Generalisability in unbalanced, uncrossed and fully nested studies

Ajit Narayanan,1 Michael Greco2 & John I. Campbell3

OBJECTIVES There is growing interest in multi-source, multi-level feedback for measuring the performance of health care professionals. However, data are often unbalanced (e.g. there are different numbers of raters for each doctor), uncrossed (e.g. raters rate the doctor on only one occasion) and fully nested (e.g. raters for a doctor are unique to that doctor). Estimating the true score variance among doctors under these circumstances is proving a challenge.

METHODS Extensions to reliability and generalisability (G) formulae are introduced to handle unbalanced, uncrossed and fully nested data to produce coefficients that take into account variances among raters, ratees and questionnaire items at different levels of analysis. Decision (D) formulae are developed to handle predictions of minimum numbers of raters for unbalanced studies. An artificial dataset and two real-world datasets consisting of colleague and patient evaluations of doctors are analysed to demonstrate the feasibility and relevance of the formulae. Another independent dataset is used for validating D predictions of G coefficients for varying numbers of raters against actual G coefficients. A combined G coefficient formula is introduced for estimating multi-sourced reliability.

RESULTS The results from the formulae indicate that it is possible to estimate reliability and generalisability in unbalanced, fully nested and uncrossed studies, and to identify extraneous variance that can be removed to estimate true score variance among doctors. The validation results show that it is possible to predict the minimum numbers of raters even if the study is unbalanced.

DISCUSSION Calculating G and D coefficients for psychometric data based on feedback on doctor performance is possible even when the data are unbalanced, uncrossed and fully nested, provided that: (i) variances are separated at the rater and ratee levels, and (ii) the average number of raters per ratee is used in calculations for deriving these coefficients.
INTRODUCTION

The aim of this paper is to explore how reliability, generalisability and decision studies can handle the growing number of psychometric designs used in professional performance assessment studies where it is not possible to control for balanced, crossed and non-nested aspects as a result of, for instance, doctors seeing varying numbers of patients or not having the same set of colleagues who are familiar with their work. Unless rigorous and robust techniques can be found in such situations, it will not be possible to assure doctors or regulatory bodies that assessment outcomes are reliable in the context of non-repeated ratings given by raters and different numbers of raters per person rated (‘ratee’). Identifying doctors for whom feedback on their professional practice or performance suggests a need for further scrutiny or enhanced personal development on the basis of ‘under-performance’ will depend on showing that the data are sufficiently reliable in the first place for these purposes, especially in terms of there being sufficient numbers of raters.

Generalisability (G) theory quantifies possible sources of rater and item variance to obtain a measure of the reliability of obtained subject scores, especially when it is not possible to repeat the measurements.1,2 The two main aims of G theory are to identify the discriminating ability of the test and to integrate this with the reproducibility of the result. Both depend on identifying and removing any extraneous variance that affects the true scores of raters. Classical G theory (and the associated G coefficient) attempts to quantify this extraneous variance in two ways: by assuming that the same raters are used for all ratees (a balanced and unnested design), and by assuming that it is possible to repeat the measurement for the same ratee (a crossed design). Together, these provide estimates of rater and item interpretation variance that can be removed to result in true score differences between ratees. The outcome of a G study is a G coefficient that provides confidence that, if the test is repeated with different ‘universes’ of raters and ratees, similar results will be returned. For instance, \( G = 0.80 \) is the accepted level of reliability in typical professional assessment studies,3 implying that 80% of the variation is caused by real differences between raters, with the remainder attributable to other sources. However, reproducibility may be compromised if the discriminating ability is not achieved as a result of the unbalanced, uncrossed and fully nested aspects of the design. There is currently very little knowledge of how to amend classical G theory to take into account these confounding aspects.

Decision (D) studies calculate the minimum number of raters or items required for a generalisable outcome.1,3 For instance, if a balanced design returns a G coefficient of 0.90 with a certain fixed number of raters per ratee, a D study may show that it is still possible to achieve a respectable 0.80 G coefficient with some reduced number of raters per ratee. D studies depend on a balanced design to allow for experimentation with smaller or larger numbers of raters per ratee and so help determine the effect on the G coefficient. It is currently not clear how to amend D theory to cope with unbalanced studies.

