

Routledge
Taylor & Francis Group

Applied Measurement in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Evaluating Human Scoring Using Generalizability Theory

Yaw Bimpeh , William Pointer , Ben Alexander Smith & Liz Harrison

To cite this article: Yaw Bimpeh , William Pointer , Ben Alexander Smith & Liz Harrison (2020) Evaluating Human Scoring Using Generalizability Theory, Applied Measurement in Education, 33:3, 198-209

To link to this article: https://doi.org/10.1080/08957347.2020.1750403

Published online: 21 Jul 2020.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hame20

APPLIED MEASUREMENT IN EDUCATION
2020, VOL. 33, NO. 3, 198–209
https://doi.org/10.1080/08957347.2020.1750403

Routledge
Taylor & Francis Group

Check for updates

Evaluating Human Scoring Using Generalizability Theory

Yaw Bimpeh, William Pointer, Ben Alexander Smith, and Liz Harrison

Research and Analysis, AQA

ABSTRACT

Many high-stakes examinations in the United Kingdom (UK) use both constructed-response items and selected-response items. We need to evaluate the inter-rater reliability for constructed-response items that are scored by humans. While there are a variety of methods for evaluating rater consistency across ratings in the psychometric literature, we apply generalizability theory (G theory) to data from routine monitoring of ratings to derive an estimate for inter-rater reliability. UK examinations use a combination of double or multiple rating for routine monitoring, creating a more complex design that consists of cross-pairing of raters and overlapping of raters for different groups of candidates or items. This sampling design is neither fully crossed nor is it nested. Each double- or multiple-scored item takes a different set of candidates, and the number of sampled candidates per item varies. Therefore, the standard G theory method, and its various forms for estimating inter-rater reliability, cannot be directly applied to the operational data. We propose a method that takes double or multiple rating data as given and analyzes the datasets at the item level in order to obtain more accurate and stable variance component estimates. We adapt the variance component in observed scores for an unbalanced one-facet crossed design with some missing observations. These estimates can be used to make inferences about the reliability of the entire scoring process. We illustrate the proposed method by applying it to real scoring data.

1. Introduction

Almost all general qualifications in the United Kingdom (UK) involve examinations that are developed and rated (scored) by an external examination board. The kind of tasks mostly used in large-scale assessment in the UK require response scripts to be rated by subject experts who score each question using a set of criteria to judge how well a candidate has performed on each task or question. Unfortunately, errors in expert judgment such as leniency, halo effect or severity create systematic variation in the awarded scores. Previous research has shown that raters' errors are not negligible in large-scale assessments and they can affect valid interpretation of scores (Koretz, Stecher, Klein, & McCaffrey, 1994). Engelhard (1994) and Zhang (2013) described different types of rater error (e.g., rater severity, halo effect, central tendency, and restriction of range).

There are different ways in which variability in a candidate's score can arise. It might be due to variations in the rater's (or raters') scoring, in candidates' performance or in the sampling of items (Shavelson, Gao, & Baxter, 1996). High inter-rater reliability is desirable so that raters can be considered interchangeable (i.e. a score awarded is not dependent on the specific raters that are assigned to score the item) and it is required for exam scores that have more significant consequences for test takers (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

CONTACT Yaw Bimpeh ybimpeh@aqa.org.uk AQA Stag Hill House, Guildford, GU2 7XJ, Surrey, UK
© 2020 Taylor & Francis Group, LLC

In this paper, we are interested in the rating quality when constructed-response tests are scored by humans. We look at the consistency of scoring across raters (inter-rater reliability) and the level of consistency of individual raters (intra-rater reliability).

Numerous plausible models have been proposed in the literature for analyzing rating data collected from multiple raters (Eckes, 2017). Wind and Peterson (2018) carried out a systematic review of the main methods used to evaluate rating quality in different application areas. The most common method identified for evaluating ratings based on the scaled ratings is Rasch measurement theory, and methods based on the observed ratings are dominated by inter-rater reliability, rater agreement, and generalizability theory (G theory). Wind and Peterson also highlighted the dependence on aggregate-level information that is not specific to individual raters or specific factors of an assessment.

The increase in on-screen rating by the UK exam boards opens up possibilities to gauge inter-rater reliability. According to the UK National Foundation for Educational Research (NFER) 2013 report on the review of the literature on rating reliability (Tisi, Whitehouse, Maughan, & Burdett, 2013), there is no current agreement about the best means of quantifying rater reliability. The report also stated that the UK assessment community is not any closer to achieving a key recommendation from the Meadows and Billington (2005) review, which was to publish reliability information alongside the results of examinations. The NFER systematic review of the best available and relevant literature published on rating reliability indicates that data from routine monitoring of on-screen rating has the potential to be used for regular measurement of rating reliability. The NFER review also indicates that G theory can be used to separate sources of error in test results and to model the effects on reliability of changing parameters, such as the number of raters. Johnson and Johnson (2009) explored the potential for using G theory to study the reliability of assessment outcomes and measurement errors from various sources, such as items and raters in the context of the examination results used by UK general qualification systems.

