Mesure et évaluation en éducation

The versatility of generalizability theory as a tool for exploring and controlling measurement error

Sandra Johnson

Volume 31, Number 2, 2008

URI: https://id.erudit.org/iderudit/1025007ar
DOI: https://doi.org/10.7202/1025007ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Johnson, S. (2008). The versatility of generalizability theory as a tool for exploring and controlling measurement error. *Mesure et évaluation en éducation*, *31*(2), 55–73. https://doi.org/10.7202/1025007ar

Article abstract

Measurement error arises from many sources in educational assessment. It is important to estimate the importance of this error, and, if appropriate, to seek ways to reduce it. Generalizability theory represents a powerful tool in this sense, allowing identifiable error contributions to be separately quantified, and measurement error to be estimated and even predicted in response to possible changes in the measurement procedure. The paper offers examples of generalizability analysis of numeracy attainment data deriving from the Scottish Survey of Achievement, with the aim of illustrating the versatility of the methodology for error estimation and prediction in this type of sample-based programme.

The versatility of generalizability theory as a tool for exploring and controlling measurement error

Sandra Johnson
Assessment Europe, Scotland

Key words: Generalizability theory, national assessment, attainment surveys, numeracy assessment, test reliability, domain sampling, matrix sampling

Measurement error arises from many sources in educational assessment. It is important to estimate the importance of this error, and, if appropriate, to seek ways to reduce it. Generalizability theory represents a powerful tool in this sense, allowing identifiable error contributions to be separately quantified, and measurement error to be estimated and even predicted in response to possible changes in the measurement procedure. The paper offers examples of generalizability analysis of numeracy attainment data deriving from the Scottish Survey of Achievement, with the aim of illustrating the versatility of the methodology for error estimation and prediction in this type of sample-based programme.

Mots clés: Théorie de la généralisabilité, évaluation nationale, enquêtes pour le suivi des acquis, évaluation des notions de calcul, fiabilité d'un test, échantillonnage par domaine, échantillonnage matriciel

L'erreur de mesure découle de nombreuses sources dans l'évaluation en éducation. Il est important d'estimer l'ampleur de cette erreur et, si c'est le cas, de chercher les moyens de la réduire. La théorie de la généralisabilité représente dans ce sens un outil puissant qui permet d'identifier les sources de l'erreur et de les quantifier séparément, d'estimer l'erreur de mesure et même de prédire la réponse à d'éventuels changements dans la procédure de mesure. Cet article offre des exemples d'application de l'analyse de la généralisabilité sur des données pour le suivi des acquis des notions de calcul, données provenant de l'enquête écossaise sur la réussite, dans le but d'illustrer la polyvalence de la méthodologie d'estimation et de prévision de l'erreur dans ce programme d'évaluation basé sur un échantillonage.

Author's note – Assessment Europe is Technical Adviser to the Scottish Government's Assessment for Learning Programme, and in particular to the Scottish Survey of Achievement. All correspondence should be addressed to [Sandra.Johnson@Assessment-Europe.com].

Palavras-chave: Teoria da generalizabilidade, avaliação nacional, estudos das aprendizagens adquiridas, avaliação das noções de cálculo, fiabilidade de um teste, amostragem por domínio, amostragem matricial

Na avaliação em educação, o erro de medida decorre de numerosas fontes. É importante calcular a amplitude deste erro e, se for o caso, procurar meios para a reduzir. A teoria da generalizabilidade representa, neste sentido, um instrumento poderoso que permite identificar as fontes do erro e quantificá-las separadamente, calcular o erro de medida e mesmo prever a resposta a eventuais mudanças nos procedimentos de medida. Este artigo fornece exemplos de aplicação da análise da generalizabilidade sobre os dados das aprendizagens adquiridas de noções de cálculo, dados provenientes do Estudo Escocês sobre o Sucesso, com o objectivo de ilustrar a versatilidade da metodologia de cálculo e previsão do erro neste programa de avaliação baseado numa amostra.

Generalizability theory

Generalizability theory (G-theory) uses information about quantified contributions to measurement error to estimate, and even to predict, measurement precision (Brennan, 1992, 2001; Cardinet & Tourneur, 1985; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991; Thompson, 2003). Given this, G-theory clearly has a useful role to play in any sample-based assessment exercise, including experimental design applications and sample-based attainment surveys.

