‘I’m pickin’ up good regressions’: the governance of generalisability analyses

Jim Crossley,1 Jean Russell,2 Brian Jolly,3 Chris Ricketts,4 Chris Roberts,5 Lambert Schuwirth6 & John Norcini7

CONTEXT Investigators applying generalisability theory to educational research and evaluation have sometimes done so poorly. The main difficulties have related to: inadequate or non-random sampling of effects, dealing with naturalistic data, and interpreting and presenting variance components.

METHODS This paper addresses these areas of difficulty, and articulates an informal consensus amongst medical educators from Europe, Australia and the USA, who are familiar with generalisability theory.

RESULTS We make the following recommendations.

- Ensure that all relevant factors are sampled, and that the sampling meets the theory’s assumption that the conditions represent a random and representative sample of the factor’s ‘universe’. Research evaluations will require large samples of each factor if they are to generalise adequately.
- Where feasible, conduct 2 separate studies (pilot and evaluation, or Generalisability and Decision studies).

- For unbalanced data, use either urGENOVA, or 1 of the procedures minimum norm quadratic unbiased estimator, (MINQUE), maximum likelihood (ML) or restricted maximum likelihood (REML) in SPSS or SAS if the data are too complex.
- State which mathematical procedure was used and the degrees of freedom (d.f.) of the effect estimates. If the procedure does not report d.f., re-analyse with type III sum of squares ANOVA (ANOVA ss III) and report these d.f.
- Describe and justify the regression model used.
- Present the raw variance components. Describe the effects that they represent in plain, non-statistical language.
- If standard error of measurement (SEM) or Reliability coefficients are presented, give the equations used to calculate them.
- Make sure that the method of reporting reliability (precision or discrimination) is appropriate to the purpose of the assessment. This will usually demand a precision indicator such as SEM. Consider a graphical presentation to combine precision and discrimination.

KEYWORDS education, medical/*standards; *research, biomedical; research design/*standards; data collection; analysis of variance.

INTRODUCTION

The series ‘The metric of medical education’ attempted ‘to enable educators to better understand the basic concepts of assessment’ and included ‘a “how to” guide on designing an educational experiment’. Since then, ever-increasing numbers of assessment studies have been submitted to Medical
Overview

What is already known on this subject

Generalisability theory provides a psychometric approach for quantifying factors that influence the reliability of assessment data and for modelling approaches to reduce measurement error.

The theory is sometimes poorly applied in educational research papers.

What this study adds

This paper describes how to appropriately sample and evaluate all relevant factors; how to deal with relatively unstructured naturalistic data such as workplace assessment data; ways to meaningfully interpret and present the variance components that a Generalisability study produces.

Suggestions for future developments

Development of more user-friendly programmes that incorporate dataset management, control card management and the actual analyses.

THEORETICAL BACKGROUND

This light theoretical section introduces the main ideas that will be used in later sections.

A perfectly reliable test produces a score which is only influenced by the construct of interest (e.g. ‘non-verbal reasoning’ in an IQ test) and is completely unaffected by the circumstances of the test (e.g. when, where and by whom it is administered). A test score like this might be called a ‘true’ score. Reliability diminishes as circumstantial or ‘error’ factors increasingly affect the test score in the real world.

Generalisability is an inductive statistical method falling within the family of regression techniques – techniques which model and quantify relationships between variables to make predictions. It builds on the mathematical approach ‘variance component analysis’, which was developed at the same time as Cronbach developed his theory. Generalisability theory assumes that there is no arbitrary variation: a test score is entirely determined by the condition of the ‘true’ construct being measured and the condition of the ‘error’ factors which influence the score. (The terms ‘factor’ and ‘effect’ are equivalent to the statistical term ‘variable’. The assessment score is the test variable. Assessor, candidate, case, etc. are explanatory variables. The term ‘condition’ is equivalent to the statistical term ‘level’ and refers to the nominal or ordinal value of a variable when a given measurement is made.) In essence, the analysis quantifies the impact that relevant factors exert on the assessment score in an assessment ‘experiment’. The relative size of each factor’s contribution to the score variance reveals valuable information: if candidates’ scores are consistent across different examin-
ers or challenges, then they are probably reliable; if different examiners give very different scores to the same candidate on the same challenge, a single score will be unreliable.