The purpose of this paper is to estimate rater reliability by applying G theory to datasets from the routine monitoring of on-screen human rating with missing observations.

2. Brief Background to UK Rating System

The on-screen rating system used by most UK exam boards involves the scanning and segmentation of scripts into images: the candidates' responses are viewed on computer screens, rather than in their original paper format. This technology allows question papers to be split into individual items. Examiners, using rating software that is installed on their computers, then access the images for the items that they are required to rate. The question papers are scored on screen at either the item or script level (whole-paper level). For a reliability check, samples of these responses are randomly double "blind" rated or seeded.

Seed responses are pre-selected responses rated by the senior examiner (or a senior examiner panel), which are then introduced randomly into the rating. The rater is not aware that a particular response is a seed and has no knowledge of the score that has been awarded. The same seed responses may be scored by many raters. Seed responses are closely monitored by the senior examiner, or senior examiner panel, throughout the rating session to ascertain whether examiners' scores are within a given tolerance (i.e. maintain desirable levels of accuracy). The senior examiner's score is taken as the correct score. Examiners are expected to be within a given tolerance of seed scores or else they will be flagged by the system and then they may be either retrained or stopped from rating. All candidates in the monitoring dataset are evaluated by many raters, but each item gets a different candidate–rater combination.

In double "blind" rating, neither rater is aware of the score awarded by the other. In general, about 5% of item responses are randomly double "blind" rated. When the scores given by the two raters are out of tolerance from each other, a senior rater will decide what score should be given to the candidate's work.

3. Brief Description of G Theory

G theory provides a statistical framework for evaluating the reliability of measurement and other related factors (Brennan, 2001; Shavelson & Webb, 1991). G theory allows the researcher to establish, for example, whether the main cause of measurement error is due to the test items or the rating; how the reliability of the measurement can be improved; and whether the test scores are sufficiently reliable to make decisions about the level of a candidate's performance for a qualification.

In G theory, a test measurement (e.g., a test score) is conceived of as a sample from a universe of admissible observations. The universe refers to the condition of measurement from which the decision maker is willing to generalize. A measurement condition has characteristic features such as test form, test item, rater, and/or test occasion. Each characteristic feature is called a facet of a measurement. A universe of admissible observations, then, is defined by all possible combinations of the facets (e.g. items, raters). The objects of measurement, for example persons, are not a facet. In order to use G theory to estimate the effects of the facets or factors of concern for a given testing situation, we conduct a generalizability study (G study) – in which the effects of these different facets are clearly distinguishable – that provides sample statistics indicating the degree of these effects in the sample used. This is done through a linear model that relates the obtained measurement to the facets that are assumed to affect the measurement. In the model, the facets can be either crossed (all levels of a facet are observed with all conditions of another source of variation), or nested (some levels of a facet are observed with some conditions of another source of variation in specific context).

The G study sample statistics provide the basis for a decision study (D study), in which the goal is to estimate how changes in the number of items and/or number of raters would improve the reliability of a candidate's score, and how best to optimize future measurement of test scores.

3.1. *Inter-Rater Reliability Under a G Theory Framework*

Consider, for example, a two-facet crossed person by item by rater (*p* x *i* x *r)* random-effects design. Here, the candidates (i.e. persons) are the object of measurement and are therefore not a facet. In the *p* x *i* x *r* design (with generalization over all admissible test items and raters taken from an indefinitely large universe), if X_{pir} denotes the score awarded to person p on item i by rater r, then the linear model is:

$$X_{pir} = \mu + p + i + r + (pi) + (pr) + (ir) + (pir) + e \tag{1}$$

where μ = the grand mean, *p* = the person effect, i = the effect of item on the observed score, r = the effect of rater on the observed score, e = random error, and other terms indicate interactions between facets.

The term μ is the grand mean (or expected value) over all persons in the population, and all items and raters in the "universe of admissible observations", where the number of persons in the population is infinite, and the number of items and raters in the universe of admissible observations are infinite. The variance of the scores given by Equation (1), over the population of persons and the conditions in the universe of admissible observations may be decomposed as:

$$\sigma^2(X_{pir}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(r) + \sigma^2(pi) + \sigma^2(pr) + \sigma^2(ir) + \sigma^2(pir, e) \tag{2}$$

The seven terms on the right-hand side of Equation (2) are called "variance components". These variance components provide a partitioning of$\sigma^2(X_{pir})$. By partitioning variance of the scores in this way, G theory helps to identify the main sources of measurement error, estimate the total magnitude of error, and form a reliability coefficient.

The first variance component in Equation (2), $\sigma^2(p)$, reflects differences among candidates' performance. The differences in the difficulties of the items are captured by the second variance component $\sigma^2(i)$. The third variance component $\sigma^2(r)$represents the differences in the severities and leniencies of raters. The candidate-by-item interaction, which reflects the differential performance of

candidates on different items, is explained by the fourth variance component $\sigma^2(pi)$. The variance component $\sigma^2(pr)$ indicates the interaction due to the fact that different raters may apply the scoring scheme differentially across candidates. The component $\sigma^2(ir)$ represents the rater-by-item interaction, which shows the degree to which some raters apply different rating rules to items that have the same level of difficulty. The seventh variance component $\sigma^2(pir, e)$ reflects randomness, other unidentified error, or both.