A generalizability study, or G-study, first requires the identification of observable factors that can be assumed or suspected to affect the dependent variable, which might be a reading test score, an ICT skills score, an attitude scale score, or whatever. Relevant factors might include curriculum covered, teaching effectiveness, adequacy of revision, test length, marking accuracy, gender, actual ability or attitude (the real focus of the measurement), mood on the day of assessment, etc. An appropriate G-study is then designed with which to investigate the contributions to score variation of those factors which can in practice be identified and whose effects can be observed. For example, test length and gender can clearly be observed. Curriculum coverage can also be observed to some extent. But teaching effectiveness would be difficult to address, and mood on the day of assessment even more so. Relative influences on the dependent variable are quantified in the form of estimated variance components, using classical analysis of variance (ANOVA) or some other appropriate methodology. The component information is in turn used to calculate

measurement errors and "generalizability coefficients" (G-coefficients) – ratios of linear combinations of adjusted components that indicate the technical reliability of target measurements. Finally, a "what if?" facility, or decision-study (D-study), permits predictions of reliability and measurement error when those features of the current design that can be changed *are* changed in a future application – such as increasing the numbers of items in a test – so that we might optimize measurement by maximally increasing precision.

Pupil attainment surveys are an interesting context for the application of G-theory. The principal purpose of such surveys is typically to estimate the ability, achievement or attainment of some *population* of pupils in some area of the curriculum – perhaps pupils' numeracy attainment at the end of primary schooling. In this context, sampled pupils merely represent their population – they are not of special interest as individuals. Since they *are* a sample, pupils, and their schools and classes, contribute to error variance, as do the assessment tasks administered to them (that is if one wishes to generalize beyond the specific set used in the survey). Clearly, in such applications it is essential to offer some indication of the size of estimation error for reported population attainment estimates. It is also important to explore ways of reducing estimation error in future surveys, within financial, logistic and other constraints. Attainment surveys, by virtue of their scale and design, are also able to furnish additional information of value in other assessment contexts, including item banking. Here, too, reliability indices and measurement error estimates are useful.

In this paper, numeracy data from the Scottish Survey of Achievement are used to illustrate the versatility of G-theory through examples of application. The G-theory software package EduG[1] was used to analyse the data.

The Scottish Survey of Achievement

The Scottish Survey of Achievement (SSA) is a sample-based survey programme that constitutes one element in a coherent framework of assessment in Scotland (Hayward, 2007). Launched in 2005, the SSA continues Scotland's long history of sample-based attainment surveys, in that it has evolved from the long-running Assessment of Achievement Programme (AAP), that began life in the mid-1980s (Condie, Robertson & Napuk, 2003). The difference between the two programmes is that the SSA was initially designed to offer attainment reporting at the level of local authorities as well as nationally, whereas the AAP offered only national attainment reporting.

There are always multiple objectives for any large-scale survey of the type considered here. The *principal* survey objective in both 2005 and 2006 was to assess pupils' attainment levels in numeracy and reading at national and local authority level for four key stages in schooling : P3 (7-8 year olds), P5 (9-10 year olds), P7 (11-12 year olds, end of primary schooling) and S2 (13-14 year olds). Pupil attainment was reported with reference to the progressive level framework in the 5-14 National Curriculum Guidelines (*e.g.*, SOED, 1991 ; SOEID, 1999 for mathematics), as percentages of pupils attaining particular levels (A to F). While there are measurement challenges associated with this form of reporting, the benefits are that all stakeholders – policy makers, teachers, pupils and parents – are able fully to understand the reported findings, given their shared understanding of the 5-14 level framework. At each stage, pupils were assessed at three of the six progression levels. These are the levels most appropriate for the stage concerned : Levels A, B and C for P3 ; Levels B, C and D for P5 ; Levels C, D and E for P7 ; and Levels D, E and F for S2.

Domain sampling was employed to select numeracy items from the 5-14 National Assessment Bank[2] for survey use, supplemented through new item development where necessary. Almost 600 items in total were administered in each of the two surveys, that is 80-90 items per level ; just under half the items used in 2005 were incidentally common to the two surveys. In 2005, the items were distributed among 40 different numeracy test booklets, 10 per stage ; in 2006 there were 48 booklets in total, 12 per stage. An important constraint had to be met when creating the booklets. This was that no pupil should be faced with a test booklet entirely composed of items from one level only, since the level concerned might be far below their capabilities or dauntingly above them. Every test booklet therefore contained items at three consecutive levels, with items presented in a randomized order throughout the booklet. And every booklet came in two versions, the second simply reversing the item presentation order of the first (this was to minimise the possible impact of fatigue effects on particular items, since item-level statistics were to be an important secondary outcome). Multiple matrix sampling was used to allocate items to test booklets, and test booklets to pupils. Each pupil took two different test booklets, and in consequence attempted three different single-level tests, where a single-level test comprised those items at the same level across the booklet pair.

