for the relative measurement of persons (scripts) using multiple markers



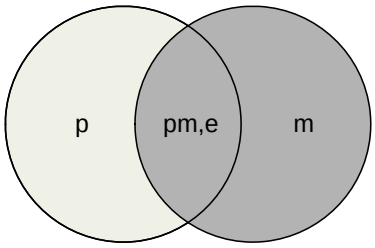
The G coefficient for relative measurement, ρ^2, is in this case:

$$[3.9] \quad \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pm,e}^2 / n_m}$$

where σ^2_p is, as usual, the between-persons (in practice the between-scripts) variance, which reflects the variation between the overall marks awarded to the evaluated examination scripts; $\sigma^2_{pm,e}$, is the confounded person-marker (script-marker) interaction variance, which reflects the degree to which different markers agree or disagree about the marks to award to one script compared with another; and n_m is the number of markers involved. As noted by Bramley (2007), this coefficient is identical in form to Cronbach's α coefficient, the internal consistency of markers becoming the focus in place of the internal consistency of test questions (cf. [3.6]).

If it is the absolute value of the person's score that is important, rather than person ranking, then we should be exploring the reliability of absolute measurement, which the between-marker (intermarker) variance will additionally influence (see Figure 3.9).

Figure 3.9 Variance attribution diagram for the absolute measurement of persons (scripts) using multiple markers



The expression for Φ, the absolute G coefficient is:

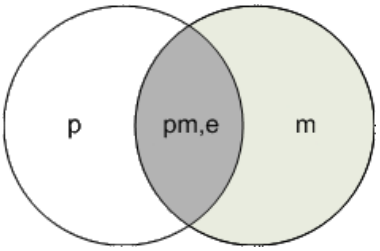
$$[3.10] \quad \frac{\sigma_p^2}{\sigma_p^2 + \left(\sigma_m^2 / n_m + \sigma_{pm,e}^2 / n_m\right)}$$

The aim of marker standardisation is clearly to minimise, if not to eradicate, the between-markers variance. The person-marker interaction effect – already reduced by replacing 'aberrant' markers – could then be minimised through the multiple marking of scripts (although the relationship with marker numbers is not linear, and there will eventually be diminishing returns for higher marker investment).

As an interesting aside, this same two-factor design can be used to provide an indication of the number of common scripts that markers should ideally be asked to mark in order for markers themselves to be reliably identified as 'functioning adequately' or not, in the sense of their relative severity or leniency. For if we were to ask several markers independently to mark, say, 15 scripts, we might identify one or other as lenient, normal or severe using some given criterion. If we increased the number of scripts to, say, 25 we might begin to see a different picture. This will be because of the influence of script sample size on the comparative outcomes. Through the 'principle of symmetry', identified by Cardinet, Tourneur and Allal (1976, 1981,

1982), we could use the same study data to see how reliably markers had been positioned relative to one another on the common measurement scale, rather than looking at how well scripts had been relatively located. Attention would switch from persons (scripts) to markers. Figure 3.10 illustrates this shift in focus.

Figure 3.10 Variance attribution diagram for the relative measurement of markers on the basis of marked persons (scripts)



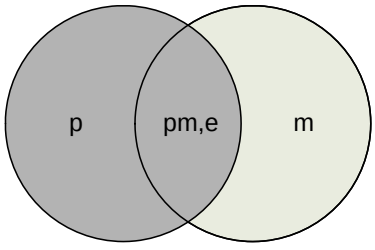
The *relative* G coefficient for marker measurement is

$$[3.11] \quad \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{pm,e}^2 / n_p}$$

where σ^2_m represents the between-markers variance and n_p is the number of persons (scripts) evaluated by each marker.

But perhaps relative measurement is not the most appropriate form for comparing markers' standards? We are surely more interested in how different the markers' absolute marking standards are than whether or not different markers come out with the same script ranking. We should, therefore, be thinking instead about absolute measurement when we explore marker reliability. The effect on measurement error of the between-person (between-scripts) variance therefore needs to be taken into account as well – see Figure 3.11 for variance attribution.

Figure 3.11 Variance attribution diagram for the absolute measurement of markers on the basis of marked persons (scripts)



The G coefficient for absolute marker measurement is then:

$$[3.12] \quad \frac{\sigma_m^2}{\sigma_m^2 + \left(\sigma_p^2 / n_p + \sigma_{pm,e}^2 / n_p\right)}$$

If these coefficients could be calculated both before and after marker standardisation then the impact of the standardisation could be measured. But however effective marker standardisation might be, there will always remain differences in the overall marking standards of markers and in the degree to which different markers are influenced to move away from any general standards in an inconsistent way when faced with individual scripts. This is why it is always advisable to multiple mark scripts in cases where mark schemes and other factors leave room for marker subjectivity. It is in consequence problematic that even dual marking is apparently impracticable to implement within the examinations system in the UK, because of the sheer proliferation of qualifications combined with a finite number of potential examination markers and time constraints (Meadows & Billington, 2005, p.58).

3.4 A comment on 'hidden' factors

All of these two-factor (one-facet) designs are seriously limited as regards score reliability. This is because, putting aside the impact on question or test scores of the persons themselves, i.e. of the knowledge, ability or skill of the individuals being assessed, each two-factor design focuses on only one other potentially influencing factor, and in so doing ignores other factors that might be at least as influential as the one investigated. Thus, the persons by tests design (pt) looks only at the impact of entire tests, excluding any consideration of questions within the tests or of marker effects, where these, or interactions involving them, might have an influence. The persons by questions (pq) design looks at the impact of test questions, but not at the impact of alternate tests or of markers. The persons by markers (pm) design explores marker effects but to the exclusion of test and question effects, and indeed of any possible mark scheme effects.

By failing to include more than one score impactor at a time, all variations of the one-facet design fail to take into account the influence on scores of other main effects. They equally fail to explore the influence on scores of several potential interaction effects, among main factors and also between main factors and the test performances being evaluated. For example, a marker reliability study in which multiple markers mark a set of student essays might result in a high reliability coefficient. But all that this is telling us is something comforting about the degree of consistency in marking that can be achieved for that one essay topic. Should essays from the same students but on a different essay topic have been rated, would the reliability outcome have been the same? Possibly. But then again possibly not. By focusing on one single essay topic we have essentially chosen to ignore any potential contributions to relative measurement error of person by topic interaction, along with the additional potential contributions to absolute measurement error of topic variance and of marker by topic interaction.

When potentially influencing variables do not feature in an analysis design, we call them 'hidden factors'. Where hidden factors do have a potential impact on scores, then ignoring them, and their effects, will reduce the validity of the reliability measurement. We would be estimating reliability as consistency, or replicability, but only for a restricted set of conditions of assessment. We return to this issue later.

4 Extending variance analysis in reliability estimation

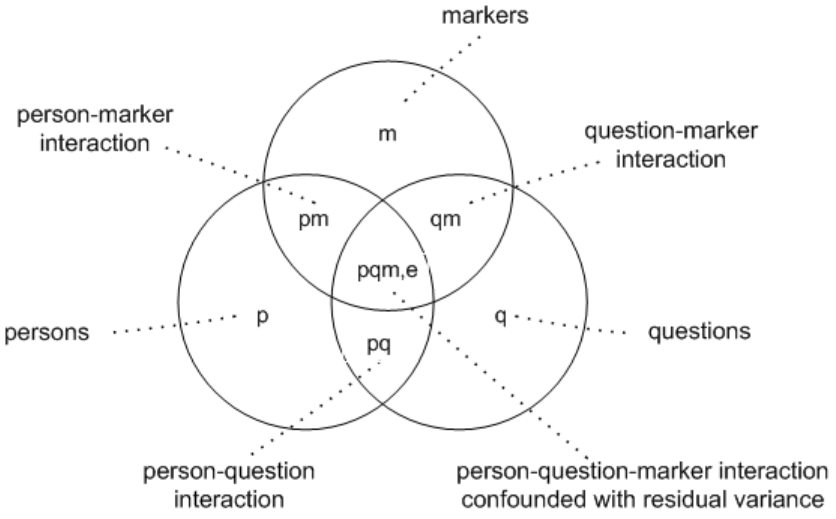
4.1 Investigating question and marker effects simultaneously

The principles of G-theory are readily extendable to more complex designs, accommodating broader composite universes of generalisation. Thus, we can move away from the "one variable at a time" approach of reliability investigation described in the previous chapter and look instead at more comprehensive models of reality. In particular, instead of looking separately at the influence of questions or of markers on measurement error we can explore both potential influences simultaneously, as well as their possible interaction.