Typically, an estimate of each variance component is obtained using analysis of variance (ANOVA) – other methods such as maximum likelihood (ML) or restricted maximum likelihood (REML) can also be used (see, for example, Searle, Casella, & McCulloch, 1992). The estimates are obtained by equating observed and expected mean squares from an analysis of variance and solving the resulting equations. If the data are balanced, the ANOVA estimators have many appealing properties. For unbalanced data, the property of unbiasedness that is implicit in the ANOVA does not hold, which creates a number of problems. Various procedures are available in the literature for estimating variance components for unbalanced data that yield unbiased estimators. However, they can all produce awkwardly negative estimates for the variance. The typical treatment for negative variance estimates is to set them to equal zero (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

From the variance components, the inter-rater reliability can be assessed through the calculation of the generalizability coefficient, which is the proportion of the variance of the universe score to the variance of the observed score. The classical inter-rater reliability estimate is obtained by correlating the scores assigned by two different raters to a group of candidate responses to the same item. In G theory, the standard coefficient uses a G study design (person by item by rater) with two raters rating the same items. In terms of variance components in Equation (2), the inter-rater reliability is given by the generalizability coefficient denoted $E\rho^2$ (see Brennan, 2001, p. 129):

$$E\rho^2 = \frac{\sigma^2(p) + \sigma^2(pi)/n_i}{\sigma^2(p) + \sigma^2(pi)/n_i + [\sigma^2(pr)/n_r + \sigma^2(pir, e)/n_i n_r]} \tag{3}$$

The number of raters and items are denoted by n_r and n_i, respectively. For a D study interested in p x i x r design with a single random rater and a single fixed item, the n_r and n_i in Equation (3) can be set to one. The generalizability coefficients may be used in evaluating the degree to which raters are operating as intended and also the reliability of candidate scores based on two ratings.

4. Analyzing Human Rating Data under a G Theory Framework

In order to analyze high-stakes rating data in the UK using a G study, one has to, firstly, identify facets or factors in the intended universe, with particular focus on which factors are effectively random (i.e. allowed to vary) or effectively fixed (i.e. constant for all observations); secondly, decide on or identify the data-collection design; and finally, evaluate the extent to which raters are functioning. Two facets can be identified in the rating: item and rater. The analysis of variations in raters' scoring under G theory ideally requires a fully crossed or a nested measurement design: candidates are either fully crossed or else they are nested with raters and items. In such designs, the effects of candidates, raters and items can be separately estimated. When the design is neither fully crossed nor nested, some effects may be confounded with others. The fully crossed design requires that each candidate in the sample be rated by all raters. However, this is not feasible or cost effective in the high-stakes operational rating settings. With the on-screen rating design, each item is scored by a different set of randomly assigned raters. This can increase the number of raters scoring each candidate and thus helps to control any lack of consistency among raters.

The double-rating procedure currently used is a more complex design that consists of cross-pairing of raters and overlapping of raters for different groups of candidates. In a typical double-rating procedure, raters are nested within candidates such that different groups of candidates are rated by different pairs of raters. However, the on-screen double-rating design is neither fully crossed

nor nested. The seeded rating design is also unbalanced with the same seed items being rated by many examiners but with each item receiving a different candidate–rater combination. Due to the complexity of the monitored rating datasets, the above method from G theory, and its various forms for estimating inter-rater reliability, cannot be directly applied to the high-stakes rating data in the UK. It is generally known that G theory is unstable, and potentially invalid variance component estimates may result from using unbalanced or sparse data (Brennan, 2001). However, unbalanced data or missing observations are common in operational performance-assessment settings (e.g., Brennan, 1992; Shavelson & Webb, 1991) because of the nature of the assessment design. This is the case for high-stakes rating data in the UK.

In light of this problem, we propose a method that takes double and seeded rating data as given, and analyzes the datasets by item in order to obtain more accurate and stable variance component estimates. In estimating item-level variance components, we use Henderson's Method I (1953), which is best known to provide unbiased estimates of variance components for unbalanced random-design models (see, for example, Brennan, 2001; Searle, 1987). The variance components for the entire paper are then obtained by pooling the estimates from the item-level variance components. The method used to aggregate variance components from each item dataset is similar to the one used by Chiu and Wolfe (2002). The rater reliability for the entire paper is then calculated based on weighted estimated variance components.

Below, we describe the proposed sampling model for double-rated or multiple-rated data, and also how to estimate the inter-rater reliability and absolute standard error of measurement for an item. We then discuss methods for synthesizing double- and multiple-rating datasets to obtain the inter-rater reliability for the entire paper. Following this, we discuss procedures for adjusting rater reliability coefficients for sample range restriction when this has occurred. Finally, we illustrate the proposed method.