After inevitable pupil losses from the intended samples, around 28,000 pupils in almost 1500 mainstream schools across Scotland were tested in each of the surveys, this is approximately 12% of the pupil population at each stage. Around half the pupils took numeracy tests and half reading tests. Pupils were classified into attainment bands at each level on the basis of their test performances, specifically in terms of the proportions of items they successfully answered at that level in their two booklets[3]. Pupils correctly answering 80% or more of the test items at a level were deemed to have shown "very good" attainment at that level. Pupils answering 65% or more of the items at a level correctly but fewer than 80% were classified as having "well-established" skills at the level (*i.e.*, "secutity"). Pupils correctly answering 50% or more of the items at a level correctly but not as many as 65% were deemed to have made a "good start" at the level. Weighted proportions of pupils in each band at each level were computed, and margins of error estimated using the jackknife technique. The results are presented in detail in the survey reports (Scottish Government, 2006, 2007).

Measurement reliability

The SSA, like other national and international survey programmes, is not concerned primarily with assessing the abilities or skills of individual pupils. But individual pupils were inevitably assessed in the process of producing population attainment estimates, and their achievement results can therefore be used to explore the reliability of the test scores that underpin those population estimates. They can also be used to explore the reliability of item summary scores – facility values in this case – that might be useful in other types of application.

Such information is especially useful in the Scottish context for two reasons. The 5-14 National Assessment Bank furnishes assessment materials for use in the SSA, and reciprocally benefits post-survey from new item input from the SSA. It also serves as a resource for teachers, who can download National Assessments, single-level tests, for use in their classrooms to confirm their own judgements of their pupils' attainment levels. There is no mechanism in place at present for teachers to feedback the item-level results that could be used to explore the "fitness for purpose" of the National Assessments. This means that the SSA has a role to play here, in investigating the reliability of the downloadable tests for the assessment of individual pupils.

Secondly, even though item statistics are not used when creating test booklets for the SSA or for the National Assessments programme, this situation might change in the future. In addition, other item-banking applications might be launched at some point for which such statistics might be useful or even necessary. The SSA provides item performance statistics on the basis of large nationally representative pupil samples. But how reliable are the resulting statistics? The answer is very reliable, as the first example G-theory application presented in this paper confirms.

The simplest possible ANOVA model, that involves pupils on the one hand and items on the other, is the crossed design P x I, where P represents pupils and I represents items, and all pupils attempt all items. Figure 1 is a typical variance partition diagram for this design, following Cronbach et al. (1972). There are three potential sources of score variation: between-pupil variance (the pupil effect in ANOVA terminology), the between-item variance, (the item effect), and the pupil by item interaction. This interaction is confounded with residual variance, given that we typically have one single observation per interaction cell, since any one pupil attempts any particular test item once only: again following Cronbach et al.(1972), the residual variance element in this confounded variance component is represented by the letter *e*.

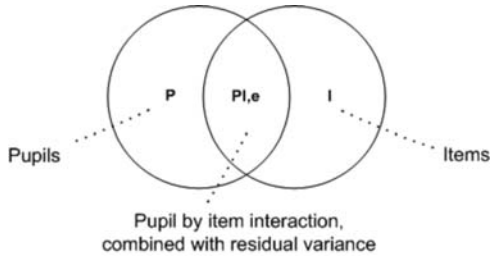

Figure 1. ***Variance partition diagram for the crossed design P x I***

This is an appropriate point at which to introduce three important G-theory terms: relative measurement, absolute measurement, and criterion-referenced measurement. In a context of relative measurement the aim is to locate objects, in this case pupils or items, relative to one another on some scale of measurement, with maximum reliability; a typical application would be selection, or normative grading. Absolute measurement is concerned rather with the precision with which the objects are individually measured, irrespective of where any other objects might be on the measurement scale;

domain-referenced mean scores are relevant here. In criterion-referenced measurement, the aim is to make mastery decisions with maximum confidence when applying criterion cut-off scores to test results, such as the 65% "well-established" score in the SSA.

In G-theory terminology, the factors of ANOVA become "facets", a term change apparently introduced to avoid confusion with the factors of factor analysis (Cronbach et al., 1972, p. 2 footnote). "Generalization facets" contribute to measurement error. Which facets are generalization facets will depend on which facet is the object of measurement (*i.e.*, the "differentiation facet"), as well as on the sampling status of the facet itself (fixed, finite, random). Jean Cardinet and his colleagues were instrumental in promoting the property of ANOVA "symmetry" to expand the application of G-theory beyond the pupil assessment that was its original focus (see Cardinet & Tourneur, 1985; Cardinet, Tourneur & Allal, 1976, 1981, 1982). Symmetry simply indicates that any of the factors/facets in a design can in principle be identified as the object of measurement. Even in this simplest crossed design there are two possible choices for a differentiation facet: pupils and items.