In most situations generalisability theory treats the conditions of the factors as if they were randomly sampled from an infinitely large universe of possible conditions. For example, it would treat the marks of 3 examiners on 1 written assignment as if they represented a sample \((n = 3)\) of marks from an infinite pool of examiners. Thus, one way to know the ‘true’ score would be to aggregate the assignment marks from this whole infinite pool. The more randomly sampled examiners’ scores are aggregated, the closer we get to the ‘true’ mark. This assumption breaks down when the examiners are not a random or representative sample.

The theory can also address fixed factors that do not fit this assumption, but they are rarely used to evaluate assessment data and are not discussed here.

**PRACTICAL ISSUES**

**Avoiding errors in sampling effects**

This section is intended to help investigators avoid the mistake of inadequate or non-random sampling, or ignoring important error factors in an evaluation.

Our first example problem is shown in Fig. 1.

A director of undergraduate assessment, aware of the feasibility benefits of patient assessors and the regulatory importance of reliable assessment, wishes to know how reliable are patient assessors’ ratings of students’ performances on simulated practical procedures.

**Figure 1** Problem 1

Examiners often give different scores to the same candidate responses. Examiner variation can be subdivided into examiners’ consistent tendencies to be harsh or generous – their ‘stringency’ \((V_{\text{examiner}})\) and the varying views that examiners take of a particular candidate because of their particular perspective or preconception – their ‘taste’ \((V_{\text{examiner}*\text{candidate}})\).

The same considerations apply to samples of candidates’ performance (cases, items, or simulated procedures in Problem 1). Candidates often vary across their performances more than examiners do in their judgements. Just like examiner variation, case variation can be subdivided into case ‘difficulty’ \((V_{\text{case}})\) and the varying ‘case aptitude’ that candidates display for a particular case \((V_{\text{candidate}*\text{case}})\).

Furthermore, examiners will often vary in their perceptions of the challenges of a particular case. In our example, a patient who has recently undergone cannulation may have certain strong views about that particular procedure. These views may give him idiosyncratic expectations which influence his scoring of cannulation, but do not affect his scoring of other procedures. This effect too can be subdivided into \(V_{\text{examiner}*\text{case}}\), where it has a consistent effect on how the assessor scores this procedure across all candidates (e.g. hawkish about cannulation) and \(V_{\text{candidate}*\text{examiner}*\text{case}}\), where it causes him to mark candidates idiosyncratically (e.g. only happy if the candidate warns about bruising). The latter is indistinguishable from residual, unexplained error.

The decisions about which effects contribute error to the final score depend completely upon the intended meaning of the score (its universe of generalisation). In quality assuring a norm-referenced test at a particular point in time, the score should reflect the relative performance of the candidates compared with each other. The first order effects (case difficulty \([V_{\text{case}}]\) and examiner stringency \([V_{\text{examiner}}]\)) do not contribute error provided that candidates are marked by the same examiners and perform the same challenges. Even \(V_{\text{examiner}*\text{case}}\) will cancel itself out across candidates in assessments where every candidate performance is multiple-marked by all examiners. However, examiner taste \((V_{\text{examiner}*\text{candidate}})\) and candidate case aptitude \((V_{\text{candidate}*\text{case}})\) will cause error. If these effects are large, they imply that the judgements are subjective and the performance domain is variable, respectively. Check that the examiners and cases are representative of their universes (the ‘average examiner on the street’ and the intended performance domain). If so, the main solution to the variation problem lies in sampling the views of more examiners and testing more cases. Excluding ‘outlier’ examiners and cases can have an unintended effect in that it can reduce the validity of score by making the sample less representative of its universe.

However, in a criterion-referenced test, the score is intended to reflect the performance of the candidates compared with any implementation of a similar test with any sample of examiners and cases. In this case, all the above factors contribute error to the
score as the particular sample of examiners and cases are only a sample from the intended universe of generalisation.