In achievement testing there will typically be differences in the difficulty of test questions, and there can be expected to be person-question interaction as well. The larger and the more complex the questions the larger these influences on measurement error are likely to be. Structured questions, of the type that feature in science examinations, and essay questions, so popular in the humanities and social sciences, are also particularly vulnerable to between-marker differences and to marker inconsistency. So, too, are the creative products that are the outcomes of assessments in subjects like art and drama, and probably also the elements that comprise a portfolio of class or workplace based work. Simultaneously investigating the separate and the combined effects of questions and markers on assessment reliability will provide more valid reliability estimates, since we will be looking at the replicability of assessment results over both the universe of possible test questions and the universe (population) of potential markers.

Figure 4.1 illustrates the three-factor model pqm, with p representing persons (for instance, GCSE candidates, Key Stage 2 pupils, workplace trainees), q representing questions, and m representing markers.

Figure 4.1 Variance partition diagram for the crossed design p × q × m



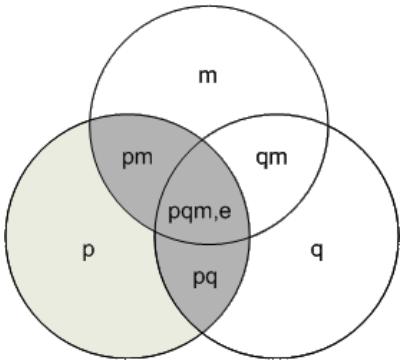
For generalization purposes we assume that both markers and test questions are samples – preferably random representative samples – drawn from within their respective populations or universes. Moreover, we assume here for simplicity that

these populations would be of infinite size (this is not a *necessary* assumption for G-theory application) whether the set physically exists in its totality or not. In the case of test questions the universe could be, for example, the entire set of GCSE A level mathematics questions, some of which will already have been used in past examinations but most of which remain to be developed. Or, in the case of the unit 'Installation of Electrical Systems' in a Level 3 Diploma in Electrotechnical Services, it might be all the possible practical tasks from which a handful might be selected to form the basis for assessment through practical demonstrations. Then again, it might be all the possible examples of created products that an individual might offer within a portfolio of evidence in the 'Design' unit of a Level 1 Award in Creative Techniques in Design.

In the case of markers, for which we can readily substitute workplace assessors, the universe of generalisation would be all those individuals with the appropriate characteristics, practising school teachers, for example, or experienced professionals in the vocational field concerned, including any that might have served in this role in the past or who might serve as such in the future.

Any score variation that arises from differences between markers, from differences between questions, or from interactions between markers and examinees, between questions and examinees, and between markers and questions and examinees, will constitute noise in the system, and will contribute to error variance, for both relative and absolute measurement. Variation between markers, variation between questions, and interaction between markers and questions, will be additional contributors to error variance in the case of absolute measurement. Figures 4.2 and 4.3 illustrate this clearly. Variance sources within the person (examinee) circle, excluding p itself, contribute both to relative and to absolute error variance, while variance sources outside of the p circle are additional contributors to absolute error variance.

Figure 4.2 Variance attribution diagram for the relative measurement of p in the crossed design p × q × m



As before, the universe score variance is simply σ^2_p. The error variances, however, become more complex as the number of facets in the design increases. Here, the error variance for relative measurement, σ^2_δ, is a combination of the three sample-based interaction variances (see Figure 4.2). In the error variance expression each estimated interaction variance component is divided by the number of observations made for each person – the number of markers who marked the person's essays (or practical tasks or creative works) and/or the number of questions, tasks or creative works that the person attempted or produced and that were evaluated by the markers:

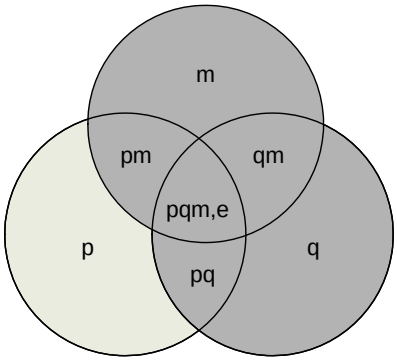
[4.1] $\sigma^2_{pm} / n_m + \sigma^2_{pq} / n_q + \sigma^2_{pqm,e} / n_q n_m$

The square root of this error variance is the SEM for relative measurement. The G coefficient for relative measurement is, as usual, the between-persons variance divided by the combination of between-persons variance and relative error variance:

[4.2] $$\frac{\sigma^2_p}{\sigma^2_p + \left(\sigma^2_{pm} / n_m + \sigma^2_{pq} / n_q + \sigma^2_{pqm,e} / n_q n_m\right)}$$

The absolute error variance is comprised of the interaction variances involving persons, plus contributions from the main effects of markers and questions and the interaction between these two factors (see Figure 4.3).

Figure 4.3 Variance attribution diagram for the absolute measurement of p in the crossed design p × q × m



In other words, the error variance for absolute measurement is given by:

[4.3] $\sigma^2_m / n_m + \sigma^2_q / n_q + \sigma^2_{qm} / n_q n_m + \sigma^2_{pm} / n_m + \sigma^2_{pq} / n_q + \sigma^2_{pqm,e} / n_q n_m$

The square root of expression 4.3 is the SEM for absolute measurement at the level of a person by question by marker score: as usual we multiply by n_q to find the SEM for a person's total test score (averaged over markers).

The absolute G coefficient is given by:

[4.4] $$\frac{\sigma^2_p}{\sigma^2_p + \left(\sigma^2_m / n_m + \sigma^2_q / n_q + \sigma^2_{qm} / n_q n_m + \sigma^2_{pm} / n_m + \sigma^2_{pq} / n_q + \sigma^2_{pqm,e} / n_q n_m\right)}$$

As before, once we have expressions for the G coefficients and for the SEM we can substitute any values that we choose into them for the factor sample sizes (here changing the values of n_q and n_m to indicate alternative numbers of questions and of markers, respectively) to see how changes might affect estimation precision.

This particular 3-factor, or 2-facet, design has featured quite often in real-life applications. Interestingly, a typical finding is that the person by question, or, using more appropriate terminology for the applications concerned, the person-task

interaction variance has proved to be the largest contributor to measurement error, with contributions from raters being more modest, which could be explained by rater standardisation (Brennan, 2000).

For example, Shavelson, Baxter and Gao (1993) implemented this design using science performance data from the California Assessment Program. The science assessment involved five different science tasks and was organised in the familiar 'circus' arrangement. The five tasks were set up at five different stations, and the students rotated around the station at timed intervals. A given rubric was used by teacher raters to score the students' performances on a 5-point scale, every task performance being independently rated by three different raters. The analysis results revealed the student-task interaction to have the highest estimated variance component. This component reflected inconsistent performances across the tasks by individual students, some students doing better on some tasks than others and vice versa. The next largest estimated variance component was associated with between-students variance. Estimated components relating to the between-rater variance, the rater-task variance and the student-rater variance were negligibly small. The absolute generalizability coefficient was 0.70. Further analysis confirmed that increasing the number of assessment tasks whilst reducing the number of independent raters would increase the reliability.

In the context of large-scale assessment of science performance in the UK, relatively important amounts of between-marker variation and marker-question interaction variance emerged. (Johnson, 1989, Chapter 7), as did between-question and pupil-question interaction variance.

A particularly important area of application for G-theory continues to be the health sciences, and in particular the assessment of medical students' diagnostic and patient relationship skills. This type of assessment typically takes a form similar to the circus arrangement described above for science practicals. A number of stations are set up, the task stimulus at each station being a trained actor simulating a patient with particular medical problems and personality. Students move from one station to another, dialoguing with the patient and performing a physical examination before eventually arriving at a medical diagnosis. As the dialogue and examination progress the student is rated, by the 'simulated patients' and/or by medical staff, generally using a rubric to rate the aspects of performance of interest. In such assessment G-theory has been used to explore the contributions to measurement error from the tasks and the raters, both contributions proving to be important (Govaerts, van der Vleuten, & Schuwirth, 2002; Solomon & Ferenchick, 2004; Burch, Norman, Schmidt & van der Vleuten, 2008; Murphy, Bruce, Mercer & Eva, 2009).