Classical G and D studies use variance (squared standard deviation [SD]) as the main component of analysis. Alternative methods to G exist for estimating true variance for unbalanced studies. Crossley et al.4,5 recommend calculating a ‘true’ score as an ‘observed score’ (the mean score) plus or minus the standard error of measurement (SEM) multiplied by 1.96. A traditional ANOVA approach to calculating true score variance is to determine whether the mean scores of all ratees are equal and to partition any variances into different components (‘means model’). However, traditional ANOVA requires a balanced design so that each rater’s score for a ratee is compared (sums of squares [SS] approach) with the equivalent rater’s score for another ratee (e.g. rater 1 for ratee X with rater 1 for ratee Y; rater 2 for ratee X with rater 2 for ratee Y, etc.). If ‘cell’ sizes are unequal, the resulting SS will no longer partition by rater, by ratee and by item. If the data are culled by, for example, limiting the raters to a fixed number to achieve a balanced design, useful data for doctors with more than this fixed number of raters might be excluded in the variance calculations.

Extensions to traditional ANOVA have been proposed for dealing with this problem, including assuming a fixed, independent effect (a variable measured without error, which is not possible for many psychometric measurements including the case studies discussed here) and regarding cells as independent groups, the means of which deviate with the effect (‘effects model’). However, all such methods must guard against cells with large numbers of raters influencing cells with small numbers of raters when the amount of imbalance is high. Introducing weighted means (i.e. means weighted by sample size)
adds another layer of complexity and ambiguity to the study.

urGENOVA6 adopts statistical methods derived from Henderson7 for estimating random (i.e. not fixed) effects variance components using the analogous ANOVA procedure (also called Henderson’s Method 1) based on SS. Introducing random effects allows the analysis to generalise away from the actual raters and ratees used in the study and to make inferences about the ‘universe’ of raters and ratees. However, under a random-effects approach questionnaire items may have to be considered independently of one another, which is not the case for typical psychometric studies. In addition, urGENOVA is difficult to handle missing data, special nested-effect control cards must be used. Finally, all these alternative approaches to variance cannot generate D predictions in a naturalistic way for unbalanced studies. The lack of D predictions can lead to a lack of confidence that a sufficient number of raters has been achieved in unbalanced designs for reliable assessments of doctor performance.

This paper presents revisions to existing reliability and G formulae, as well as new formulae for D studies, for use by researchers involved with unbalanced, uncrossed and fully nested designs. To help readers understand the formulae and their applicability, two datasets resulting from actual, multi-sourced feedback evaluations of doctors are analysed later in the paper, together with fully worked examples (‘Case Studies’ section). A small artificial dataset (also unbalanced, uncrossed and fully nested) is used in the ‘Methods’ section to help readers interpret the formulae as they are introduced.

METHODS

Reliability

Cronbach’s alpha8 is widely adopted as a statistic for checking the reliability (also called ‘internal consistency reliability’) of psychometric instruments (questionnaires) and is classically9 defined as:

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum_{i=1}^{n} \sigma_{Yi}^2}{\sigma_X^2} \right)$$

(1)

where $n$ is the number of questionnaire items, $\sigma_X^2$ is the variance (squared SDs) of the total set of item scores, and $\sigma_{Yi}^2$ is the variance of item $i$. If all items have identical scores (all $\sigma_{Yi}^2$ are equal), then $\sigma_X^2 = n^2 \sigma_{Yi}^2$ so that $\alpha = 1$. The alpha formula measures ‘true variance’ as given by individual items as a proportion of total variance across all items. The formula can be alternatively stated using standard variance components:

$$\alpha = \frac{n \times \bar{a}}{\left( \bar{v}_i + (n-1) \times \bar{a} \right) \times \bar{a}}$$

(2)

where $\bar{a}$ is the average of all covariances between the items and $\bar{v}_i$ is the average variance across all items. Covariance is a measure of the amount by which each item varies in accordance with another item. However, Formula 2 does not account for repeated measurements of the same subject by varying numbers of non-overlapping raters.8,9 A revised measure of questionnaire reliability at the unaggregated (U) level (i.e. the rater level) can be estimated if Formula 2 is supplemented with the average rater variance $\bar{v}_r$ as a source of measurement error:

$$\alpha_U = \frac{n \times \bar{a}}{\left( \bar{v}_i + (n-1) \times \bar{a} + \bar{v}_r \right) \times \bar{a}}$$

(3)

Consider the artificial data in Table 1, where Formula 3 returns 0.59 (worked example) and Formula 2 would return 0.76. In other words, adding average rater variance explicitly as an error takes into account the fact that all raters are different and has the overall effect of reducing the reliability produced by classical alpha.