4.1. *Sampling Model for Double- or Multiple-Rating Items*

The appropriate model for this double or multiple "blind" rating design is the unbalanced persons (p) nested within items (i) crossed with raters (r) random-effect design (i.e., $p{:}i$ x r design). However, due to the sparse data matrix for this rating data, and the fact that items are fixed, we analyze each item separately using a persons (p) crossed with raters (r) random-effect design with missing data. Generally, there are two approaches to analyzing designs with a fixed facet. The first approach involves running an analysis of variance considering all facets as random (even the fixed facet); identifying the random parts of the mixed design and the associated variance components. The variance components of the random parts are then adjusted to reflect the interaction between the source of variance and the fixed facet. The second approach is to run separate G studies within each condition of the fixed facet. This approach allows different conclusions to be reached about the item rating. Here, since each item takes a different set of candidates, and candidates' item responses are evaluated by all of the raters, we used the second approach. We analyze each item separately and report results for each item separately; therefore, the interaction between the source of variance and the item does not feature.

We conduct item-level G studies and then average the variance components from several G studies to get more accurate and stable variance component estimates for the entire paper. Thus, we model the conditional sampling design for each item by an unbalanced persons (p) crossed with raters (r) random-effect design for a given item (i.e., p x $r|\ i$ design) and then obtain its variance component. The variance components for the unconditional design ($p{:}i$ x r design) can then be obtained by weighting the conditional variance components of the data subsets (p x $r|\ i$ design) by their sample sizes, across all items.

For our conditional sampling design (p x $r|\ i$ design), the observed score X_{pr} assigned to any person (p) by any rater (r) for an item (i) conforms to an unbalanced (a design with missing data) random linear model:

$$X_{pr} = \mu + p + r + (pr, e) \tag{4}$$

where (pr, e) is residual effect. Variance components within each item represent a one-facet design (raters). Accordingly, only person main effects, rater main effects, and the residual can be estimated within an item. The target of measurement (here, candidates) are crossed with a facet of measurement (here, raters), but not all conditions of that measurement facet are associated with all targets: scores are missing for some pairings of raters and candidates. We therefore adapt the variance component in observed scores for an unbalanced one-facet crossed design with some missing observations, as discussed in Brennan (2001, p. 236). Using Brennan's (2001) formulation, the expected value of the observed variance in candidate scores in Equation (4), $E\sigma_i^2(Xpr)$, for an item (i) that has been averaged across h_{ri} raters per candidate can be expressed as:

$$\mathbf{E}\sigma_i^2\left(X_{pr}\right) = \sigma_i^2(p) + \left[\frac{1}{h_{ri}} - \frac{\sum_p \sum_{p'} \frac{\tilde{n}_{pp'}}{\tilde{n}_p \tilde{n}_{p'}}}{n_{pi}\left(n_{pi} - 1\right)}\right]\sigma_i^2(r) + \frac{1}{h_{ri}}\sigma_i^2(pr, e) \tag{5}$$

for item i = 1, …, k, where $\tilde{n}_p$ is the number of raters for person p, $\tilde{n}_{pp'}$is the number of raters each pair of candidates p and p' $(p \neq p')$ share in common for item i, $n_{pi}$is the total number of candidates whose item i was randomly double or multiple rated, and $h_{ri}$is the harmonic mean of the number of raters per candidate $\tilde{n}_p$.

$$h_{ri} = \left[\frac{1}{n_{pi}}\sum_{p'}\frac{1}{\tilde{n}_p}\right]^{-1}$$

The residual $\sigma_i^2(pr, e)$ confounds the p x r interaction effect with other sources of unsystematic variation (here, called e). For a double-rating design, $h_{ri} = 2$. Equation (5) allows for conditions with a different number of raters per candidate.

Variance components are estimated using Henderson's Method I, which does not require distributional assumptions (e.g. normality), or the specification of *a priori* weights, and it yields the same results as the standard procedures when the design is balanced. Henderson's Method I estimates variance components by equating the expected value of quadratic forms to sample counterparts and solving for the unknown population parameters. Searle et al. (1992) provide an in-depth discussion of the available statistical procedures for estimating variance components with unbalanced or sparse designs.

4.2. *Item-level Inter-rater Reliability for Randomly Double- or Multiple-rated Items*

The coefficient of generalizability is defined as the ratio of universe (true) score variance $(\sigma_i^2(p))$ to expected observed score variance, $E\sigma_i^2(Xpr)$. Using Equation (5), the inter-rater reliability for any given item (i) that is randomly double or multiple rated can be approximated by the per-item generalizability coefficient:

$$E\rho^2(i) = \frac{\sigma_i^2(p)}{\mathbf{E}\sigma_i^2(Xpr)}$$

The estimates of $E\rho^2(i)$ are obtained by the sample estimates of the parameters in Equation (5). Thus:

$$E\hat{\rho}^2(i) = \frac{\hat{\sigma}_i^2(p)}{\hat{\sigma}_i^2(p) + \left[\frac{1}{h_{ri}} - \frac{\sum_p \sum_{p'} \frac{\tilde{n}_{pp'}}{\tilde{n}_p \tilde{n}_{p'}}}{n_{pi}\left(n_{pi} - 1\right)}\right]\widehat{\sigma}_i^2(r) + \frac{1}{h_{ri}}\hat{\sigma}_i^2(pr, e)} \tag{6}$$