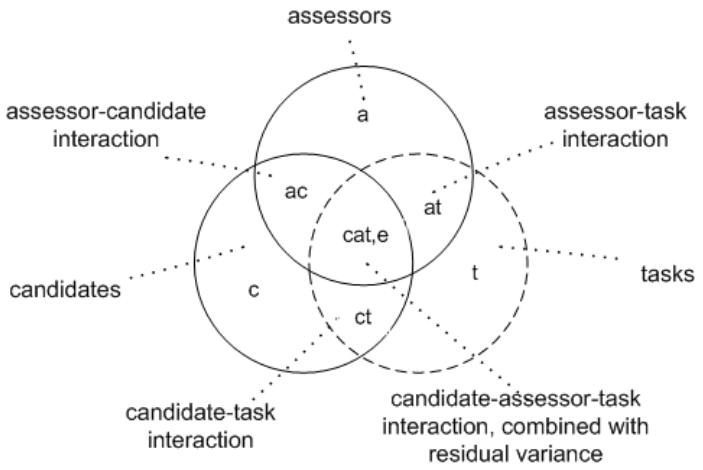
This might be an appropriate point to note that this 3-factor (2-facet) design, or more comprehensive ones, could be applied in situations where different examination candidates attempt essentially different examinations within the same qualification. An example would be a 'layered' examination, where students are entered by their teachers into particular sets of question papers within an examination, each set providing access to a restricted sets of grades. Another would be test components in which candidates are offered a question choice, such as two essay topics from five in a history paper, or three structured questions from six in physics. Here the design could be analysed separately for the different candidate subgroups, and measurement

consequently estimated for the candidates in that group. An important interpretational issue here, however, concerns identification of the student population that the teacher-selected or self-selected examination candidates represent, and identification of the restricted 'task universe', or curriculum, to which the results can validly be generalised.

So far we have considered designs in which all the factors are *random*. In other words, the levels of each factor that actually feature in each study are considered to be random samples from some larger population (or universe, or domain). The persons themselves, whether primary pupils, examination candidates or workplace trainees, are considered to be a representative sample of all such individuals, whether past, present or future. The test questions, the creative products and the practical tasks are similarly considered to be merely samples of all such elements that could have been used and evaluated in the assessment process. And the markers, product evaluators or workplace assessors, are assumed simply to represent all such individuals, holding no special interest in themselves. It is on the basis of these assumptions that we can generalise the analysis results of such a design, in particular when attempting to identify how to reduce measurement error in a future assessment application.

But there are situations in which the assumption that a factor is sampled is not appropriate. For example, suppose that in a performance art course, such as ballet or gymnastics, there are a certain number of movements that all students must learn to master, and the number of these is relatively small. A certifying examination could feature all of these movements. Similarly, consider a workplace assessment, in which assessors evaluate the performance of all trainees on each of a number of required job-related tasks. If all the important dance or gymnastic movements, and all the relevant job-related tasks, were assessed for every aspiring qualification holder, then the factor 'tasks' (or 'movements') would be considered *fixed*. Figure 4.4 illustrates this 'mixed model' design.

Figure 4.4 Variance partition diagram for the mixed model crossed design c × a × t, where c (candidates) and a (assessors) are random factors and t (tasks) is fixed

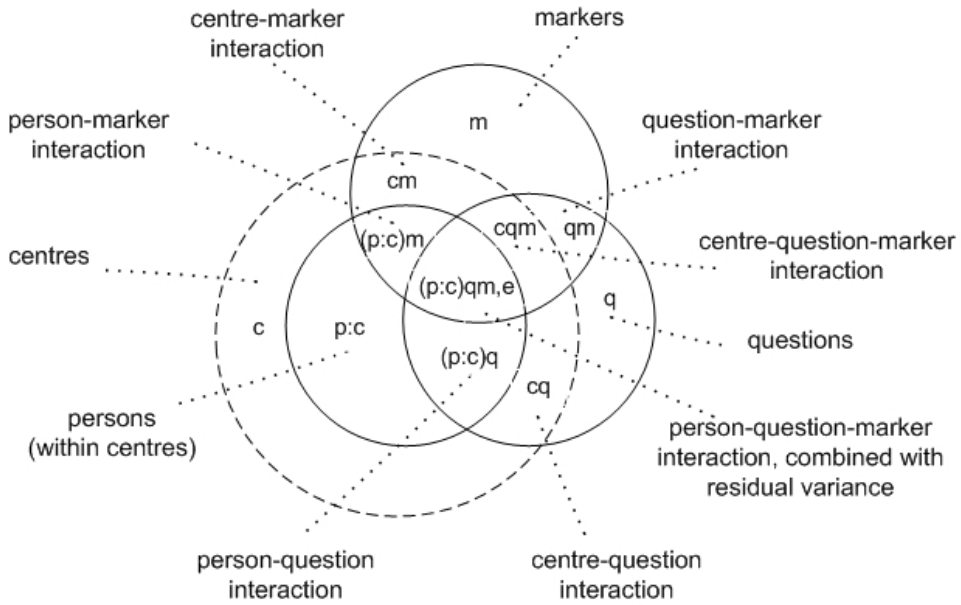


In this situation tasks will not be a source of sampling variance, and in consequence the between-tasks variance would contribute to absolute measurement error.

4.2 Taking account of demographics and other grouping characteristics

In real-life testing situations we rarely have completely crossed measurement designs. Students, for example, are *nested* within schools (and often also in classes within schools), in the sense that any particular student attends one and only one school. The students in a particular school might share certain characteristics to do with that school – its learning environment, the curriculum taught, and so on. Schools are in turn nested within local authorities and often draw their students from particular socio-economic areas. Students are further nested within gender, ethnic group and socio-economic status. Sometimes shared group characteristics can themselves have an effect on test results, and if they are not identified they, too, become hidden facets in the design, which might or might not be inflating the apparent effects of one or more of the variables that *are* identified. Let us illustrate this by adding examination centre to the pqm model discussed in section 4.1, examination centres being schools, colleges or businesses. Figure 4.5 illustrates the new picture of variance partition, with p:c indicating that persons are nested within centres.

Figure 4.5 Variance partition diagram for the crossed design (p:c)qm, where p (persons), q (questions) and m(markers) are random factors and c (centres) is fixed



Since persons are nested within centres, centre by default joins persons as a contributor to 'valid' variance. Universe score variance in this design is therefore given by the composite expression: $\sigma^2_{p:c} + \sigma^2_c$. The expression for relative error variance, σ^2_δ, will include contributions from all the interactions involving persons and/or centre:

$$[4.5] \quad \sigma^2_{cm}/n_m + \sigma^2_{cq}/n_q + \sigma^2_{cqm}/n_q n_m + \sigma^2_{(p:c)m}/n_m + \sigma^2_{(p:c)q}/n_q + \sigma^2_{(p:c)qm,e}/n_q n_m$$

with the main effect sources of variance as additional contributors to the absolute error variance, σ^2_Δ:

$$[4.6] \quad \sigma^2_\delta + \sigma^2_m/n_m + \sigma^2_q/n_q + \sigma^2_{qm}/n_q n_m$$

The SEM is as usual given by the square root of the error variance, while the generalizability coefficients are as usual found by substituting variance component estimates and factor sample sizes as appropriate into the respective ratio expressions. Optimization proceeds as before, this time by substituting alternative numbers of questions *and* markers into the error expressions and coefficient formulae.

We could elaborate this type of design by substituting or adding different nesting variables for persons, such as gender or socio-economic group. The principle for analysis would be the same. All the factors in which persons are nested will contribute to 'valid' variance, all interaction effects between the factors within the person nesting hierarchy and those outside it that involve a random facet would contribute to relative and to absolute error variance, and all main effect factors outside of the person nesting hierarchy would be additional contributors to absolute error variance (as long as these could be assumed to be random or finite random factors, i.e. factors whose levels in the data set represent samples from infinite or finite populations). Should any of the factors outside of the person nesting hierarchy be fixed, i.e. should the levels in the dataset be the only ones that exist or the only ones of interest, then the situation changes.

Factors other than persons might also usefully be recognised as being nested. Test questions, for example, might be nested within curriculum objectives, component papers, formats, and so on. Markers, too, might be nested, perhaps within subject area, length of teaching experience, gender or age group. Any of these nesting variables could potentially have an influence on assessment results. If they do, and if they are not recognised in the variance analysis, then their effects will simply be wrapped up in those of factors and factor interactions that *are* analysed. It isn't always possible, however, to incorporate all the potentially influencing variables in a reliability study, even if they could all be identified, observed and somehow categorised. This is because at some point the size of the available dataset will become insufficient to provide robust estimates for all the main and interaction effect variance components.