In research, representativeness is even more important. The sampling considerations mirror an hypothesis-testing experiment. For example, we are not interested in how much this sample of examiners varied in their scoring; instead, we want to know how much any sample of examiners would vary in their scoring. The same applies to each and every factor which influences the test score – candidates, cases and all their interactions. Hence, in order to accurately estimate the effects of each relevant factor, it becomes essential to test a large and representative sample of conditions (as it would in similar multi-factor experiments in industry).³

Hypothesis-testing studies use a power calculation to calculate how large a sample is required to make it probable that random sample variation is small compared with the expected effect size. Variance component estimations make no assumptions about effect sizes, so there is no equivalent to a power calculation. Ideally, the investigator should conduct 2 studies, consisting of a pilot study (or Generalisability study) in which effect sizes are estimated on a preliminary sample (as fully crossed as possible), followed by a fuller evaluation (or Decision study) using a sample size and design determined by the pilot. As a minimum, the size and representativeness of the actual sample should be discussed and justified.

It is a common mistake to invest heavily in assessing a large sample of candidates, but to use a small and unrepresentative sample of cases/items or examiners. It would be better to spread the available resources so that all relevant factors are reasonably well sampled.

From Problem 1, a fully crossed, random-effects model on a Generalisability study with assessors \( (n = 10) \) crossed with procedures \( (n = 10) \), crossed with candidates \( (n = 10) \) produced the following effect estimates based on \( (10 \times 10 \times 10 = 1000) \) scores (Table 1).

How do these effect estimates help to plan for reliable assessment?

In comparisons across programmes, stringency and difficulty are likely to cause error but can be reduced by sampling across assessors and procedures, or by matching the procedures assessed, or by training assessors.

Case aptitude tells the investigator about the uniformity or breadth of the performance domain. Provided that the cases properly reflect the intended area of performance, it is not appropriate to limit this, but rather to sample well from across it.

Assessor taste tells the investigator about the uniformity or breadth of the judgement domain. If all judgements are appropriate, then the same considerations apply: sample rather than limit.

### Table 1: A fully crossed, random-effects model on a Generalisability study with assessors \( (n = 10) \), crossed with procedures \( (n = 10) \), crossed with candidates \( (n = 10) \) produced the following effect estimates based on \( (10 \times 10 \times 10 = 1000) \) scores (Problem 1)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Variance component estimate</th>
<th>Proportion of total variance from each factor (%)</th>
<th>Effect sample (d.f.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{\text{student}} ) (the consistent differences between students' ability across assessors and procedures)</td>
<td>0.30</td>
<td>15%</td>
<td>9</td>
</tr>
<tr>
<td>( V_{\text{assessor}} ) (the consistent differences in assessor stringency across students and procedures)</td>
<td>0.36</td>
<td>18%</td>
<td>9</td>
</tr>
<tr>
<td>( V_{\text{procedure}} ) (the consistent differences in procedure difficulty across students and assessors)</td>
<td>0.20</td>
<td>10%</td>
<td>9</td>
</tr>
<tr>
<td>( V_{\text{student*assessor}} ) (the varying tastes of assessors between students across procedures)</td>
<td>0.30</td>
<td>15%</td>
<td>81</td>
</tr>
<tr>
<td>( V_{\text{student*procedure}} ) (the varying case aptitude of students between procedures across assessors)</td>
<td>0.24</td>
<td>12%</td>
<td>81</td>
</tr>
<tr>
<td>( V_{\text{assessor*procedure}} ) (the varying case-specific stringency of assessors between procedures across students)</td>
<td>0.40</td>
<td>20%</td>
<td>81</td>
</tr>
<tr>
<td>( V_{\text{student<em>assessor</em>procedure}} ) (residual error)</td>
<td>0.20</td>
<td>10%</td>
<td>729</td>
</tr>
</tbody>
</table>

Even sampling produces the best estimates of all the effects. Two patient assessors scoring 100 students on 3 procedures would produce poor estimates of assessor and procedure variance and their interactions. Full crossing provides the best estimates of interaction effects. If 1 assessor marked each procedure, assessor and procedure would be confounded; it would then be impossible to say whether a low score on a given procedure was a result of assessor stringency or procedure difficulty. Random sampling enhances the generality of the findings. If all the patients had recently had a bad cannulation experience (or came from a particular demographic group, or shared a medical condition), this might influence their scores in a way that makes them unrepresentative of the universe of patients' views. This is a common problem with assessors used in evaluations, who are often academics from the same institution. d.f. = degrees of freedom.
Case-specific stringency appears to be a substantial source of error. Examining the actual data may allow for the identification of an idiosyncrasy that requires further investigation. If not, reduce this within a programme by keeping assessors nested within a procedure. Reduce it across programmes by sampling plenty of procedures and examiners, even to the point of using multiple examiners per procedure if procedures are a more expensive resource than examiners.