Generalisability

Generalisability (G theory) extends classical reliability by dealing with the problem of multiple sources of variance and error.5 The G formula calculates the G coefficient as:

$$G = \frac{tv}{tov} = \frac{tv}{tv + ev}$$

(4)

where $tv$ is true score variance and tov is total score variance, which in turn is tv plus any error score variance ev. For instance, Crossley et al.10 provide the following formula for calculating a general reliability coefficient R for a single observation in a fully crossed assessment (i.e. all raters see all ratees and use the same items):

$$R = \frac{v_a}{v_a + v_b + v_c + v_{a+b+c} + v_{a+c} + v_{b+c} + v_{a+b+c}}$$

(5)
where $\nu_r$ is the ratee variance (true variance), $\nu_v$ is the rater variance, $\nu_i$ is the item variance and the other terms signify variance interactions and residuals between ratees and raters, ratees and items, raters and items, and ratees by raters by items. However, these interactions can only be confidently estimated if raters rate more than one doctor and ratees are rated by the same raters.

A further complication is that the analysis must operate at two levels: unaggregated and aggregated. At the unaggregated level, the data consist of all the ratings given by raters. The number of samples at this level (the number of records in the database) equals the number of raters. At the aggregated level, the data consist of ratings received by ratees, or the mean of all raters’ scores (the number of records equals the number of ratees).

To calculate G coefficients in such situations, Formula 6 and 7 are adapted from Formulae 4 and 5 for the unaggregated (U) and aggregated (A) data, respectively:

$$G_U = \frac{tv}{tv + \nu_r + \nu_i + (v_r + v_i)}$$

where the average rater and item variances from Formula 3 are reused but their interaction is now included as a source of score error. Formula 6 is essentially a measure of the generalisability to the universe of raters at the unaggregated level of data analysis.

### Table 1
An artificial unbalanced, uncrossed and fully nested performance evaluation dataset. Three doctors, A, B and C, are rated by 16 separate patients after a consultation; eight patients rate Dr A, five rate Dr B, and three rate Dr C. Variations in the number of responses per doctor could have been caused by different procedures used by receptionists for handing out and collecting questionnaires, or by the non-completion of some questionnaire items that led to the removal of the entire rater response from analysis. The five questionnaire items adopt Likert scale responses where 1 = the lowest value (e.g. ‘unsatisfactory’) and 5 = the highest value (e.g. ‘excellent’). There is no duplication among the raters and all ratings are non-repeated. Alpha scores derived from Formulae 2 and 3 are also included and the working of Formula 3 is shown.

<table>
<thead>
<tr>
<th>Doctor</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>Rater variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.0</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Average unaggregated item variance $\nu_i = 0.24$

Average item covariance $\nu_i = 0.093$

Classical $\sigma$ (Formula 2) $\sigma = 0.76$

Revised $\sigma$ (Formula 3) $\sigma = (16 \times 0.093)/(0.24 + 15 \times 0.093 + 0.17) = 0.59$
To calculate the generalisability to the universe of ratees at the aggregated (mean score) level, Formula 7 is used, where:

- \( \bar{v}_i / n \) is the unaggregated item variance divided by the average number of raters \( n \) per ratee contributing to this variance;
- \( a\bar{v}_s \) is the aggregated average ratee variance; and
- \( a\bar{v}_i \) is the aggregated average item variance.

In other words, generalisability at the aggregated level of ratee score now includes rater variance to deal with the uncrossed and fully nested aspects. Evaluations of doctor (ratee) performance based on patient and colleague rater scores, for instance, are at the aggregated level. Table 2 contains an aggregated version of Table 1 and worked examples of Formulae 6 and 7.