A recent attempt to explore the likely influence of markers on the results of the French baccalauréat, and which involves a nesting variable for markers, is described by Suchaut (2008). This study, albeit extremely small scale and relatively informal, is a rare example in its field, no research into the reliability of the baccalauréat having apparently been published since the 1930s. As a CPD exercise in assessment, Suchaut organised a marking study in which around 30 economics and social science teachers in two French academies (Dijon and Besançon) independently marked the essays (same topic) of three students from their own academy who had passed the baccalauréat in this subject field in 2007 and 2008. The teachers who participated were generally representative of those practising teachers who might have served as baccalauréat markers in the live examination, though they had not necessarily done so. Suchaut's principal aim was to investigate the extent to which markers might agree or not in their ratings of students' work in a baccalauréat examination. Secondary interests were to see whether any differences between markers were systematic (relative severity/leniency), and what effect their specific subject background and their academy might have on their marking behaviour.

No assessment professional would be surprised to learn that there was an extremely large variation in the marks given to any one script by the 30+ markers who marked

it, on a 0-20 mark scale. Equally unsurprising is the fact that there was a large variation in the average scores of the six students. More surprising might be the fact that there was no significant difference between the markers in overall severity/leniency terms. There was, though, strong evidence of marker-script interaction, different markers having different views about the relative merits of different scripts. When invited to articulate their opinions about the scripts, the markers showed evidence of often very diverse views about the same piece of work. What one marker might consider a superficial treatment of the topic, with poor analysis and poor use of support references, another could consider to be a well-supported and convincing argument.

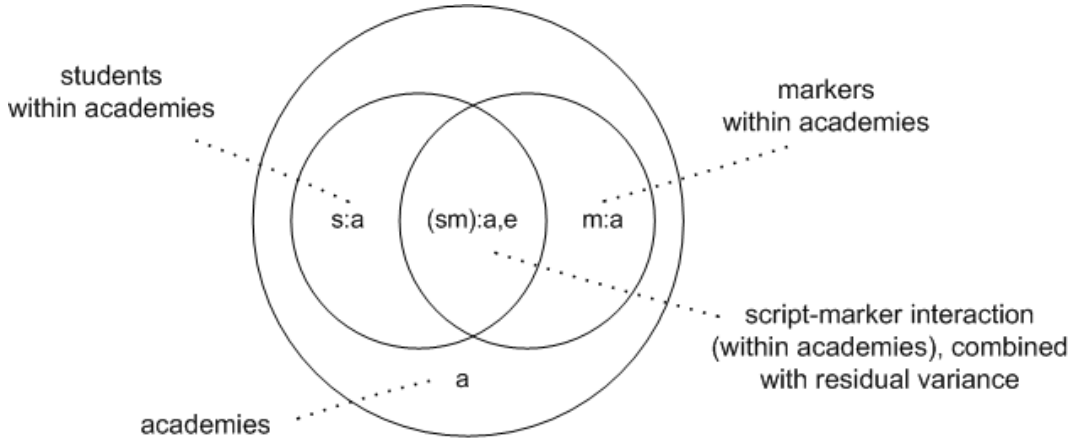
Suchaut's analysis of the resulting set of marks included comparisons of mean scores, a simple crossed analysis of variance for each academy separately, and average inter-marker correlations, all involving significance testing. Yet all of the relevant findings could have been identified more effectively through a single comprehensive G study.

Notwithstanding the extremely small number of scripts considered in the study, we can for the purpose of illustration make the assumption that these are a random sample from the larger pool of baccalauréat scripts, which they were not in practice. We also make the assumption that the markers are a random sample of all potential or actual baccalauréat economics and social science markers – they were clearly not, although they might have represented the marker population quite well. We have a choice with the academies. For maximum generalization we can also treat these two academies as a random sample of all such academies, and indeed of all schools who submit students for that particular baccalauréat examination. On the other hand, if our interest were in those particular two academies and no others then 'academies' should be considered a fixed factor. Let us assume for present purposes that all three variables, scripts, markers and academies are random factors.

The design would be (sm):a, where s, m and a represent, respectively, scripts, markers and academies. Markers and scripts are crossed factors, since every marker in each academy marked every script from that academy, and both markers and scripts are nested within academies, each marker and each script belonging to one only of the two academies (the colon in the design notation conventionally indicates factor nesting). The appropriate variance partition diagram for this situation is shown in Figure 4.6.

We can choose to focus on any sector in Figure 4.6, and calculate relative or absolute G coefficients and SEMs. For example, if we were interested in estimating how well – how reliably – academies had been measured, relatively or absolutely, the 'valid' variance that we would be interested in would be the between-academies variance. Given the assumption that both scripts and markers were randomly sampled (which in practice, of course, they were not) then the contributors to error variance here would be the between-scripts variance, the between-markers variance, and the confounded script-marker interaction variance. If, on the other hand, we were interested in how reliably the scripts were differentiated on the basis of the evaluations of all the markers who marked them, then the contributors to relative measurement error will be inter-marker variation and script-marker interaction.

Figure 4.6 Variance partition diagram for the nested design (sm):a, where s (scripts), m (markers) and a (academies) are random factors

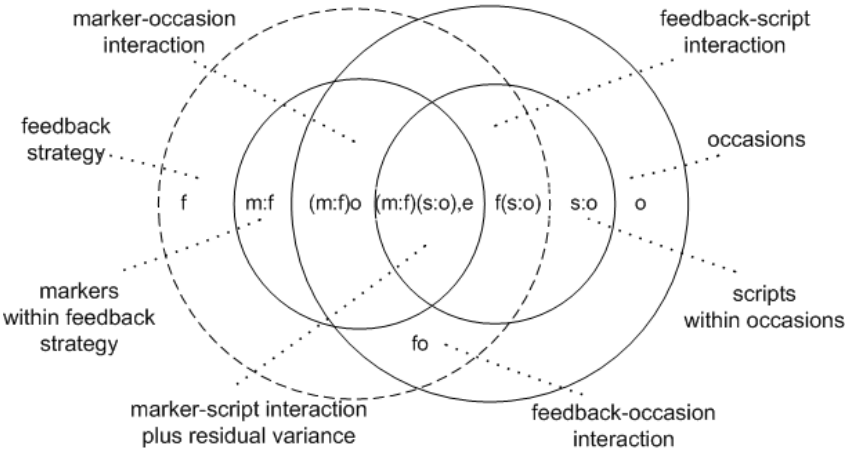


In fact the G coefficient for script differentiation is 0.97, for all 32 markers within an academy marking each script. The SEM is 0.42, giving a 95% confidence interval around a script score of just over ±0.8 marks on the 20-mark scale. Using the "what if?" approach, we find that if a single marker were to have marked any particular script then the SEM would be estimated to be 2.36, giving a margin of error of ±4.68 for each script score. Dual marking would result in a margin of error of ±1.67 and triple marking an error margin of ±1.36

This study is clearly too small-scale in terms of script numbers to be useful in practical terms. Also, not only were the markers not randomly sampled from within the national pool of potential baccalauréat markers, they had not undergone any kind of marker standardisation beforehand. Nevertheless, the results are interesting. Should this type of design be implemented on a larger scale, with more scripts and fewer multiple markers per script, and random sampling employed, then the results could be very informative.

Another study into marker reliability that could have benefited from the variance component approach is described in Sykes et al. (2009). This study aimed to evaluate the relative effect on marker consistency of four different forms of senior examiner feedback to markers. A total of 33 markers were assigned to one or other of four marker groups, or, rather, to one or other of four different forms of feedback ('treatments' in experimental design terminology). A total of 100 scripts, representing the responses of 100 candidates to one question in a GCSE English Higher Tier examination, were distributed among five batches of 20 scripts, the batches being designed to reflect similar ranges of response quality. All the markers marked all the scripts, working through the batches in the same order on consecutive days. But the groups of markers were given different amounts and kinds of feedback after marking one or more of the batches. The study design can be symbolised as (m:f) × (s:o), where m and f represent, respectively, markers and feedback strategies, s and o represent, respectively, scripts and occasions of marking (batches). Markers are nested within feedback strategies, scripts are nested within occasions, and the two nesting hierarchies are crossed with one another, since every marker in every feedback group marked every script on every occasion. The variance partition diagram is shown in Figure 4.7.

Figure 4.7 Variance partition diagram for the mixed model nested design (m:f) × (s:o), where m (markers), scripts (s) and occasions (o) are random factors and f (feedback strategy) is fixed



The data that were analysed comprised differences between markers' allocated marks and 'reference' marks, these latter being the marks that had been previously allocated to the scripts by a senior examiner (the markers, naturally, were unaware of these reference marks). As in Suchaut's case, an analysis of variance was carried out, once again accompanied by significance testing of the various main effects and interactions. But the opportunity was not taken to analyse the data set as a G study. Had it been, then G coefficients might have been used as effect size indicators for the various effects of interest to the researchers (occasions, feedback strategy by occasions, markers by occasions), and possible optimization strategies could have been explored to help design a more effective study for the future.