The $E\hat{\rho}^2(i)$in Equation (6) is the *ith* item inter-rater reliability index. For a D study design with a single random rater and a single fixed item, the h_{ri} in Equation (6) equals one. Thus, setting the

harmonic mean of the number of raters per candidate $\tilde{n}_p$to one (i.e., $h_{ri} = 1$) in Equation (6) gives the consistency of rating that we would expect if a single rater scores the items, rather than all the raters who scored the seed or double-rating data.

The absolute error, which is the error involved in using a candidate's observed mean score as an estimate of his or her universe (true) score, is defined as:

$$\Delta = Xpr - \mu_p$$

Consequently, the variance of the absolute errors for the p x r design with missing data for each seeded item i is given by (see Brennan, 2001, p. 235):

$$\sigma_i{}^2(\Delta) = \frac{1}{h_{ri}}\sigma_i{}^2(r) + \frac{1}{h_{ri}}\sigma_i{}^2(pr, e) \tag{7}$$

where $\sigma_i{}^2$ denotes the variance component for p x r design for each item i and h_{ri} is the harmonic mean of the number of raters per candidate $\tilde{n}_p$, which is the number of raters who scored common persons as defined above. Again, h_{ri} can be set to 1 in the item-level D study in order to estimate the absolute error in an individual's scoring in the wider rating process.

The square root of Equation (7), which is $\sigma_i(\Delta)$, may be interpreted as an estimate of the absolute standard error of measurement. This indicates how far scores assigned are likely to deviate from their true scores, and therefore is useful if the absolute level of candidate performance is of concern (Shavelson & Webb, 1991).

4.3. *Script-Level Inter-Rater Reliability Based on Item-Level Double or Multiple Rating*

Data subsets for each item may differ in the sample as a result of the different random process used to select a number of candidates for double or multiple rating by a relatively small number of random raters (for any item). While an individual G study provides valuable information on variations in the rater reliability for each item, a synthesis of all G studies offers insights into overall trends across script rating. The question is how to combine the information obtained from all the different item-level G studies. In what follows, a method for obtaining a composite index for script-level inter-rater reliability, based on item-level double or multiple rating, is proposed.

One can combine the information obtained from all different G studies by pooling the estimates to obtain an overall estimate for the variance components: we estimate $E\rho^2$ by pooling estimates from each of the G studies. A method similar to the one used by Chiu and Wolfe (2002) may be used to aggregate item-level variance components. The variance components for the whole paper based on item-level double or multiple rating are estimated by the weighted estimates of the variance components in Equation (5). If the sample sizes of the G studies differ, then the estimates from the larger studies will be more precise than the estimates from the smaller studies. In this case, it is reasonable to give more weight to the more precise estimates when pooling. This leads to weighted estimators of the form:

$$\hat{\sigma}^2(p) = \frac{\sum_{i=1}^{k} n_{pi}\hat{\sigma}_i{}^2(p)}{\sum_{i=1}^{k} n_{pi}} \qquad \hat{\sigma}^2(pr, e) = \frac{\sum_{i=1}^{k} n_{pi}\hat{\sigma}_i{}^2(pr, e)/h_{ri}}{\sum_{i=1}^{k} n_{pi}}$$
$$\hat{\sigma}^2(r) = \frac{\sum_{i=1}^{k} c_{r(i)} n_{pi}\hat{\sigma}_i{}^2(r)}{\sum_{i=1}^{k} n_{pi}} \tag{8}$$

where k is the total number of items and $c_{r(i)}$ is defined as:

$$c_{r(i)} = \left[\frac{1}{h_{ri}} - \frac{\sum_p \sum_{p'} \tilde{n}_{pp'}/\tilde{n}_p\tilde{n}_{p'}}{n_{pi}(n_{pi} - 1)}\right]$$

The purpose of the multiplier $c_{r(i)}$ is to scale the contribution of rater variance to the observed score variance based on the amount of overlap between the sets of raters who rate each candidate score. The factor $c_{r(i)}$ ranges from 0 to $1/h_{ri}$; it is equal to zero when the raters and candidates are fully crossed and equal to 0.5 when raters are nested within candidates with two raters per candidate.