In either case, the only contribution to error variance in a context of relative measurement will be the pupil by item interaction effect (combined, as it is, with all residual variance). If the aim is to rank pupils the general level of difficulty of the items used will simply result in the pupil rank-order moving further up or further down the measurement scale; between-item variance will be irrelevant. Similarly, the general ability of the pupil group is irrelevant in terms of item ranking. In the pupil ranking context the observed "universe score" variance is the between-pupil variance, represented in ANOVA terms by the estimated variance component for pupils. The total observed score variance is the sum of the between-pupil variance and the pupil-by-item interaction variance, after the latter is divided by the number of items involved (*i.e.*, by the number of pupil by item observations that have contributed to the pupil's total or average test score). The ratio of observed universe score to total observed score variance provides an indicator of relative measurement reliability: rho-squared, G-theory's relative and original generalizability coefficient. For this simple application, the relative G-coefficient is exactly equivalent to Cronbach's alpha coefficient (Cronbach, 1951), the most familiar indicator of score reliability, and the most frequently, if often inappropriately, reported (Hogan, Benjamin & Brezinski, 2000). A corresponding G-coefficient can be calculated for the item ranking application, this time by dividing the interaction variance by the number of *pupils* used to produce the estimated item

facilities before combining with the between-item variance to produce the total observed score variance. Conventionally, coefficient values of 0.8 or higher (on a 0-1 scale) are considered to indicate adequate degrees of score reliability, but obviously the higher the value the better.

If we are interested in locating a pupil or an item as precisely as possible on a measurement scale, then mere rank ordering is not sufficient. Difficulty variation in the items used will become relevant in the first case, between-item variance becoming a second contributor to measurement error. Variation in the ability of the pupils to whom the items are administered will become relevant in the second case, and between-pupil variance will join the pupil-by-item interaction effect as an error contributor. In other words, we need now also to take into account the additional measurement error that will have arisen from the item or the pupil sampling. The appropriate G-coefficient for absolute measurement is the phi-coefficient. When cut-off scores are to be applied the phi coefficient is replaced by the phi(lambda) coefficient, a criterion-referenced variant (Brennan, 2001). The closer that lambda, the criterion cut-off score, is to the test's mean score the lower will be the phi(lambda) coefficient.

None of these coefficients is relevant in a situation where the aim is to estimate the average attainment of a group, or population, of pupils with respect to some given item domain (such as numeracy). The parameters of interest here are the overall mean score (here the average pupil-item score), and its associated standard error of measurement. There are no differentiation facets.

Numeracy test score reliability in the SSA

While the SSA is not concerned with ranking pupils or items, it might nevertheless be of interest to note that for the majority of the over 100 single-level numeracy tests administered in one or other of the surveys of 2005 and 2006, the coefficient of relative measurement was above 0.8, with several in the 0.9s (Scottish Government, 2006, Technical Annex, Section C; 2007, Annex II). For an illustrative example, let us take just one of the numeracy tests used at P7 in 2005: an 18-item Level D test whose items were distributed across booklets N21 and N22. Table 1 provides the ANOVA table for this test.

Table 1
Analysis of variance for the design P x I
(SSA 2005, 18-item Level D numeracy test spanning booklets 21 and 22 at P7; 628 pupils)

Source of variance	Sum of squares	df	Mean square	Estimated variance component
Pupils (P)	431.30467	627	0.68789	0.03003
Items (I)	272.93064	17	16.05474	0.02533
Pupils by Items plus residual variance (PI,e)	1571.51380	10659	0.14744	0.14744
Total	2275.74912	11303		

Using the data in Table 1, we calculate the G-coefficient for relative pupil measurement, rho-squared, as:

0.03003/(0.03003 + 0.14744/18) = 0.79 (identical with coefficient alpha)

The corresponding G-coefficient for relative *item* measurement is:

0.02533/(0.02533 + 0.14744/628) = 0.99

The other Level D tests produced similar results, showing adequate score reliability for relative pupil measurement and extremely high score reliability for relative item measurement (this latter to be expected, given the very high pupil numbers involved).

To estimate what the value of the relative G-coefficient for pupil ranking might be, should a test longer than 18 items be used, we simply substitute 18 in the expression above with any feasible alternative number (a 200-item test would, for example, not be practically feasible to administer). While there are no National Assessments in numeracy, there *are* in mathematics, and at Level D they comprise 27 items. With this number, the G-coefficient (assuming primary mathematics and primary numeracy to be roughly interchangeable) would increase to around 0.85, an acceptable level; 30 items would increase the coefficient value to 0.86, 35 items to 0.88 and 40 items to 0.89.

What can we say about the reliability of *absolute* measurement, rather than relative measurement? For the test under review, the absolute G-coefficient for pupil measurement will be slightly lower than the relative coefficient, given that the item variance will now also be a contributor to the error variance:

0.03003/(0.03003 + 0.02533/18 + 0.14744/18) = 0.76

while the corresponding coefficient for item measurement is unchanged in value:

0.02533/(0.02533 + 0.03003/628 + 0.14744/628) = 0.99

Again, we can use "what if?" analysis to estimate the values of the coefficients should we change the number of items in the test or the number of pupils taking the test.