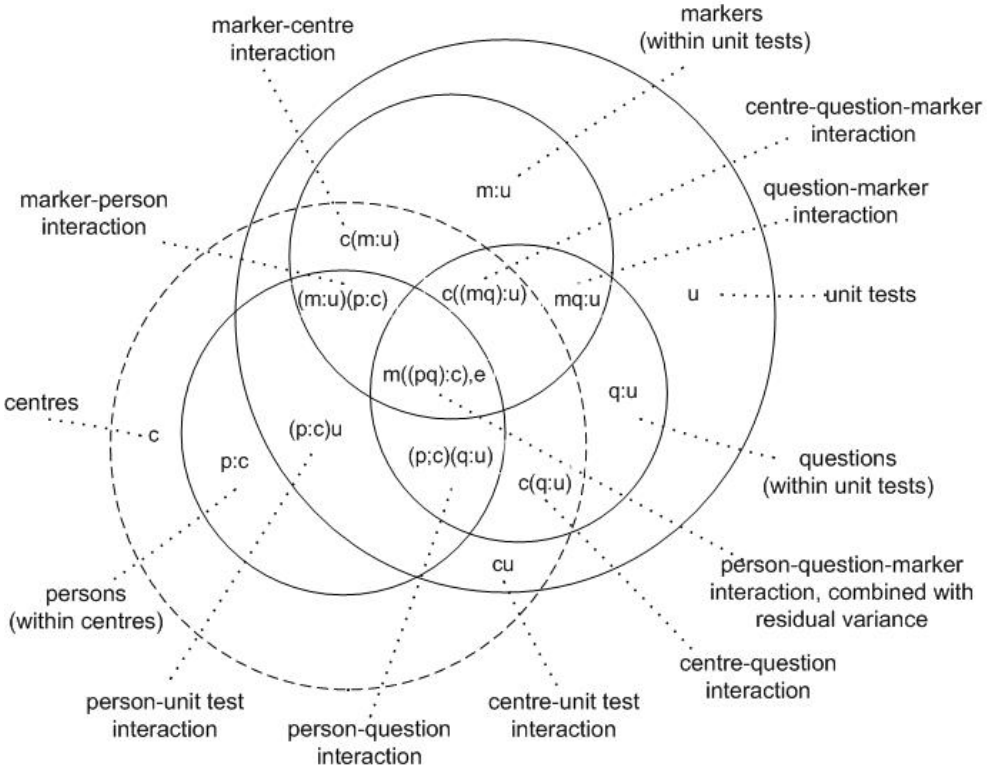
4.3 Assessment reliability at the level of whole examinations

So far we have been discussing assessment reliability with reference to a single test, where the test is of greater or lesser length in terms of constituent test questions. In some cases the same general variance analysis procedure described earlier in this chapter and in Chapter 3 can be applied to a design that incorporates the separate components of an entire examination as simply another factor. This would be possible, for example, where examination components resemble each other in form and length, and are considered to have equal weight in the whole examination. The examinations of some unit-based qualifications take this form. Test questions would be nested within unit tests, as might markers. Candidates would be nested within centres (and gender, socioeconomic group, and so on). Figure 4.8 provides flavour of modelling possibilities (this is Figure 4.5 with the addition of the extra factor of unit tests).

In the design in Figure 4.8, the universe score variance would be a combination of person (candidate) variance plus centre variance (as a nesting variable for candidates). The measurement error variance would be the combination of variance components (adjusted by dividing by the respective factor sample sizes) associated with markers and questions, the interaction between markers and questions, and all interactions involving markers, questions, centres and persons. The SEM for absolute person measurement would as usual be the square root of the measurement error variance,

with the measurement error variance itself being a function of main effects, other than persons and centres, and interaction effects involving persons and centres, each contribution suitably reduced by the relevant factor sample size(s). The generalizability study results could be used further, to estimate the effect on composite score precision of changing the number of unit tests and/or changing the number of test questions within each test.

Figure 4.8 Variance partition diagram for the mixed model nested design (p:c) × (mq):u, where p (persons), m (markers), u (unit tests) and q (questions) are random factors, and c (centres) is fixed



Often, though, the different components in an examination take very different forms, with different but related assessment aims. Thus, in a physics GCSE examination there might be an objective test, a structured question paper, a formal practical examination, and possibly also a course work element. The components might be given equal weight when brought together to produce the composite final examination mark. Or they might be differentially weighted according to some given criterion. One particular strategy that has been shown to maximise apparent score reliability for the whole examination is to give the highest weight to the most 'reliable' component, which will usually be the objective test, since this will typically have the largest number of test questions. Giving this component too much weight, however, can decrease the validity of the observed composite (the total examination mark) as a measure of the target composite (Kane & Case, 2000). In other words, the increased reliability could be spurious, and bought at the expense of assessment validity.

Where examination components differ in importance and nature (including metric) then a univariate generalizability analysis would be inappropriate, and even non-feasible. A multivariate generalizability analysis would be applicable in such cases (see, for instance Brennan, 2001b, and Raykov & Marcoulides, 2008, 2010), or some

other multivariate approach. He (2009) offers a useful comprehensive overview of issues, methodologies and applications. For an interesting recent and relevant application, in which the effects on examination reliability of assigning different weights to multiple-choice and free-response components in biology and world history examinations were explored, as well as the effects of different numbers of questions within each component, see Powers and Brennan (2009).

An aspect of whole-examination reliability that is heavily researched in the UK is grade comparability. This can be over-time comparability, between-board comparability and even between-subject comparability. Interest in over-time and between-subject comparability is not unique to the UK, but between-board comparability is. The reason for this is the presence in the qualifications market of more than one examining board offering examinations of the same general type, at the same level and in the same subjects, inviting the assumption that the respective qualifications are interchangeable. Here, too, assessment reliability, as outcome replication over different conditions of measurement, is relevant.

The question at issue is the extent to which the grading outcomes for examination candidates would vary depending on the board whose examination they had taken. The potential factors that could be hypothesised to influence these outcomes will include the examination components that comprise the different boards' examinations, the mark schemes used to evaluate candidates' performances on these components, the weighting strategies used when aggregating component outcomes, and the procedures used to determine the cut scores that lead to performance grading. It would be difficult and costly, even if possible in practice, to design a G study in which the effects on measurement error of all these factors might be simultaneously quantified. In particular, a random sample of candidates from each relevant board would be required to sit the 'parallel' examinations of the other boards. Recognising the impossibility of this requirement, a variety of different investigative approaches have been trialled and used over past decades, most proving disappointingly limited from an interpretational point of view (see Newton, Baird, Goldstein, Patrick & Tymms, 2007, for a comprehensive review). The variance analysis approach has been shown to have some potential for meaningful application in this area (Cohen & Johnson, 1982; Johnson & Cohen, 1983, 1984; Johnson, 2007).

5 Reliability estimation and reporting: the way forward?

5.1 The need to report on reliability

It is generally accepted in the UK and elsewhere that awarding bodies and other testing agencies should report information on the reliability of their assessments, whether the outcomes of this assessment take the form of test scores, mastery decisions or criterion-referenced grades. Cronbach, Linn, Brennan and Haertel (1997), for example, had this to say:

> When a public opinion poll reports what percentage of respondents favor each rival candidate, it also reports a margin of error. Agencies responsible for an educational assessment system should similarly make clear how much uncertainty is associated with any score or summary, particularly with any report released to the public or to public representatives. (Cronbach et al., 1997, p.1)

In similar vein the American 'Joint Standards' suggest that there is a duty to communicate reliability information to the public (AERA/NCME/APA, 1999, Standard 2.1, p. 31). Others would agree, whilst at the same time evincing concern about the effects that such transparency might have on public trust in the assessment system and its outcomes. Newton (2005a, b), for instance, writes about the obligations on the part of awarding bodies to communicate with the public about reliability, and offers interesting and perceptive accounts of the arguments for and against releasing reliability information for public consumption in the UK.

Hutchison and Benton (2009), too, express the view that awarding bodies should provide information about the reliability of public examinations. They emphasise the importance of transparency in reliability reporting, which they claim would:

> …offer all involved the possibility of weighing up the benefits of a more reliable assessment system against the costs in terms of time that could be spent on teaching and learning, imposition on the young people concerned, and, in fact, cost, so that decisions could be taken on the best available evidence" (Hutchison & Benton, 2009, p.34).