The inter-rater reliability for all the double- or multiple-rated items can be computed on the basis of those pooled variance components, which is approximately equal to the generalizability coefficient:

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(r) + \hat{\sigma}^2(pr, e)}$$

5. Adjustment for Range Restriction

Due to the way the responses are sampled for the purposes of monitoring, they may not be representative of the observed distribution of scores, which could affect the estimate of a rater reliability coefficient. In particular, range restriction (i.e., if the sample has a restricted range of scores) may cause the variance components to be underestimated. It is important to note that the range restriction described here is different from situations in which ratings are clustered about the mean of the rating scale, reflecting rater unwillingness to use either of the extreme ends of the rating scale. There are two widely known formulae for a "corrected" correlation coefficient, which adjust for a restriction of the range (see, for example, Gulliksen, 1950, pp. 130–140). These are adapted here to be applied to the inter-rater reliability coefficient to adjust for the curtailment of seeds and double-rated data. These procedures estimate what the rater reliability coefficient would be in the unrestricted sample (i.e., the entire cohort) when we know the variances of the single-rated data and double/multiple-rated data as well as the inter-rater reliability coefficient for a restricted sample.

If a regression line is fitted between the first and second instance of double rating, the standard error of rating (scatter about the regression line) can be estimated. We assume here that the standard error of estimate has remained constant (i.e., homoscedasticity of the regression line). It is necessary to assume that the standard error of measurement of the single-rated items will be quite similar to that for first or second instances of double or multiple rating; that is, the average standard error of measurement of each of the double- or multiple-rating incidents is assumed to be the same as the standard error of measurement for the single-rated item response population. This assumption is equivalent to assuming that:

$$\sigma_{obs.p}{}^2\left(1 - r_p{}^2\right) = \sigma_{obs.s}{}^2\left(1 - r_s{}^2\right)$$

where $\sigma_{obs.p}$is the unrestricted standard deviation of the observed scores of the entire population, $\sigma_{obs.s}$is the restricted standard deviation of the observed scores in the sample, $r_p{}^2$is the rater reliability coefficient for the whole cohort, $r_s{}^2$is the rater reliability coefficient for the sample.

Thus, a "corrected" reliability coefficient that adjusts for the restriction of range can be estimated as:

$$r_p{}^2 = 1 - \frac{\sigma_{obs.s}{}^2}{\sigma_{obs.p}{}^2}\left(1 - r_s{}^2\right)$$

According to many authors (e.g., Greener & Osburn, 1980; Gulliksen, 1950), the minimum assumptions necessary for the application of this correction formula are: (i) a regression line can be fitted between the first and second instances of double rating; and (ii) the standard error of estimate (scatter about the regression line) remains constant (i.e., homoscedasticity of the regression line). However, Gross (1982) has shown that a correction formula can be valid for a given selection procedure even if the assumptions are violated; that is, even if the regression is non-linear and the standard error of estimates is not homoscedastic. One could also use the direct range restriction correction formula Thorndike Case 1 (see Gulliksen, 1950, Eq. 18, p. 137).

6. Numerical Illustration of the Proposed Method

In this section, examples are given of the application of the theoretical results discussed above. The empirical data used for the illustration comes from the high-stakes A-level qualification in England. The General Certificate of Education Advanced Level (A-level) is a qualification available in various subjects and taken by many students in England, Wales and Northern Ireland at 18 years of age. The proposed method is applied to a sample of double/multiple-rating data: in particular, June 2017 GCE A-level rating data.

Tables 1 and 2 summarize the estimated G study percentage of variance components for each item relating to persons and raters and various statistics in terms of absolute error (this is simply the difference between a person's observed and universe scores) and the generalizability coefficient (rating reliability) for the sample of GCE A-level subject rating. These statistics provide diagnostic evidence of rating quality specific to individual items. Also provided in Tables 1 and 2 are the number of candidates whose item responses were double or multiple rated, the number of raters monitored, and the maximum score allotted for each item. The variance component estimates were obtained using the analogous-ANOVA method (Henderson's Method I). From these results, the per-item rater reliability ($E\hat{\rho}^2$) was calculated using Equation (6).

From the estimated variance components, we see that rater was a negligible source of variation for low-tariff items (i.e., items with a low maximum score allotted). Thus, only a small percentage of the total variation was attributable to rater (0.11% to 8.96% for all the items except BC33 in Table 2). This means that variability in a candidate's score due to the differences between the standards used by individual raters is small. This small percentage does not mean that differences between raters are small in an absolute sense. A meaningful way to interpret the variation across raters is to use the square root of the variance component for raters: the standard deviation. The magnitude of a standard deviation can be used to estimate the range within which raters' scores fall.

Tables 1 and 2 show relatively large values for the differences among candidates' performance (i.e. person variance). This indicates that the majority (i.e., 34% to 95%) of the variability in scores was attributable to persons (which reflects the differences in ability or competency across candidates). Also, variability in scores attributable to persons varies across items: differences between persons are generally greater for low-tariff items than high-tariff items. This indicates that some low-tariff items are good at distinguishing levels of candidate competency.

The residual, which confounds the effect of person and rater interaction with other sources of unsystematic variation, explained a moderately large proportion of the total variance in scores in some items (ranges between 2.87% and 62% depending on the item). Moderately large residual variance components suggest there is either a large person/rater interaction effect or that other sources of variation have not been captured in the model, or both. A large variance component for the person and rater interaction means that the relative score of the person differs from rater to rater.