**Dealing with naturalistic data**

This section is intended to guide investigators towards the best approach for estimating variance components from naturalistic datasets.

In evaluating a mini-CEX assessment process, the investigator receives 658 score sheets from 128 trainee doctors. A total of 174 observers have contributed; some have observed 20 doctors, but many have observed only 1 or 2 doctors. Some doctors have been observed by 5 assessors, but many have been observed by only 1 or 2. There are cases of crossover where 2 or more doctors have been seen by the same 2 or more observers. (There are also many cases of indirect crossover where doctor A saw assessors 1 and 2, doctor B saw assessors 2 and 3, and doctor C saw assessors 1 and 3.) However, in many observations, assessor and doctor are confounded.

**Figure 2 Problem 2**

The previous section examined an example of fully crossed assessment data in which every candidate saw every case and was multiple-marked by every examiner. This type of data allows for the most efficient variance component analysis, but is fairly uncommon even in controlled undergraduate assessment. Full crossing is almost unheard of in more pragmatic workplace assessment.

Data can deviate from full crossing in a number of ways. Effects may be nested or confounded and the data may be unbalanced by design or omission. Naturalistic data (as in the example shown in Fig. 2) often have all of these features. As the data become more naturalistic, the confidence of effect estimates reduces, and the potential for seriously erroneous estimates increases. However, it is still possible to produce reasonable estimates with appropriate analysis.

An effect is said to be nested if a set of scores (the test variable) represents varying levels of 1 explanatory factor (e.g. patient encounters) within 1 level of another explanatory factor (e.g. 1 doctor) that are not crossed with any other level of that factor (another doctor). This is common in workplace assessment, where, for example, patients cannot undergo multiple operations.

Two effects are confounded if 2 or more scores represent varying levels of 2 or more factors simultaneously (e.g. an observer scores doctor A on encounter 1 and doctor B on encounter 2). In this example it would be impossible to tell whether the difference between the 2 scores is produced by the different doctors or the different encounters, as in Problem 2 (Fig. 2).

A dataset can have several of these features by design. For example, if, in a controlled assessment, each of 10 supervisors scores their 1 trainee doctor on his or her performance in 10 patient encounters, then doctors and supervisors are completely confounded, and encounters are nested within both doctors and assessors. However, this is a balanced dataset because of the equal sample numbers. The data can also be systematically unbalanced, such as, for example, if 5 of the observers score 10 encounters and 5 score 15 encounters. The most complex data of all are naturalistic data such as those in the dataset described in Problem 2.

Data with systematic nesting or confounding must be analysed using an appropriate regression model. If doctor A’s encounters 1–10 are treated as if they were the same as doctor B’s encounters 1–10 when they are actually encounters 11–20, the estimates will be incorrect. How to specify an appropriate regression model will be set out in the user manual for whichever variance component estimating software is chosen, and is not the subject of this paper. Systematically unbalanced data (see above) are also straightforward to analyse using either uGENOVA\(^6,7\) or the more general software packages SPSS or SAS with appropriate regression models.

The greater problem arises when the data are naturalistic. This occurs where the assessment has been designed to a matrix but large portions of the data are missing, or where the data have been collected pragmatically with little reference to an intended matrix. Within such data there is likely to be nesting and confounding, but there are also likely to be subsets of scores representing independent variations in single factors and factor interactions. Careless analysis of this type of data can produce highly misleading estimates.