Ofqual has responded to the general call for more transparency in this area, by launching a 3-strand reliability programme (Boyle, Opposs and Kinsella, 2009; Opposs, 2009):

Strand 1 – generating evidence on reliability
Strand 2 – interpreting and communicating evidence on reliability
Strand 3 – exploring public understanding of reliability and developing Ofqual policy on reliability.

This report is one in a series of related research reports that together represent the first outcome of Strand 2.

Initial relatively informal explorations of the public understanding of examination validity and reliability, carried out under Strand 3, have confirmed that members of the public, albeit particularly interested members of the public at this stage – pupils, teachers, examiners, employers – can to some degree distinguish between 'inherent'

or 'inevitable' error in measurement and 'avoidable' error (Ipsos MORI, 2009; Boyle, Opposs & Kinsella, 2009). They tend to be intolerant of avoidable error, a prime example of which would be marker differences or inconsistency, as also, apparently, are members of the public in France (publication of the small-scale study on likely marker error in the baccalauréat examination described in Chapter 4 – Suchaut (2009) – caused a stir in the French press).

Interestingly, the evidence is that members of the public are relatively tolerant about what they perceive as 'inherent' errors, when perhaps they would be less so if they did indeed understand more than they do about the business of examining. In particular, there is recognition that tests and examinations cannot cover every detail in the curriculum, and that some topic selection must take place. Where testing space severely constrains examination content in this sense, it seems to be accepted that some candidates will likely benefit when their revised topic "comes up" and others will suffer when their topics do not. And yet measurement error that arises from this source is not 'inherent' and unavoidable. It could be avoided through use of longer tests, with the additional costs and testing time demands that would be associated with this (see Hutchison and Benton, 2009, p.34, for relevant discussion), and/or curriculum reduction.

If the public is to be educated about technical issues in assessment, and if reliability information is to be routinely published alongside examination results, then we need to decide which form of reliability measure would be the most appropriate one to use. There are basically two choices: a variance ratio (reliability coefficient), and a standard error of measurement. [We do not include here any kind of misclassification index, given their very limited value as valid comments on assessment reliability.]

It might be true that a standard error of measurement is likely to be better understood in the public domain than a reliability coefficient, especially when converted into a margin of error, from which the familiar 95% confidence interval can be calculated around test scores (given the assumption that errors are Normally distributed). Skurnick and Nuttall (1968) thought so, making a plea for SEM reporting over 40 years ago in the context of external examinations in the UK. In the US this view is also held by some. When reflecting on his lifetime's work in this field, Cronbach observed that:

> I am convinced that the standard error of measurement … is the most important single piece of information to report regarding an instrument, and not a coefficient. The standard error, which is a report of the uncertainty associated with each score, is easily understood not only by professional test interpreters but also by educators and other persons unschooled in statistical theory, and also to lay persons to whom scores are reported. (Cronbach & Shavelson, 2004, p.413).

But is the notion of a reliability coefficient really so difficult to comprehend, giving us as it does an idea in proportional terms of how much 'noise' there is in the results of an assessment process? Perhaps not. Coefficients still have a role to play in the reporting of assessment reliability in some contexts.

What remains is to decide how to produce the most valid and meaningful indicators of reliability for reporting purposes, and this depends on how we define 'reliability' itself.

5.2 Reliability as replicability

If we conceive of assessment reliability as the degree to which we can expect an assessment outcome for a single examination candidate to be repeated, or replicated, under different conditions of measurement (and evaluation), particularly when multiple sources of measurement error are at play, then the investigative approach to use is variance analysis. Variance analysis based on appropriate designs, whether prospective or retrospective, provides us with a tool to estimate score reliability as outcomes replication over some given 'universe of generalisation', by quantifying the contributions to measurement error of identifiable and 'approachable' factors, which are necessarily sampled in the assessment process. Variance analysis equally allows us to identify practical strategies for reducing measurement error, some of which will be more feasible and more cost-effective to implement than others.

The advantage of G-theory, the variance analysis approach, over calculation of classical reliability indices lies in its ability to handle several error-relevant factors simultaneously. In other words, the variance analysis approach is not restricted to the practice within classical True Score Theory of addressing "one variable at a time" – most typically questions or markers – to the exclusion of others. Analysis models reflect reality more comprehensively, and so enable the estimation of more valid reliability estimates.

For example, speaking in the context of the 'simulated patient' form of assessment that is now commonplace in medical education worldwide, Swanson, Clauser and Case (1999) comment that:

> Because multiple sources of measurement error are present, the statistical methods that classical test theory provides for investigating the reproducibility of scores (e.g., separate indices of internal consistency and inter-rater agreement) are not adequate. Instead, use of the conceptual and analytic tools of generalizability theory (G-theory) are [sic] mandatory… (Swanson, Clauser & Case, 1999, pp.75-76)

In this medical training context the principal sources of error variance are medical cases, standardised patients (the trained actors portraying patients with particular medical conditions), raters, rating schemes and occasions of testing, along with interactions among these and between these and the students being assessed. In the UK test and examinations context the variance sources will similarly include assessment questions and tasks, markers and raters, mark schemes and rating protocols, occasions of testing, modes of assessment, and so on.

The main argument of this report is that there is no longer any reason to appeal to classical reliability indicators as evidence of assessment reliability. These are all limited in scope and, in consequence, limited in usefulness. And they are anyway subsumed within the more general sampling-based variance analysis approach that is G-theory. These comments apply particularly to the well-known alpha coefficient, which is still the most used reliability indicator (Hogan, Benjamin & Brezinski, 2000), even when much of the time its use is inappropriate. The alpha coefficient is an internal consistency coefficient that was conceived for application in a norm referencing world. But the world has changed, and so has the importance of coefficient alpha. It is no longer "the gold standard for measuring reliability", as still

apparently assumed by many practitioners and as explicitly claimed by at least one commercial company whose testing services are routinely used by some vocational awarding bodies in the UK. Cronbach himself considered the alpha coefficient to have become virtually obsolete, given the modern-day predominance of absolute measurement for criterion-referenced cut score applications: "I no longer regard the alpha formula as the most appropriate way to examine most data", he pronounced (Cronbach & Shavelson, 2004, p.403).

Unless the aim of our assessment is to use arbitrary cut scores to separate pre-chosen proportions of candidates in terms of their relative attainment (the 50 highest scoring individuals, the top 10% of attainers, next 15%, etc), then we should be using some form of reliability index that is appropriate to absolute measurement. The phi coefficient, which we cover in Chapters 3 and 4, is an obvious candidate

It has to be said, though, that variance ratios can have reduced value in the context of mastery testing, where it is quite legitimate to accept that at times there will be limited, if any, between-person variance: all can pass or all can fail, with most candidates expected to be at or near the criterion cut-score. This will be the case in much of vocational skills assessment. Here the standard error of measurement, as an indicator of 'noise' in the assessment results, is particularly valuable.

Most measurement methodologies enable the calculation of standard errors of measurement. But only within the framework of G-theory is the estimated measurement error *generalisable*, both beyond the actual set of questions or tasks used in the particular assessment application, and also over other factors that potentially contribute to measurement error, including markers. While a global generalisability analysis provides an estimate of *average* measurement error for all candidates, analysis of data subsets will furnish estimates – conditional SEMs – specific to particular candidate subgroups, or even to individual candidates (for details see Feldt, Steffen and Gupta, 1985; Feldt and Brennan, 1989; Raju, Price, Oshima and Nering, 2007).

The principal advantages of G-theory are that:

1. multiple contributions to measurement error can be simultaneously quantified
2. the reliability indices that result from a generalisability study, whether coefficients or standard errors of measurement, are generalisable by virtue of the random sampling theory underpinning them
3. complex sampling designs, including domain sampling and multiple matrix sampling, are directly supported by G-theory as a consequence of its ANOVA, experimental design heritage
4. the quantified variance component information can be used to identify ways to reduce measurement error in future testing applications (within the constraints of practicality and budget), and
5. the magnitude of measurement error for different candidate subgroups, including candidates at different points on the measurement scale, can also be estimated and optimised.

5.3 Researching reliability

Whether or not reliability information is eventually routinely released to the public, and whatever form this information, if released, might take, there has to be an obligation on awarding bodies to quality assure their tests and examinations, so that their assessment outcomes are as fair as possible to the highest number of candidates. This means doing more than carrying out 'end-of-line' quality checks, which can provide only limited information about how to improve the assessment process, if necessary, in the future, and whose results are rarely available in time to rectify any injustices done to that year's examination candidates. What quality assurance essentially involves is identification of the kinds of factors that contribute to measurement error, quantification of the relative contributions of those factors to measurement error, and subsequent use of that information not only to determine current levels of reliability but also to guide the redesign of assessment tools, practices and procedures before they are used again.