As to the phi(lambda) coefficient, the mean test score for this Level D test at this stage (P7) was 13, close to the criterion cut-off score of 12 (see Figure 2), and the value of the phi(lambda) coefficient is 0.78. The Level C test in the same pair of test booklets had a mean score of 14.5, while the Level E test had a mean score of 8.1. As might be predicted, therefore, both of these single-level tests had higher phi(lambda) coefficients than the Level D test, at 0.85 and 0.88, respectively.

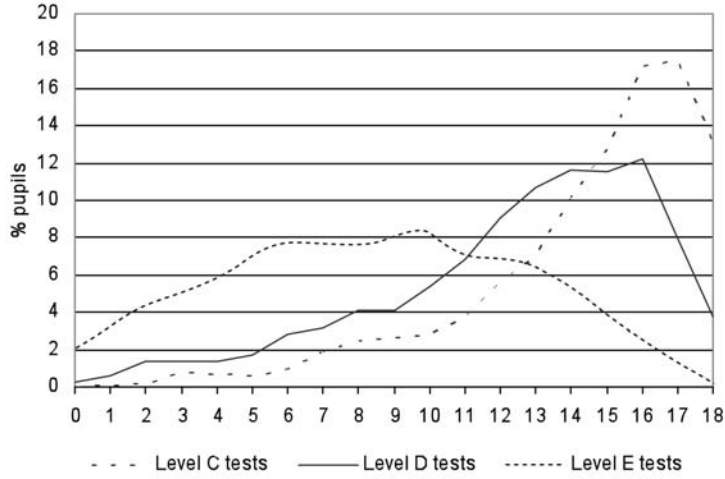

Figure 2. ***Aggregated score distributions for SSA 2005 numeracy tests at Levels C, D and E at P7*** (the 65% criterion cut-off score for 'well-established skills' is equivalent to 12 items)

Subgroup comparisons and population estimates in the SSA

The simple model shown in Figure 1 can be extended, by introducing other relevant factors. In particular, pupils are generally grouped in some way in reality – they share some common characteristics, such as gender, socioeconomic background, class, school and age-group. Items, too, can sometimes be grouped, for example in terms of topic, item format, parent test. Such grouping is termed "nesting," and is conventionally indicated by bracketing, as in P (G), or by use of a colon, as in P:G. It is recognition of nesting features in data that first led to the term "multilevel modeling". Multilevel modeling as a measurement approach is based on regression analysis, and was first developed in a context of school effectiveness research to recognise the contributions of nesting variables to the values of dependent variables, in particular to recognise the importance of class and school where pupil attainment is concerned (Snijders & Bosker, 1999).

The problem with introducing nesting into designs within a context of large-scale attainment surveys is that the attainment data are never balanced. For example, even though the intended pupil sample for a particular booklet pair in the SSA could have comprised equal numbers of boys and girls, the obtained sample would usually not. In other cases, such as socioeconomic background, the data are by nature very unbalanced. Typically, reflecting the situation in the population at large, pupil samples contain relatively low proportions of deprived pupils – defined as those living in an area that is among the 20% most deprived areas of the country, based on the Scottish Index of Multiple Deprivation (Scottish Executive, 2004). EduG, though, requires balanced data sets.

In the interests of illustrating a mixed model design involving nesting, two nesting variables for pupils feature here: gender (G) and stage (S). Stage can be included as a pupil nesting factor because every Level D test administered at P7 was also administered at S2 (pupil records were eliminated at random to produce a balanced data set comprising 250 pupils per gender per stage; record elimination could have been avoided by processing the unbalanced data set through an analysis of variance procedure provided by other software and then submitting sums of squares to EduG). Figure 3 is the variance partition diagram for the resulting design (P:GS) x I. Note that both gender and stage are fixed factors, since both genders feature in the data set, as do both stages of interest – there is no intention to generalize results to any other stages. This fixed status is indicated by dashed lines in the variance partition diagram.