### Decision studies

In the absence of \( G \), or if the study is unbalanced, the Spearman–Brown prophecy formula can be used to determine the number of raters for any given level of reliability:\(^{11}\)

\[
\tilde{n} = \frac{\rho \times (1 - \beta)}{\beta \times (1 - \rho)}
\]

where \( \tilde{n} \) is the estimated number of raters (rounded-up integer), \( \rho \) is the desired reliability level and \( \beta \) is a suitable reliability measurement, such as intraclass correlation coefficient (ICC; e.g. an ICC of 0.15 would result in 23 raters being required to achieve a desired reliability of 0.8). The ICC is a more specific measure of questionnaire reliability than alpha that checks that raters’ responses are in the same response range and not just based on simple correlation.

However, an unbalanced, uncrossed and nested design does not easily fit any of the Shrout–Fleiss\(^{12}\) categories of ICC because of the unbalanced nature of varying numbers of raters per ratee. Shavelson and Webb\(^{13}\) provide a formula for calculating a \( G \) coefficient that can also be used for \( D \) studies (GD):

\[
GD = \frac{\bar{v}_{ratees}}{\bar{v}_{ratees} + (\bar{v}_{raters} / n_{raters})}
\]

where \( \bar{v}_{ratees} \) is the variance for subjects (ratees), \( \bar{v}_{raters} \) is the rater variance and \( n_{raters} \) is the hypothetical

### Table 2: The aggregated (mean scores) version of the artificial data in Table 1, with unaggregated and aggregated \( G \) results and the workings of Formulae 6 and 7. The average doctor variance is the average ratee variance and the average aggregated item variance is the average variance of the mean scores for each doctor.

<table>
<thead>
<tr>
<th>Doctor</th>
<th>Aggregated item 1</th>
<th>Aggregated item 2</th>
<th>Aggregated item 3</th>
<th>Aggregated item 4</th>
<th>Aggregated item 5</th>
<th>Ratee variance</th>
<th>Number of raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.50</td>
<td>3.63</td>
<td>3.38</td>
<td>3.25</td>
<td>3.38</td>
<td>0.02</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>4.00</td>
<td>3.80</td>
<td>3.60</td>
<td>3.60</td>
<td>3.40</td>
<td>0.05</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>3.67</td>
<td>3.67</td>
<td>3.33</td>
<td>3.33</td>
<td>3.00</td>
<td>0.08</td>
<td>3</td>
</tr>
<tr>
<td>Average aggregated item variance</td>
<td>0.06</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>Average ratee variance</td>
</tr>
</tbody>
</table>

Unaggregated \( G \) (Formula 6) \( G_6 \) = \( (0.17 + 0.24)/(0.17 + 0.24 + 0.24 \times 0.24) = 0.91 \)

Aggregated \( G \) (Formula 7) \( G_7 \) = \( (0.05 + 0.035 + 0.17)/(0.24/5 + 0.05 + 0.035 + 0.17 + 0.05 \times 0.035 + 0.05 \times 0.17 + 0.035 \times 0.17 + 0.05 \times 0.035 \times 0.17) = 0.80 \)
number of raters varied to produce different values of \( G_D \). However, Formula 9 does not take into account either two-level or unbalanced designs in the way that Formula 7 does.

A further complication is that D studies for varying numbers of raters (unbalanced designs) have two applications. First, for the actual study undertaken and actual average numbers of raters per ratee, what is the minimum number of raters required to produce an acceptable G coefficient in that study? This posterior calculation (D1) is important for identifying how far to go down the scale of average responses per ratee in a current study and still obtain a reliable result. This is different from a follow-on prior or expected calculation (D2): if the study is repeated using new raters and ratees, what is the minimum number of average raters required to produce a reliable and generalisable outcome, given what we know already from D1 about the average number of raters used?

In Formula 10, the posterior D1 calculation is ‘summative’ in terms of building a coefficient out of the sum of individual components, where:

- \( a\bar{v}_i \) is the total variance of the average number of raters per ratee at the aggregated level;
- \( \bar{v}_i \) is the total variance of the average number of raters per ratee at the unaggregated level;
- \( n \) and \( p \) are the actual average number of raters and actual number of items, respectively; and
- \( m \) and \( k \) are hypothetical and variable numbers of items and raters, respectively, for determining the effects on G.

When \( l = p \) and \( m = n \), Formula 10 reduces to Formula 7 for calculating the aggregated G coefficient.