Among awarding bodies in the UK the greatest research investment in recent years has been in the area of marker reliability, which is a legitimate operational concern. Very much less attention seems to have been devoted to investigating the effect on score reliability of the number and nature of the test questions that are put before examination candidates, or to exploring the effect of alternative mark schemes. Yet examining agencies have long been urged to begin comprehensively investigating the separate and joint influences on test and examination scores of sources of measurement error other than, though still including, markers, in particular by using G-theory (see, for example, Wood, 1991, p.144).

It will typically not be possible to cover all relevant error-contributing variables in a single reliability study, or even within an entire programme of research. Indeed, there will certainly be occasions when variance analysis is not an option – for example when an assessment process cannot adequately be modelled. In other cases an analysis might not be worth doing, because its results will not be amenable to useful interpretation. But where a generalisability analysis is feasible and potentially meaningful, then the higher the number of 'facets' that can simultaneously be investigated – so that relative contributions to measurement error can be quantified – the more valid and interpretable the resulting reliability estimation is likely to be.

When appropriate research is carried out, and reliability findings become available, the next step should be to evaluate these findings in terms of implications for future examining practice, including the design of examination components and examination procedures. The research might suggest that tests should be longer, or that several short tests should replace a single long one. Multiple marking might be indicated as essential, or confirmed as being unnecessary. Standard setting procedures might be modified, and so on.

There will be times, though, when the results of an optimization study cannot in practice be fully implemented, given the logistic, temporal and financial constraints under which assessment agencies typically operate. Multiple marking is a case in point. It is recognised that, despite the serious attention that is given to marker standardisation and equally to workplace assessor standardisation in the vocational sector, intermarker and intramarker variation (at question level) will inevitably still be

present to some extent in the live situation. Indeed, it is also widely accepted that there might be 'drift' in markers' overall marking standards and/or in their marking consistency as marking progresses through time. If any one candidate's work is evaluated by one and only one marker, then clearly injustices are likely to occur more often than if more than one marker marked the work. But the time within which thousands of scripts in numerous different qualifications must be processed, combined with a limited, albeit large, pool of suitable markers from which to draw, has meant that even dual marking has not apparently been the norm in external examining in the UK:

> Awarding bodies struggle to recruit enough examiners to mark scripts once, let alone twice. Double marking of all examination papers is not a feasible option. (Meadows & Billington, 2005, p.58)

The rapidly developing introduction into academic examinations of online marking could help to alleviate this particular problem, as will the employment of individuals other than subject specialists to mark examination components that do not demand expert input in the marking process.

Given the particular context of workplace assessment, it is easy to imagine that multiple evaluation of task performances will not be any easier for vocational qualifications, and here online marking will seldom be an option. Yet being observed at work by more than one assessor could be quite intimidating, even if this might be feasible in practice. Video recording could unnerve some candidates, is expensive to implement in a standardised way, and the resulting recordings are time consuming to process. In addition, there are unique factors here that might impinge on the validity and the reliability of assessor judgements in the workplace. These are the social and professional relationships that must exist between qualification candidates and their assessors, when often the assessor is the candidate's workplace instructor (see Wolf, 1995, for an insightful discussion on this particular issue). Nevertheless, however well internal and external verifiers do their job of assessment regulation, without formal studies designed to explore and quantify assessment reliability in this area the actual quality of workplace assessment will remain unknown (see Greatorex & Shannon, 2003, and Greatorex, 2005, for some first tentative explorations). This is a challenge that merits attention.

The impact on measurement error of the particular selection of examination questions put in front of examinees, or the particular work-relevant tasks that workplace trainees are asked to perform, is another aspect calling for investigation. What difference would it make to a candidate's outcome if the topic of an essay question, or the nature of the workplace task, were to be changed? What would be the outcome if different pieces of achievement evidence had been put together to produce an evidence portfolio? There will be limits to the degree to which such questions can be answered in practice, but some attention deserves to be given to them nonetheless.

5.4 A final note on the validity risks for reliability

When factors that are known or suspected to impact on test scores are ignored in an analysis, i.e. when the underpinning measurement model only partially reflects reality, then the resulting reliability measures will be of reduced value, and could even be

misleading. We are speaking here of 'hidden' factors. Examples include question effects and interactions between markers and questions in traditional marker reliability studies, candidate demographics and their interactions with conditions of assessment, and nesting factors for questions (such as skill areas or formats).

We have described in this report how limited all the classical reliability indicators are in this sense. Each explores just one single error-inducing variable: 'occasions' in the test-retest index, 'tests' in indices involving alternative test forms, 'questions' (items or tasks) in the case of split-half and internal consistency coefficients, including alpha. Outcome agreement measures developed for use with nominal scales, of which kappa (Cohen, 1960) seems to be the most used, suffer from the same problem.

If we accept reliability as repeatability, or replicability, then we need to be clear what the 'universe of generalisation' is that this repeatability is referenced against. Some of the important factors that in principle comprise the appropriate universe of generalisation might be able to be fully represented in the assessment process, and hence in the resulting dataset, and would therefore not be measurement error contributors. Different item formats, for instance, could all be included, as could every major topic within a chemistry curriculum, and all the tasks that a windscreen repairer would need to be able to carry out in that occupation. Other factors would essentially be sampled, explicitly or implicitly. Principal among these, as we have mentioned numerous times throughout the report, will be markers and questions. Where factors that could potentially contribute to measurement error are excluded from the model underpinning a reliability calculation then the actual universe of generalisation could be reduced, and this should be recognised.

On a final note, it is important to remember that in G-theory 'random factors' are assumed to be factors whose levels are randomly sampled, so that the levels present in the dataset (items, markers, essay topics, and so on) can be assumed to represent their respective populations or domains, however these are defined. Where the 'randomness' of the factor sampling is not guaranteed, for example when senior examiners manually build an examination paper with newly developed questions, then the degree to which a reliability coefficient might be generalised will be in doubt.

References

Ackerman, T.A. (1991). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

AERA/NCME/APA (1999). *Standards for educational and psychological testing*. The 'Joint Standards' of the American Educational Research Association, National Council on Measurement in Education and American Psychological Association. Washington, D.C.: American Psychological Association.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Baker, K.H. (1939). Item validity by the analysis of variance: an outline of method. *Psychological Record, 3*, 242-248.

Bechger, T., Béguin, A., Maris, G. & Verstralen, H. (2003). Using classical test theory in conjunction with item response theory. *Applied Psychological Measurement, 27,* 319-334.

Béguin, A.A. & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychmetrika 66*, 541-562.

Bock, R.D., Brennan, R.L. & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364-375.

Boyle, A., Opposs, D. & Kinsella, A. (2009). No news is good news? Talking to the public about the reliability of assessment. Paper presented at the 35th annual conference of the International Association for Educational Assessment (IAEA), Brisbane.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters*, Issue 4, 22-28.

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

Brennan, R. L. (2000) Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339-353.

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295-317.

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

Brennan, R.L. (2003). *Coefficients and Indices in Generalizability Theory.* CASMA Research Report: No.1. Iowa City: University of Iowa Center for Advanced Studies in Measurement and Assessment. (Available on http://www.education.uiowa.edu/casma)

Brennan, R.L. & Kane, M.T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277-289.

Brennan, R.L. & Kane, M.T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika*, *42*, 609-625.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322.

Brown, W. (1911). The Essentials of Mental Measurement. Cambridge: Cambridge University Press.

Burch, V.C., Norman, G.R., Schmidt, H.G., & van der Vleuten, C.P.M. (2008). Are specialist certification examinations a reliable measure of physician competence? *Advances in Health Sciences Education, 13,* 521-533.

Burt, C. (1947). Factor analysis and analysis of variance. *British Journal of Psychology*, *1*, 3-26.

Cardinet, J., Johnson, S. & Pini, G. (2009). *Applying Generalizability Theory using EduG*. New York: Routledge.

Cardinet, J. & Tourneur Y. (1985). *Assurer la mesure.* Berne: Peter Lang.

Cardinet, J., Tourneur, Y. & Allal, L. (1976). The symmetry of generalizability theory: applications to educational measurement. *Journal of Educational Measurement*, *13* (2), 119-135.