Table 1. Estimates of item-level rater reliability for GCE AS/A-level subjects.

				Percentage of variance				
Item	Maximum score	No. of candidates	No. of raters	Person	Rater	Residual*	Absolute error	Ep2
B11	4	30	63	88.44	1.18	10.38	0.19	0.88
B12	5	25	62	62.95	6.39	30.66	0.66	0.63
B13	5	25	62	57.33	3.84	38.82	0.72	0.57
B14	6	23	61	64.89	0.83	34.28	0.95	0.65
C15.1	9	21	63	76.33	2.87	20.80	1.32	0.76
C15.2	16	436	62	64.79	4.60	30.61	2.67	0.65
C16.1	9	21	62	60.93	4.40	34.67	1.66	0.61
C16.2	16	432	59	58.69	4.00	37.31	4.04	0.59
B17	4	32	93	88.35	1.64	10.02	0.19	0.88
B18	4	21	92	54.21	7.40	38.39	0.55	0.54
B19	6	22	90	34.01	4.02	61.96	1.30	0.34
B20	9	27	86	83.28	0.41	16.31	1.70	0.83
B21	9	19	85	77.27	2.42	20.31	1.91	0.77
CC_Essay	25	672	87	43.81	6.79	49.40	15.50	0.44

Table 2. Estimates of item-level rater reliability for GCE AS/A-level subjects.

				Percentage of variance				
Item	Maximum score	No. of candidates	No. of raters	Person	Rater	Residual*	Absolute error	Ep2
DD_Essay	25	665	83	38.11	7.53	54.36	13.31	0.38
BB21	3	23	34	89.85	2.59	7.55	0.13	0.90
BB22	4	28	29	92.43	4.70	2.87	0.20	0.93
BB23	4	25	35	73.95	3.15	22.90	0.60	0.74
BB24	4	38	34	76.59	5.92	17.50	0.34	0.77
BB25	10	26	31	47.79	1.88	50.33	2.07	0.48
BB27	3	22	33	93.17	0.07	6.76	0.11	0.93
BB28	4	29	35	93.04	1.28	5.68	0.12	0.93
BB29	4	26	33	72.11	0.91	26.98	0.41	0.72
BB30	4	28	34	77.69	1.89	20.42	0.29	0.78
BB31	10	32	31	56.91	5.93	37.17	1.75	0.57
BB32	25	222	32	68.10	8.26	23.64	6.21	0.68
BC22	4	26	28	93.08	0.11	6.81	0.16	0.93
BC23	4	29	27	76.27	2.60	21.13	0.51	0.76
BC24	4	35	28	85.31	1.95	12.74	0.27	0.85
BC25	10	26	27	57.24	2.59	40.17	1.73	0.57
BC26	25	173	26	73.80	5.69	20.51	5.75	0.74
BC31	10	30	27	67.83	7.02	25.14	2.36	0.68
BC32	25	108	27	66.03	8.96	25.01	9.32	0.66
BD31	10	21	81	54.18	4.14	41.68	1.90	0.54
BD32	15	516	77	43.92	6.94	49.15	5.36	0.44
BD33	25	523	77	46.83	14.46	38.71	10.18	0.47

The inter-rater reliability, $E\hat{\rho}^2$, indicates the degree to which scores awarded by raters are reliable. The $E\hat{\rho}^2$ estimates are for a D study design; that is, p x r with a single random rater and a single fixed item. Thus, the estimates can be interpreted as the consistency of rating that we would expect if a single rater scored the items. This helps to inform the interpretation and use of item-level ratings in the wide sense. The raters providing scores are assumed to be sampled from a population of raters.

The estimated rater reliability ranges between 34% and 95% for all the cases evaluated. The reliability of performance ratings given in practical settings is usually quite low, often falling below 60% (see, for example, Houston, Raymond, & Svec, 1991). This was the case in 1 in 3 of the samples evaluated. The absolute error variance provides different information about the degree of uncertainty in the scores awarded by raters. Thus, the error variances show the magnitude of error when generalizing from the raters' scores to the candidates' true scores. It is important to interpret the rater reliability together with the standard error of measurement. The absolute error variance for item-level score ranges 0.1 to 15.5 depends on the item tariff (see Tables 1 & 2, column 9). When interpreting the raters' performance in an absolute sense, the standard error of measurement for an item is expected to be the square root of absolute error.

The representativeness of the sample was investigated by examining the effects of range restriction or sample variability on rater reliability estimates. The adjusted and unadjusted reliability coefficients are reported in Table 3. The standard deviations of the sample ratings actually evaluated and of the target population are also given. Results are presented for a sample of items from the A-level rating data.

It may be noted that there are a variety of range restriction effects on the rating reliability coefficient. The effects on observed rating reliability were relatively small: in most cases, the adjusted value of the rating reliability was within 5% of the original estimate. However, there were a handful of cases where the effects of range restriction were more noticeable. In these extreme cases, the effect was to underestimate or overestimate the rating reliability and might be attributed to a restricted selection of responses used for rating monitoring.