The prior or expected D2 calculation in Formula 11 for subsequently repeated studies, by contrast, is ‘subtractive’ in that it assumes a desirable, idealised G coefficient of 1 and removes from this the variances that have been found to already exist:

\[
G_{D2} = 1 - \frac{a\bar{v}_i + \bar{v}_i + \bar{v}_i}{(a\bar{v}_i/k) + (a\bar{v}_i \times p) + (\bar{v}_i \times k) + \bar{v}_i}
\]  

where the prior generalisability \( G_{D2} \) is the idealised generalisability 1 minus (the numerator in Formula 11) the sum of the ratee, aggregated item, rater, unaggregated item and unaggregated item-by-rater variance, divided by the sum of (the denominator in Formula 11):

- the existing average ratee variance divided by a hypothetical and variable number of raters \( k \);
- the average aggregated item variance multiplied by the number of items \( p \) to be included in the questionnaire;
- the effect of \( k \) raters at the unaggregated level; and
- the average unaggregated item variance.

Worked examples of Formulas 10 and 11 are provided for the case study data (Tables 4 and 5).

CASE STUDIES AND RESULTS

To test the formulae, multi-source datasets from two unbalanced, uncrossed and fully nested performance evaluation studies were analysed. The first dataset comes from a large-scale pilot study exploring the potential utility of patient and colleague feedback obtained using questionnaires developed for the purpose of multi-source feedback and proposed as a component of the process for relicensing UK doctors \(^1\) (‘GMC study’). Feedback relating to 309 index doctors was obtained from 4269 colleague respondents using 17 generic assessment Likert scale-based items within a colleague questionnaire. In addition, feedback on 380 index doctors was obtained from 13 754 patient responses using nine generic assessment Likert scale-based items within a patient questionnaire. Both patient and colleague responses were available for 291 doctors and these make up the case study. Classical alpha was originally reported as 0.920 for the colleague data and 0.898 for the patient data. The study adopted ICC to predict a minimum of 22 completed patient questionnaires and eight completed colleague questionnaires, using a desired reliability of 0.85 and the Spearman–Brown proph-
When reliability and G coefficients are re-calculated (Table 3) for the GMC study using Formulae 3, 6 and 7, the revised alpha formula (Formula 5) produces lower but acceptable reliability coefficients (0.85 for colleagues, 0.81 for patients). The original G coefficients are now shown to have underestimated the generalisability when data are aggregated to the level of the doctor ratee (now 0.87 and 0.93 for aggregated colleague and patient data, respectively). The GMC D1 study (Table 4) indicates, for the actual study, that at least 12 colleagues and 30 patients are required for a doctor’s assessment to be reliable at G = 0.80, given the actual average response rate of 14 colleagues and 36 patients. The D2 studies (Formula 11), by contrast, predict that, in future, nine and 16 fully completed colleague and patient responses, respectively, will be sufficient for producing an aggregated G of 0.80 (Table 5). In other words, there is some evidence of over-sampling in the GMC study, particularly in the case of patients.

The second case study involves a general practitioner (GP) study using the Colleague Feedback Evaluation Tool (CFET) and the patient-completed Doctors’ Interpersonal Skills Questionnaire (DISQ) instruments. The data focused on feedback regarding the performance of UK GPs. This feedback consisted of 1927 colleague questionnaire responses (from an average of 13 colleagues per doctor) regarding the performance of 145 index GPs, and 7738 patient questionnaire responses (from an average of 48 patients per doctor) regarding the performance of 161 GPs. Responses from both colleagues and patients were available for 124 GPs and these data represent the case study. The CFET and DISQ consisted of 17 and 10 generic assessment Likert scale items, respectively. No previous analysis of these data has been reported. The results of Formulae 2, 3, 6
and 7 are provided in Table 3, those of Formula 10 in Table 4 and those of Formula 11 in Table 5, together with worked examples.