Cardinet, J., Tourneur, Y. & Allal, L. (1981, 1982). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, *18*, 183-204, and Errata *19*, 331-332.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, L., & Johnson, S. (1982). The generalizability of cross-moderation. *British Educational Research Journal, 8*, 147-158.

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

Cronbach, L J. (1947). Test "reliability": its meaning and determination. *Psychometrika, 12,* 1-16.

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. & Hartmann, W. (1954). A note on negative reliabilities. *Educational and Psychological Measurement, 14*, 324-346.

Cronbach, L.J., Linn, R.L., Brennan, R.L. & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373-399.

Cronbach, L.J., Rajaratnam, N. & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, *16*, 137-163.

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.

Doran, H.C. (2005). The information function for the one-parameter logistic model: is it reliability? *Educational and Psychological Measurement*, 65, 665-675.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics, 3*, 1-21.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. New York: Psychology Press.

Evers, A., Van Vliet-Mulder, J.C., Resing, W.C.M., Starren, J.C.M.G., Van Alphen de Veer, R.J. & Van Boxtel, H. (2002). COTAN testboek voor het onderwijs [COTAN test book for the educational field]. Amsterdam: Boom.

Feldt, L.S. & Brennan, R. L. (1989). Reliability. In Linn, R. L. (ed), *Educational measurement*, 105-146. New York: American Council on Education/Macmillan.

Feldt, L.S. Steffen, M. & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9*, 351-361.

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

Govaerts, M.J.B., van der Vleuten, C.P.M. and Schuwirth, L.W.T. (2002). Optimising the Reproducibility of a Performance-Based Assessment Test in Midwifery Education. *Advances in Health Sciences Education,* 7, 133–145.

Goldman, B.A., Mitchel, D.E. & Egelson, P.E. (1977). Directory of unpublished experimental mental measures (volume 7). Washington, DC: American Psychological Association.

Goldstein, H. (2003). *Multilevel Statistical Models*. (Third edition). London: Arnold.

Greatorex, J. (2005). Assessing the Evidence: different types of NVQ evidence and their impact on reliability and fairness. *Journal of Vocational Education and Training, 57*, 149-164.

Greatorex, J. & Shannon, M. (2003). How can NVQ assessors' judgements be standardised? Paper presented at the annual conference of the British Educational Research Association, Edinburgh, September.

Gulliksen, H.O. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L.A. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

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

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. Reprinted as Chapter 4 in Thompson (2003).

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577-601.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160.

Hutchison, D. & Benton, T. (2009). *Parallel universes and parallel measures: estimating the reliability of test results*. Coventry: Office of the Examinations and Qualifications Regulator (Ofqual).

Huysamen, G.K. (2006). Coefficient Alpha: unnecessarily ambiguous; unduly ubiquitous. *South African Journal of Industrial Psychology, 32*, 34-40.

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

Jackson, R.W.B. (1939). Reliability of mental tests. *British Journal of Psychology, General Section, 29*, 267-287.

Johnson, S. (1989). *National Assessment: The APU Science Approach*. London: HMSO.

Johnson, S. (2007). Commentary on judgemental methods. In Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (eds), *Techniques for monitoring the comparability of examination standards,* 295-300. London: Qualifications and Curriculum Authority.

Johnson, S. (2008). The versatility of Generalizability Theory as a tool for exploring and controlling measurement error. In M. Behrens (ed.), Special Issue: Méthodologies de la mesure. Hommage à Jean Cardinet. *Mesure et Evaluation en Education*, *31*, 55-73.

Johnson, S. & Bell, J. (1985). Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement, 22*, 107-119.

Johnson, S. & Cohen, L. (1983). *Investigating grade comparability through cross moderation*. London: Schools Council.
Johnson, S. & Cohen, L. (1984). Cross-moderation: a useful comparative technique? *British Educational Research Journal, 10*, 89-97.
Kane, M.T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.
Kane, M.T. & Case, S.M. (2000). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221-240.
Kelley, T.L. (1923). *Statistical method*. New York: Macmillan.
Kuder, G. & Richardson, M. (1937). The theory of estimation of test reliability. *Psychometrika*, 2, 151-160.
Linacre, J.M. (1994). *Many-Facet Rasch Measurement*, 2nd ed. Chicago: MESA Press.
Linacre, J.M. & Wright, B.D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 2*, 486-512.
Lord, F.M. (1955). Estimating test reliability. *Educational and psychological measurement, 15*, 325-336.
Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Lord, F.M. & Novick, M.R (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
Marcoulides, G.A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics, 18*, 197-206.
Marcoulides, G.A. (1995). Designing measurement studies under budget constraints. Controlling error of measurement and power. *Educational and Psychological Measurement, 55*, 423-428.
Marcoulides, G.A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement, 57*, 808-812.
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.
Mehrens, W.A. & Lehmann, I.J. (1984). *Measurement and Evaluation in Education and Psychology*. New York: Holt, Rinehart and Winston.
Messick, S. (1989). Validity. In Linn, R. L. (ed.), *Educational Measurement.* Washington, DC: American Council on Education/Macmillan Series on Higher Education.
Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38, 114-117.
Murphy, D.J., Bruce, D.A., Mercer, S.W. & Eva, K.W. (2009). The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Advances in Health Sciences Education, 14*, 219-232.
Newton, P. (2005a). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31*, 419-442.
Newton, P. (2005b). Threats to the professional understanding of measurement error. *Journal of Education Policy, 20*, 457-483.
Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (eds) (2007). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Nunnally, J.C. (1967). *Psychometric Theory*. New York: McGraw-Hill.
Opposs, D. (2009). *Ofqual's reliability of results programme*. Paper presented at the Chartered Institute of Educational Assessors' Third National Assessment Conference, London, May. Available online at: http://www.ciea.org.uk/upload/conference_2009/presentations/seminar%203.ppt.
Pearson, K. (1896). Mathematical contributions to the theory of evolution: III Regression, heredity and panmixia. *Philosophical Transactions, A, 187*, 253-318.
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)
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.
Raykov, T. & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis.* New York: Taylor & Francis.
Raykov, T. & Marcoulides, G. A. (2010, in press). *Psychometric Theory.* New York: Routledge.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
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.
Schmitt, N. (1996). Uses and abuses of Coefficient Alpha. *Psychological Assessment, 8*, 350-353.
Searle, S.R., Casella. G. & McCulloch, C.E. (2006). *Variance Components*. (Second edition). Hoboken: Wiley.
Shavelson, R.J., Baxter, G.P. & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.
Shavelson, R. & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
Shavelson, R. & Webb, N. (2006). Generalizability theory. In Green, J.L., Camilli, G. & Elmore, P.B. (eds), *Handbook of Complementary Methods in Education Research*, Chapter 18. London: Lawrence Erlbaum Associates.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.
Skurnik, L.S. & Nuttall, D.L. (1968) Describing the reliability of examinations. *The Statistician*, *18*, 119-128.
Snijders, A.B. & Bosker, R.J. (1999). *Multilevel analysis*. London: Sage.
Solomon, D. J., & Ferenchick, G. (2004). Sources of measurement error in an ECG examination: Implications for performance-based assessments. *Advances in Health Sciences Education, 9*, 283-290.
Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology,* 15, 72-101.
Spearman, C. (1904b). General intelligence objectively determined and measured. *American Journal of Psychology,* 15, 201-292.
Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology,* 18, 161-169.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology,* 3, 271-295.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Stanley, J.C. (1971). Reliability. In R. L. Thorndike (ed.), *Educational Measurement*, (Second edition), Chapter 13, pp 356-442. Washington DC: American Council on Education.

Stigler, S.M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science, 4(2)*, 73-86.

Suchaut, B. (2008). *La loterie des notes au bac. Un réexamen de l'arbitraire de la notation des élèves.* IREDU Working paper. Dijon: Institute for Research in the Sociology and Economics of Education.

Swanson, D.B., Clauser, B.E. & Case, S.M. (1999). Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating and security. *Advances in Health Sciences Education, 4*, 67-106.

Sykes, E., Novakovic, N., Greatorex, J., Bell, J., Nadas, R. & Gill, T. (2009). How effective is fast and automated feedback to examiners in tackling the size of marking errors? *Research Matters,* Issue 8, 8-14.

Thomson, B. (2003). *Score reliability.* Thousands Oaks, California: Sage Publications.

Van der Linden, W.J. & Hambleton, R.K. (1997). *Handbook of modern item response theory.* New York: Springer.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

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

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D. & Stone, M.H. (1979). Best test design. Rasch measurement. Chicago: MESA Press.

First published by The Office of Qualifications and Examinations Regulation in 2010.