Table 3. Empirical evaluation of the effect of range restriction on marking reliability.

Item	Max mark	Pop SD	Sample SD	Abs. error	Ep^2	Adj. Ep^2
B12	5	1.34	1.34	0.66	0.63	0.63
B13	5	1.15	1.30	0.72	0.57	0.46
B14	6	1.52	1.65	0.95	0.65	0.59
C15.1	9	1.80	2.36	1.32	0.76	0.60
C15.2	16	2.82	2.75	2.67	0.65	0.66
C16.1	9	1.88	2.06	1.66	0.61	0.53
C16.2	16	3.08	3.13	4.04	0.59	0.58
B16	3	1.34	1.23	0.14	0.91	0.92
B17	4	1.60	1.29	0.19	0.88	0.92
B18	4	0.99	1.09	0.55	0.54	0.44
B19	6	1.47	1.40	1.30	0.34	0.40
B20	9	2.69	3.19	1.70	0.83	0.77
B21	9	2.77	2.90	1.91	0.77	0.75
CC_Essay	25	5.18	5.25	15.50	0.44	0.42
DD_Essay	25	4.36	4.64	13.31	0.38	0.30
A3	4	1.02	1.13	0.27	0.79	0.74
A4	4	1.17	1.22	0.24	0.84	0.82
A5	2	0.62	0.56	0.12	0.63	0.69
A6	12	2.49	3.02	1.58	0.83	0.74
BB10	2	0.52	0.38	0.08	0.42	0.70

7. Discussion and Conclusions

Data from the routine monitoring of on-screen rating has the potential to be used for the regular measurement of rater reliability using G theory. The analysis of variations in raters' scoring under G theory ideally requires a fully crossed or nested rating design. However, the rating procedures currently used in the UK constitute a more complex design, consisting of the cross-pairing of markers and overlapping of markers for different groups of candidates. G theory cannot be applied directly to multiple- and double-rated data. We therefore used unbalanced random-effects designs (with candidates crossed with raters). However, as scores were missing for some pairings of raters and candidates, we adopted the method discussed in Brennan (2001, p. 236) to estimate potential sources of error in the item rating, and provide item-level rating reliability indices as well as aggregated item-level inter-rater reliability for the entire paper. We also discussed a procedure for adjusting rater reliability coefficients for sample range restriction when this has occurred.

In the majority of items analyzed, raters exhibited a relatively small or negligible effect on candidate scores; however, for high-tariff items, there was some evidence of non-negligible rater effects. Overall, the impact on scores of rater effects was trivial for the sample of data analyzed.

Most commonly, the residual effect (which confounds the effect of person and marker interaction with other sources of unsystematic variation) explained a moderately large proportion of total variance in scores on some items. This suggests the possibility of a moderately large variance component for the person and rater interaction, which means that the relative scores of persons might differ from rater to rater or that there are other sources of variation that have not been captured in the rating design, or both.

The method used here is suitable for analyzing operational rating data with missing observations. We use the method operationally to provide valuable statistics, including item function statistics, to aid better understanding of the types of items and rating criteria that foster high rating reliability. It is also used to alert item writers, assessment design managers and principal examiners to items that have problems with rating reliability. In addition, the rating reliability statistics are at item level, potentially highlighting where changes could be made to improve reliability in future.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. doi:10.1111/j.1745-3992.1992.tb00260.x

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

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement*, *26*(3), 321–338. doi:10.1177/0146621602026003006

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

Eckes, T. (2017). Guest editorial rater effects: Advances in item response modeling of human ratings–Part I. *Psychological Test and Assessment Modeling*, *59*(4), 443–452.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. doi:10.1111/j.1745-3984.1994.tb00436.x

Greener, J. M., & Osburn, H. G. (1980). Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and non-linear distributions. *Educational and Psychological Measurement*, *40*(2), 337–346. doi:10.1177/001316448004000208

Gross, A. L. (1982). Relaxing assumptions underlying corrections for range restriction. *Educational and Psychological Measurement*, *42*(3), 795–801. doi:10.1177/001316448204200311

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

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*(2), 226–252. doi:10.2307/3001853

Houston, W. M., Raymond, M. R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement*, *15*(4), 409–421. doi:10.1177/014662169101500411

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability* (Ofqual Report No. 10/4706). Coventry England: Ofqual.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–6. doi:10.1111/j.1745-3992.1994.tb00443.x

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability* (Report for the National Assessment Agency by AQA Centre for Education Research and Policy, England).

Searle, S. R. (1987). *Linear models for unbalanced data*. New York, NY: Wiley.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Shavelson, R. J., Gao, X., & Baxter, G. (1996). On the content validity of performance assessments: Centrality of domain-specification. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 131–141). Boston: Kluwer Academic Publishers.

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

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A review of literature on marking reliability research* (Report for Ofqual). Slough England: NFER.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192. doi:10.1177/0265532216686999

Zhang, M. (2013, March). *Contrasting automated and human scoring of essays* (R&D Connections, No. 21). Princeton, NJ: Educational Testing Service.