Firstly, missing data (by omission) could be filled in (imputed). However, the most common systems for imputation have a direct impact on apparent reli-
ability. Imputation based on the mean of the ‘true’ effect (candidate) will artificially increase reliability; imputation based on the mean of any ‘error’ effect (e.g. examiner) will artificially reduce reliability. This will produce misleading effect estimates if more than perhaps 5% of the data are missing. Alternative approaches are described at the end of this section.

Secondly, some mathematical procedures run the risk that a very provisional estimate for 1 effect (e.g. an estimate of examiner-examiner variation based on 1 or 2 non-confounded conditions of the factor ‘examiner’) will distort the calculation of all the other effects just because it precedes the other factors in the command syntax of the software. Clearly, this will lead to erroneous conclusions.

Disciplines other than education have a history of working with this type of data. For example, in agriculture it is common to apply variance component analysis to naturalistic data to ascertain the effect of combinations of climatic conditions, soil type, pesticides and fertilisers to the yield of different seed types. There is a large statistical literature on the subject. A recent paper reviewed the estimates and components much more rapidly, but it is impossible in practice to programme control cards either in DOS (or in Windows using ‘G_string’ software) if the dataset is large and highly naturalistic because every ‘cell’ must be specified in the card.

There has been a tendency in the educational literature to analyse naturalistic data of this type using 1 of 4 other approaches. The first 3 have serious problems.

1 Imputing significant amounts of data has the side-effect of artificially increasing or reducing reliability and probably introduces unacceptable error if more than 5% of the data are imputed.

2 Culling the data to leave only the ‘balanced’ sections used to be necessary before software was available to handle unbalanced data. Now, however, it simply reduces the confidence of the variance component estimates and runs the risk of excluding assessment data from a relevant group of outliers, such as candidates who do not finish.

3 Simplifying the regression model (e.g. to only ‘candidate’ and ‘error’ effects) involves the risk that an important explanatory variable (e.g. ‘examiner’) may be written out of the model and much of its variance attributed to the wrong category.

Table 2 Results of several different variance component procedures, applied to the real data in Problem 2

<table>
<thead>
<tr>
<th>Source of variance: procedure</th>
<th>Variance component/percentage of total variance (d.f. in the estimation)</th>
<th>$V_{\text{doctor}}$</th>
<th>$V_{\text{observer}}$</th>
<th>$V_{\text{doctor}+\text{observer}}$</th>
<th>$V_{\text{residual}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA ss I (simplified regression model)</td>
<td>0.36/44% (127 d.f.)</td>
<td>–</td>
<td>–</td>
<td>0.46/36% (529 d.f.)</td>
<td></td>
</tr>
<tr>
<td>ANOVA ss I (partial regression model, observer specified before doctor)</td>
<td>0.22/27% (41 d.f.)</td>
<td>0.34/41% (173 d.f.)</td>
<td>–</td>
<td>0.27/32% (442 d.f.)</td>
<td></td>
</tr>
<tr>
<td>ANOVA ss I (partial regression model, doctor specified before observer)</td>
<td>0.08/9% (127 d.f.)</td>
<td>0.49/59% (87 d.f.)</td>
<td>–</td>
<td>0.27/32% (442 d.f.)</td>
<td></td>
</tr>
<tr>
<td>ANOVA ss III (full regression model – any factor sequence)</td>
<td>0.04/5% (41 d.f.)</td>
<td>0.25/31% (87 d.f.)</td>
<td>0.26/33% (20 d.f.)</td>
<td>0.24/31% (442 d.f.)</td>
<td></td>
</tr>
<tr>
<td>MINQUE type I (full regression model – any factor sequence)</td>
<td>0.13/15%</td>
<td>0.19/22%</td>
<td>0.32/36%</td>
<td>0.24/28%</td>
<td></td>
</tr>
<tr>
<td>ml (full regression model – any factor sequence)</td>
<td>0.15/17%</td>
<td>0.20/22%</td>
<td>0.30/33%</td>
<td>0.24/28%</td>
<td></td>
</tr>
<tr>
<td>reml (full regression model – any factor sequence)</td>
<td>0.15/17%</td>
<td>0.20/22%</td>
<td>0.30/33%</td>
<td>0.24/28%</td>
<td></td>
</tr>
</tbody>
</table>