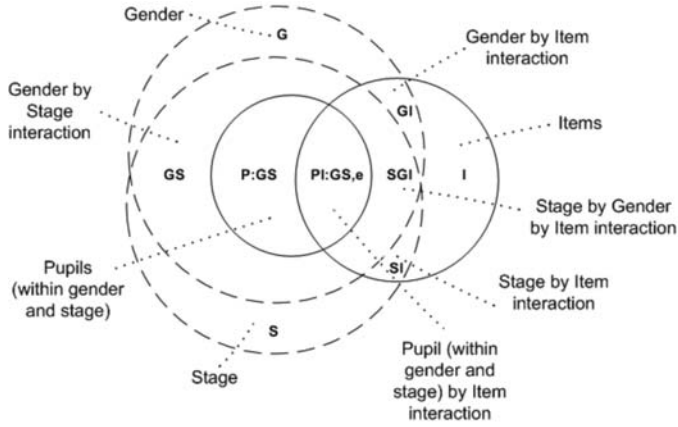

Figure 3. ***Variance partition diagram for the mixed model design P(GS) x I, with both Gender and Stage fixed factors***

Following the principle of symmetry, any one of the sources of variance shown in Figure 3 could, in principle, be the differentiation facet of interest. We might, for example, be interested, as before, in exploring the reliability of item ranking or of pupil ranking. Or we might want to calculate the phi(lambda) coefficient for this particular set of items attempted by this particular set of pupils. Or we might, rather, be concerned to see how well the assessment has managed to differentiate gender or stage. Let us focus on stage. Table 2 is the ANOVA table for this design for the same Level D test as before, while Table 3 provides the G-study results.

Only three of the sources of variance in Figure 3 actually contribute to error variance for relative measurement of stage attainment. These contributors are SI, the stage by items interaction effect, P(GS), the pupils within stage and gender effect, and (P:GS x I,e), the pupil by item interaction effect (confounded as usual with residual variance). As gender is a fixed factor, the variance sources GS and SGI do not contribute to relative measurement error, nor indeed to absolute measurement error. Dividing the relevant variance component estimates by the appropriate numbers of observations (*e.g.*, dividing 0.00058, the variance component for SI, by the number of stages, 2, and the number of items, 18) and summing the resulting adjusted variance contributions will provide the estimate of relative measurement error (see Table 3).

The relative G-coefficient is 0.91. The coefficient of absolute measurement, however, is just 0.49, the dramatic reduction explained by the contribution to absolute measurement error of the items effect (*i.e.*, of between-item variation

corresponding to the variance component, 0.01850, divided by the number of items, 18). These coefficients are technically omega coefficients, a consequence of the fact that the differentiation facet, stages, is a fixed factor (EduG calculates coefficients using whichever formula is appropriate, given the status of the differentiation facet).

Table 2

Analysis of variance for the design P:GS x I

(SSA 2005, 18-item Level D numeracy test spanning booklets 21 and 22 at P7, and booklets 31 and 32 at S2; 250 pupils per gender per stage, gender and stage fixed factors)

Source of variance	Sum of squares	df	Mean square	Estimated variance component
Stage (S)	20.60450	1	20.60450	0.00108
Gender (G)	1.07339	1	1.07339	0.00000
Items (I)	316.86850	17	18.63932	0.01850
Gender by Stage (GS)	0.28006	1	0.28006	-0.00003
Stage by Items (SI)	12.18850	17	0.71697	0.00058
Gender by Items (GI)	9.15561	17	0.53857	0.00040
Stage by Gender by Items (SGI)	4.88094	17	0.28711	0.00015
Pupils within stage and gender (P:GS)	669.47400	996	0.67216	0.02964
Pupils within Stage and Gender, by Items (PI:GS,e)	2347.96200	16932	0.13867	0.13867
Total	3382.48750	17999		

While the two stages, P7 and S2, have been differentiated reliably with this one test, absolute measurement of the stage means is inadequate. Clearly, since it is between-item variance that is the main contributor to absolute measurement error, the obvious strategy to improve the precision of the measurement is to increase the number of items used, in order to reduce the contribution of the items variance to measurement error.

Table 3
G-study results for the design (P:GS) x I

Differentiation facet	Differentiation variance	Sources of error variance	Relative error variance	Absolute error variance
S	0.00108			
		I		0.00103
		SI	0.00003	0.00003
		P:GS	0.00006	0.00006
		(P:GSx I,e)	0.00002	0.00002
Total variance	0.00108		0.00011	0.00113
Standard deviation	0.03279		0.01034	0.03368

In the 2005 survey five different, and in principle equivalent, numeracy tests were administered at each relevant level at each stage (Figure 4 illustrates this), so that 90 items actually represented a level and not 18. With five tests used at Level D, the coefficient for relative stage measurement increases from 0.91 to 0.97, while the coefficient for absolute stage measurement increases from 0.49 to 0.74, a more acceptable value.