For the GP data, a separate and fully independent dataset was analysed to provide a validation of D2 predictions in Table 5 (‘Expected Aggregated G’). The new GP dataset used the same questionnaires and consisted of 2421 colleague questionnaire responses concerning the performance of 179 doctors (an average of 14 responses per doctor) and 8784 patient questionnaire responses for these same 179
components. Applying Formula 12 to the GMC study summed from each source. The denominator includes aggregated item and unaggregated rater variances for the case of colleague and patient data), and average ratee, variance into account when calculating internal

\[
G_m^A = \frac{\sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum v_i}{\sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum \alpha_i f_j^l + \sum v_i}
\]

(12)

A similar validation analysis was not possible on the GMC data because of the lack of a second, independent GMC dataset. In addition, it was not possible to extend the GP patient validation as a result of the lack of doctors with fewer than 40 patient questionnaire responses. Finally, an overall, multi-source G for ratees \(G_m^A\) based on their multi-sourced rater data can be estimated (Formula 12) using a generalisation of Formula 7. Multi-source brings together into one dataset the rater data derived from different sources (e.g. patients and colleagues) for specific ratees (doctors), thus providing a ‘multi-source feedback’ for those ratees.

In Formula 12 \(t\) is the number of sources \((t = 2\) in the case of colleague and patient data), and average ratee, aggregated item and unaggregated rater variances are summed from each source. The denominator includes estimates of the error caused by these summed components. Applying Formula 12 to the GMC study and GP data resulted in a GMC \(G_m^A\) of 0.89 and a GP \(G_m^A\) of 0.80. Whether such a multi-sourced \(G\) is anything more than a statistical curiosity will depend on further work with multi-sourced professional assessment data.

DISCUSSION AND CONCLUSIONS

If rater feedback is used to judge doctor performance, we must be confident that there is reliability and precision in raters’ scores as well as sufficient numbers of raters for rates. The revised alpha coefficients (Formula 3, Table 3) for the GMC pilot data (0.85 and 0.81 for colleague and patient data, respectively) and the GP study data (0.86 and 0.94 for colleague and patient data, respectively) indicate that classical alpha (Formula 2) overestimates the reliability (0.92, 0.89, 0.93 and 0.97, respectively) of unaggregated and fully nested data by not taking rater variance into account when calculating internal

questionnaire consistency at source. The precision of rater scores will actually be lower, raising questions about the justifiability of any further scrutiny of a doctor’s performance if only classical alpha is used.

Unaggregated G (Formula 6) returns 0.90 and 0.91 for GMC colleague and patient data, respectively, and 0.82 and 0.88 for the corresponding GP data (Table 3), indicating that 90%, 91%, 82% and 88% of variances are generalisable to the universes of the respective colleague and patient rater domains. In other words, these are the generalisable percentages that can be attributed to real differences in the way that raters see ratees, with the remainder attributable to item and rater variation.

Table 3 also shows that the GMC pilot data have higher unaggregated G (0.90, 0.91) than revised alpha (0.85, 0.81), whereas the GP data have higher revised alpha (0.86, 0.94) than unaggregated G (0.82, 0.88). This may reflect the fact that the GMC pilot study contained many more colleague and patient raters than the GP study and its results are therefore more generalisable to the universes of colleagues and
patients. Researchers adopting unbalanced, uncrossed and fully nested designs will need to be aware that introducing more raters to raise generalisability may also have the effect of introducing more item interpretation inconsistency, as well as more rater variability, leading to lower precision in source scores.

The revised aggregated $G$ (Formula 7), which is the unbalanced, uncrossed and fully nested version of classical $G$, makes it possible to explicitly identify and isolate variances at the unaggregated and aggregated levels (both case studies) and so to generate familiar generalisability estimates at the ratee level. For the GMC and GP data, the aggregated $G$ coefficients of 0.87, 0.93, 0.80 and 0.89 (Table 3) for corresponding colleague and patient data indicate that 87%, 93%, 80% and 89%, respectively, of the differences between doctors are attributable to real differences in the scores received by doctors, with the rest attributable to rater and item variances as well as deviations from the average number of raters per ratee (Formula 7). The two $D$ formulae (Formulæ 10 and 11) allow for experimentation with variable sample sizes (both case studies) in a way that is familiar to researchers used to balanced designs (Tables 4 and 5).