The MINQUE, ml and reml methods do not produce an indication of d.f., but, by using indirect methods of estimation, operate with at least as many d.f. as ANOVA ss III. Even in the full regression model, the residual term includes $V_{\text{doctor}+\text{observer}+\text{encounter}}$ but these are inseparable with this study design because encounters are totally confounded with the other factors.

d.f. = degrees of freedom; MINQUE = minimum norm quadratic unbiased estimator; ml = maximum likelihood; reml = restricted maximum likelihood

© Blackwell Publishing Ltd 2007. MEDICAL EDUCATION 2007; 41: 926–934
4 **Partitioning** unbalanced datasets into a series of balanced designs for variance component analysis and taking a weighted average of the variance components (Sirotnik’s approach) produces similar estimates to the ML procedure, but is probably harder work.9

Table 2 shows the results of several different variance component procedures, applied to the real data in Problem 2. It demonstrates how, using a simplified model, much of the observer variance is wrongly attributed to the doctor. It also shows how the sequence of factor specification has a profound impact on the estimates using ANOVA ss I or II. By contrast, ML, REML and MINQUE procedures give almost identical estimates.

**Interpreting and presenting variance components**

This section is intended to guide the investigator who has calculated variance component estimates towards an appropriate presentation of his or her conclusions.

A programme director whose trainees participated in the mini-CEX evaluation described in Problem 2 above is concerned about 1 of the trainees. This doctor scored poorly, but the director doesn’t know how much of this was caused by the doctor’s true performance and how much was caused by errors inherent in the assessment.

**Figure 3 Problem 3**

The last difficulty relates to the interpretation and presentation of variance component results (Fig. 3). Broadly, there are 3 possibilities:

- present the raw variance component estimates;
- take the error estimate and convert it to a measure of precision along the score scale (standard error of measurement [SEM]), or
- present the estimates as a proportion of total score variance that represents variation in the theoretical true score (coefficient of reliability or generalisability – R or G).

**Presenting raw variance component estimates**

We showed raw variance components in the previous 2 sections of this paper. Table 1 describes what the effects mean in plain English. The raw estimates contain the maximum amount of information. They show, for example, how much assessor error results from consistent differences in the stringency of assessors ($V_{assessor}$), and how much results from their subjectivity ($V_{assessor*doctor}$ and $V_{assessor*case}$). However, in their raw form, they do not give a ready indication of either the precision or the discriminating capacity of the score. (They would not directly assist the programme director in our example in Fig. 3). Moreover, the units can be difficult to grasp. The variance components are related to the score scale, but their absolute values are not as helpful as their relative values. When the relative values are converted to proportions or percentages for clarity, these too can be misread as measures of reliability or specificity.

**Standard error of measurement**

This is easy to calculate for simple assessments and can be presented in a very intuitive way as a confidence interval along the score scale. $SEM = \sqrt{\text{error variance}}$, and the 95% confidence interval for a ‘true score’ is ‘observed score’ ± ($SEM \times 1.96$). The SEM fits readily with a criterion-referenced minimum standard where comparison with other candidates is not important. However, although it indicates precision along the score scale, it provides no indication of discriminating capacity as it does not compare score precision with the observed spread of scores. A very precise result (e.g. 75% ± 5%) would be a poor discriminator if half the observed scores in a population lay between 70% and 80%.

**Reliability coefficient**

This is equally easy to calculate: $R = \frac{\text{‘true variance’}}{\text{‘true variance’} + \text{‘error variance’}}$. It varies between 0 and 1. In an ideal test, all score variance would indicate varying conditions of the ‘true’ construct and R would be exactly 1. The coefficient summarises both precision and discrimination in 1 number. However, the coefficient can be hard to conceptualise, and there is no intuitive standard: what level of reliability is enough? (Some authors have made helpful suggestions for particular circumstances.)10,11 Crucially, its very strength in combining precision and spread is also its weakness: the coefficient makes no distinction between an imprecise test applied to a heterogeneous population and a precise test applied to a homogeneous population. Criterion-referenced regulatory tests are likely to be more concerned with precision than with ranking.

**A fourth way**

The confidence interval can be presented graphically alongside the spread of scores. This allows for an
observed by 8 different assessors, each of whom observed 2 encounters. How reliable was his score?