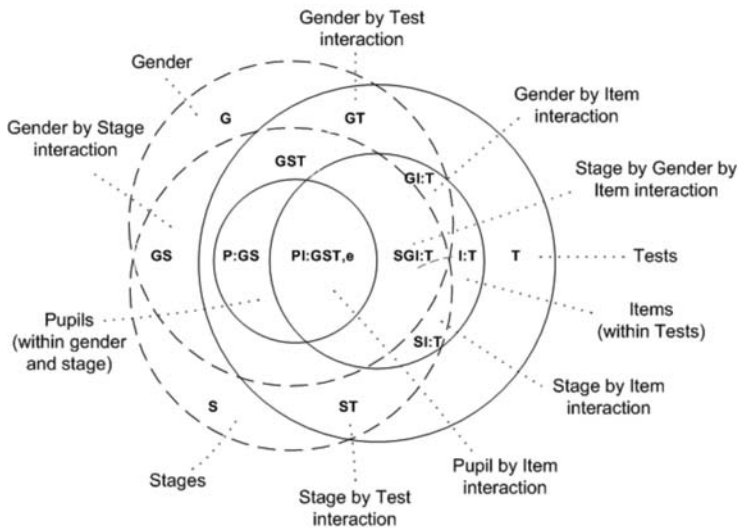

Figure 4. ***Variance partition diagram for the design (P:GST) x (I:T)*** *(SSA 2005, five 18-item Level D numeracy tests used at P7 and S2; 250 pupils per gender per stage per test, with gender and stage fixed factors)*

Stage comparisons are interesting and useful, but the principal objective of the survey was to produce population attainment estimates for each relevant 5-14 level at each stage assessed, with maximum precision. Imagine the variance partition diagram in Figure 4 with stage removed: all the remaining sources of variance, with the exception of the fixed facet gender and its interactions with other facets, will contribute to the standard error of measurement for the domain-referenced population attainment estimate at one stage – there are no differentiation facets. EduG allows us to calculate the standard error of measurement for each stage separately, using the data associated with the design in Figure 4, by requesting an observation design reduction to eliminate each stage in turn from the data set. The resulting mean scores are 70% for P7 and 76% for S2, with standard errors of measurement of 1.9 percentage points and 1.6 percentage points, respectively.

Can the measurement error be reduced? Available optimizing options are to change the numbers of items used, by changing the number or length of tests used, and/or to change the numbers of pupils tested (see Johnson & Bell, 1985; Johnson, 2003, for previous examples). Table 4 shows the results of a 'what if?' analysis for the P7 Level D data. Just three alternative facet sampling strategies are investigated here for illustration (many others are possible):

1. increasing the numbers of items per test from 18 to 20 whilst reducing the number of pupils per gender per test from 250 to 150;
2. increasing item numbers to 20 per test and the number of tests from 5 to 10, whilst reducing pupil numbers per gender per test to 100;
3. increasing item numbers to 30 per test, with five tests as before, and with 150 pupils per gender per test.

Table 4
SSA2005 D-study results for Level D numeracy assessment at P7

	G-study	Plan I	Plan II	Plan III
Gender	2	2	2	2
Number of tests	5	5	10	5
Number of items within tests	18	20	20	30
Number of pupils within gender and test	250	150	100	150
Total number of observations	45,000	30,000	40,000	45,000
Absolute error variance	0.00034	0.00032	0.00017	0.00022
Standard error of measurement	0.019	0.018	0.013	0.015
95% confidence interval in percentage points	± 3.7	± 3.5	± 2.5	± 2.9

As Table 4 shows, the number of pupils per test could have been almost halved, from 500 to 300, with little loss in measurement precision. Increasing the number of tests administered, however, would have increased precision markedly, even with a further cut in pupil numbers per test, to 200 (this number, though, would be too low for other reasons – firstly, because it would not be large enough to produce reliable item statistics, and, secondly, because it would reduce the potential for gender comparisons at the item level).

An alternative strategy could be to reduce the item variance, along with pupil by item and gender by item interaction (see Figure 5), by "culling" items appropriately, as in Item Response Theory (IRT) applications. The effect of such a strategy, however, might well be to narrow the definition of the item domain, thus reducing assessment validity, and in turn assessment value, if the nature of the domain narrowing cannot be made explicit. This would not be acceptable in this particular programme, where broad and valid curriculum coverage is considered vital to ensure assessment validity and usefulness.

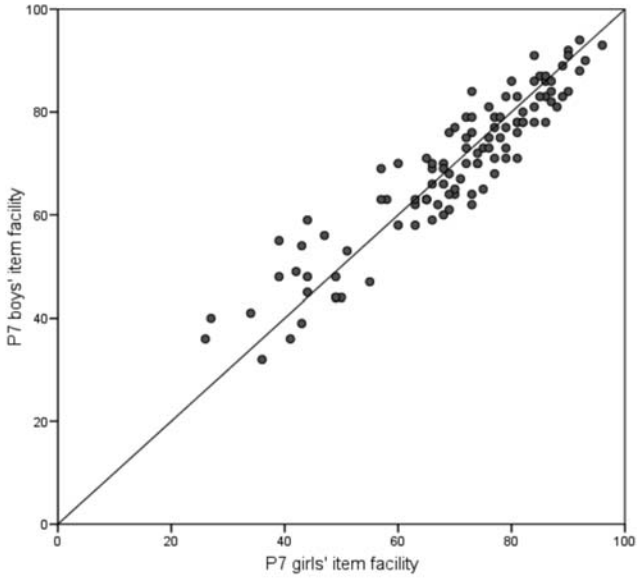

Figure 5. ***Item variance and gender by item interactions for Level D at P7***
(SSA 2005, 90 Level D items in five numeracy tests;
250 pupils per gender per item)