The validation of the $D_2$ predictions on a totally independent dataset (Table 6) shows that, despite the different statistical approaches adopted by Formula 11 (for predicting future $G$) and Formula 7 (for calculating actual $G$), there is very close

| GMC D2 | Patients | | GP D2 | Patients |
|--------|----------| |        |----------|
| Colleagues | Expected aggregated | | Colleagues | Expected aggregated |
| Raters | G | Raters | G | Raters | G |
| 9 | 0.80* | 4 | 0.55 | 10 | 0.73 | 16 | 0.74† |
| 10 | 0.81 | 8 | 0.70 | 11 | 0.74 | 20 | 0.78 |
| 11 | 0.82 | 12 | 0.78 | 12 | 0.76 | 24 | 0.81 |
| 12 | 0.83 | 16 | 0.82 | 13 | 0.77 | 28 | 0.83 |
| 13 | 0.84 | 20 | 0.85 | 14 | 0.78 | 32 | 0.84 |
| 14 | 0.85 | 24 | 0.87 | 15 | 0.79 | 36 | 0.86 |
| 15 | 0.85 | 28 | 0.89 | 16 | 0.80 | 40 | 0.87 |
| 16 | 0.86 | 32 | 0.90 | 17 | 0.81‡ | 44 | 0.88 |
| 17 | 0.86 | 36 | 0.91|| | 48 | 0.89 |

* $1 - ((0.05 + 0.08 + 0.18 + 0.28 + (0.18*0.28))/(0.05/9) + (0.08*17) + (0.18*9) + 0.28)]$
† $1 - ((0.02 + 0.08 + 0.17 + 0.69 + (0.69*0.17))/(0.02/16) + (0.08*10) + (0.17*16) + 0.69)]$
‡ $1 - ((0.11 + 0.13 + 0.35 + 0.56 + 0.35 + (0.56*0.35))/(0.11/17) + (0.13*17) + (0.35*17) + 0.56)]$
§ $1 - ((0.01 + 0.03 + 0.14 + 0.28 + (0.28*0.14))/(0.01/36) + (0.03*9) + (0.14*36) + 0.28)]$

The significance of bold, italic entries for GP D2 is explained in Table 6.

Table 5. The results of the $D_2$ studies (Formula 11) on the General Medical Council (GMC) study and general practitioner (GP) study datasets, providing estimates of expected aggregated $G$ coefficients should the study be repeated for different average numbers of raters. Two particular entries in each set of predictions are expanded to show how the $G$ estimates are calculated.
correspondence in the G estimates, indicating that these formulae are accurately identifying the variances caused by an unbalanced, uncrossed and fully nested design. Furthermore, the D calculations in the case studies indicate that there may be a tendency to over-sample patients, perhaps in the belief that patients’ responses are less reliable (more varied) than colleagues. Table 3 shows that, on average, patient responses tend to be less varied than colleague responses (0.14 versus 0.18 for GMC study data, 0.17 versus 0.35 for GP study data) because of the usual and limited role that patients have in their interaction with doctors. Colleagues, however, will have many different roles and so their responses will vary more. Both case studies use colleague and patient questionnaires that measure different aspects of doctor performance. For instance, in the GMC study, colleagues rate aspects such as clinical ability and knowledge, medical record keeping, supervision and teaching capability, which may not be measured by patients, who focus instead on, for example, communication skills and explanations of treatment. As multi-source data will typically evaluate non-overlapping aspects, it is quite possible that different sources will have their own minimum sample numbers for the same set of doctors.

In conclusion, the revised alpha and G formulae introduced in this paper build on classical techniques but extend them by, firstly, distinguishing between unaggregated and aggregated sources of error and, secondly, using the average number of raters per ratee as a reference point for G and D studies. Psychometric approaches to assessing doctor performance can continue even when the data are unbalanced, uncrossed and fully nested, provided that:

1. variances are distinguished by ratings given and ratings received;
2. such variances are introduced at the right level of analysis for questionnaire reliability, generalisability and decision studies, and
3. the average number of raters per ratee is used in calculations for deriving G and D estimates.

These formulae can allow the application of subsequent techniques for identifying ‘under-performing’ doctors (e.g., refs. 2,11) on the basis that the data are sufficiently reliable for this purpose at the aggregated level.

Contributors: all authors made substantial contributions to the study conception and design, the acquisition, analysis and interception of data, and the drafting and critical revision of the article. All authors approved the final paper for publication.

Acknowledgements: the authors are grateful to the various referees whose critical and constructive comments have resulted in an improved paper.

Funding: none.

Conflicts of interest: none.

Ethical approval: not applicable.
REFERENCES


Received 30 July 2009; editorial comments to authors 16 September 2009, 6 November 2009, 19 November 2009; accepted for publication 24 November 2009.