The poorly scoring doctor described in Fig. 3 was observed by 8 different assessors, each of whom observed 2 encounters. How reliable was his score?

Avoiding errors in sampling effects

All 4 approaches allow some mathematical modelling of different assessment situations. Building on the principle of random sampling, any error variance component, such as V_assessor, will contribute less and less error variance as the score aggregates a bigger and bigger sample of conditions. Thus:

\[
\text{SEM} = \sqrt{(\frac{V_{\text{assessor}}}{\text{number of assessors}}) + \frac{V_{\text{error b}}}{n_{\text{error b}}}} \text{ etc.}
\]

Similarly:

\[
R = \frac{V_{\text{true}}}{(V_{\text{true}} + \frac{V_{\text{assessor}}}{\text{number of assessors}})} \frac{V_{\text{error b}}}{n_{\text{error b}}} \text{ etc.}
\]

Great care is required in interpreting variance components from complex assessments and applying them to different questions. For example, if a doctor’s gender affects his or her mini-clinical evaluation exercise (mini-CEX) score, it is not immediately obvious whether this should be treated as a source of error variance or not. In the section Avoiding errors in sampling effects, we described how the same variance estimates require different interpretation in quality assurance and research reporting.

The poorly scoring doctor described in Fig. 3 was observed by 8 different assessors, each of whom observed 2 encounters. How reliable was his score?

Taking the MINQUE estimates, and based on the worst-case scenario where assessors are completely nested within doctors, the calculations are as follows.

The SEM is:

\[
\sqrt{(0.19/8) + (0.32/8) + (0.24/16)} = 0.28.
\]

This gives a confidence interval (CI) of (0.28 x 1.96 =) ± 0.55.

The reliability coefficient (R) is:

\[
0.13/(0.19/8 + 0.32/8 + 0.24/16) = 0.62.
\]

Figure 4 shows a histogram of actual scores alongside the score CI (i.e. the fourth way). The CI has been placed around the minimum satisfactory standard on the scale (4.0). Any score within a 95% CI of this standard cannot be said to be better than unsatisfactory with 95% confidence.

From the diagram, the assessment appears to contain enough precision to indicate true concern over 1 doctor, and to exclude true concern for the other relatively weak performers. However, the CI crosses 3 quartiles, giving less than 95% confidence that a doctor in the top quartile is truly better than a doctor in the third quartile. It is clearly not sufficiently precise for precise comparison or ranking purposes.

CONCLUSIONS

We make the following recommendations.

- For both quality assurance and research evaluations, ensure that all relevant factors are sampled, and that the sampling meets the theory’s assumption that the conditions represent a random and representative sample of the factor’s ‘universe’. Research evaluations will require large samples of each factor if they are to generalise adequately.
- Where feasible, conduct 2 separate studies (pilot and evaluation, or Generalisability and Decision studies).
- For unbalanced data, use either UGENOVA (if it is possible to specify the arrangement of data in the card), or 1 of the procedures MINQUE, ML or REML in SPSS or SAS if the data are too complex.
- State which mathematical procedure was used and the degrees of freedom (d.f.) of the effect estimates. If the procedure does not report d.f.

**Figure 4** Histogram of actual scores alongside the score confidence interval, illustrating the ‘fourth way’
re-analyse with ANOVA ss III and report these d.f. as the other procedures have at least the same level of effect sampling.

- Describe and justify the regression model used.
- Present the raw variance components. Describe the effects they represent in plain, non-statistical language.
- If SEM or R coefficients are presented, give the equations used to calculate them.
- Make sure that the method of reporting reliability (precision or discrimination) is appropriate to the purpose of the assessment. This will usually demand a precision indicator such as SEM. Consider a graphical presentation to combine precision and discrimination.

Contributors: JC recognised the need for such a paper in discussions with BJ and CRo. JC produced the first draft. All other authors made substantial suggestions, resulting in fundamental changes to the scope and content of the first draft. All authors approved the final manuscript.

Acknowledgements: none.

Funding: none.

Conflicts of interest: none.

Ethical approval: not applicable.

REFERENCES


Received 19 December 2006; accepted for publication 16 May 2007