In conclusion

The SSA is a sample-based national monitoring programme, with multiple aims, principal among them being to provide as precise estimates as possible of the attainment levels of pupil populations in key curriculum areas. But attainment estimates that are "as precise as possible" are not necessarily the "most precise" estimates, because maximizing precision, within given constraints, might only be achieved by reducing content validity and in turn assessment value.

The programme is intended to monitor pupil attainment, and change in this attainment over time, in terms of the curriculum being taught in the schools. The domain sampling of level-validated items to create the tests used in the programme contributes to broad curriculum coverage within each survey. But broad curriculum coverage is usually associated with high inter-item variance, and with high pupil-item interaction effects. Even in a very narrowly defined aspect like addition involving 2-digit numbers, items with very different facilities can be written, and different pupils might find one or other more or less difficult than the next (think of 24 + 42, 24 + 42 + 35, 13 + 73 + 7, all perfectly valid items under the definition). It is highly likely that Item Response Theory (IRT) would not be useful in the SSA, because in order for the data to fit an IRT model, even one with three or more parameters, a proportion of the existing items, valid as they are in curriculum terms, might have to be eliminated. To explore the applicability of IRT models in this context, a 'spiral' booklet administration strategy was used in 2006 to provide appropriate linked-item data.

G-theory is useful in this kind of sample-based application, in providing a versatile methodology for investigating sources of measurement error, and quantifying their relative contributions so that strategies for reducing the error might be identified. For researcher and measurement practitioners alike EduG is a welcome tool in this respect.

NOTES

1. EduG is the result of a collaboration between research groups in Switzerland and Canada, driven by the efforts and enthusiasm of Jean Cardinet. The software, along with its supporting documentation, is downloadable free from the website [www.irdp.ch/edumetrie/logiciels.htm], in both French and English versions.
2. [http://www.aifl-na.net/]
3. The banding criteria, while in principle arbitrary, are based on the professional judgment of subject specialists. They were first used for reporting reading attainment in the 2001 AAP English Language survey, and have continued into the SSA.

REFERENCES

Brennan, R.L. (1992). *Elements of generalizability theory* (2nd edition). Iowa City : ACT Publications (First edition : 1983).

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

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*(4), 183-204, and *Errata*, *19*(4), 331-332.

Condie, R., Robertson, I.J., & Napuk, A. (2003). The assessment of achievement programme. In T.G. K. Bryce & W.M. Humes (Eds), Scottish Education (pp. 766-776).Edinburgh : Edinburgh University Press.

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.

Hayward, E.L. (2007). Curriculum, pedagogies and assessment in Scotland : the quest for social justice. 'Ah kent yir faither'. *Assessment in Education*, *14*(2), 251-268.

Hogan, T.P, Benjamin, A., & Brezinski, K.L. (2000, 2003). Reliability methods : A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Johnson, S. (2003). Une application de la théorie de la généralisabilité à la planification des enquêtes sur les acquisitions des élèves. In J. Cardinet (Ed.), Special Issue : Generalizability Theory. *Mesure et évaluation en éducation*, *26*(1-2), 37-59.

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

Scottish Executive (2004). *Scottish Index of Multiple Deprivation 2004 : Summary technical report*. Edinburgh : Scottish Executive.

Scottish Government (2006). *The 2005 Scottish Survey of Achievement (SSA) : English language and core skills.* [http://www.scotland.gov.uk/Publications/2006/06/29141936/0]

Scottish Government (2007). Scottish Survey of Achievement : 2006 Social subjects (enquiry skills) and core Skills. [www.scotland.gov.uk/Publications/2007/08/15104710/0]

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

Snijders, T.A.B, & Bosker, R.J. (1999). *Multilevel analysis : An introduction to basic and advanced multilevel modelling*. London : Sage Publications.

SOED (1991). National guidelines : *Mathematics 5-14*. Edinburgh : Scottish Office Education Department. [http://www.ltscotland.org.uk/5to14/guidelines/mathematics.asp]

SOEID (1999). *National guidelines : Mathematics 5-14 Level F.* Edinburgh : Scottish Office Education and Industry Department. [http://www.ltscotland.org.uk/5to14/guidelines/mathematics.asp]

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Ed.), *Score reliability* (pp. 43-58). Thousand Oaks, CA : Sage publications.