Creating Test Score Bands for Assessments Involving Ratings using a Generalizability Theory Approach to Reliability Estimation

Charles Scherbaum  
*Baruch College, City University of New York*

Marcus Dickson  
*Wayne State University*

Elliott Larson  
*Baruch College and Graduate Center, City University of New York*

Brian Bellenger  
*Personnel Board of Jefferson County*

Kenneth Yusko  
*University of Maryland*

See next page for additional authors

**Recommended Citation**

Scherbaum, Charles; Dickson, Marcus; Larson, Elliott; Bellenger, Brian; Yusko, Kenneth; and Goldstein, Harold (2018) "Creating Test Score Bands for Assessments Involving Ratings using a Generalizability Theory Approach to Reliability Estimation," *Personnel Assessment and Decisions*: Number 4 : Iss. 1 , Article 1.  
DOI: https://doi.org/10.25035/pad.2018.001  
Available at: https://scholarworks.bgsu.edu/pad/vol4/iss1/1

This Invited Article is brought to you for free and open access by the Journals at ScholarWorks@BGSU. It has been accepted for inclusion in Personnel Assessment and Decisions by an authorized editor of ScholarWorks@BGSU.
Creating Test Score Bands for Assessments Involving Ratings using a Generalizability Theory Approach to Reliability Estimation

Authors
Charles Scherbaum, Marcus Dickson, Elliott Larson, Brian Bellenger, Kenneth Yusko, and Harold Goldstein
Creating Test Score Bands for Assessments Involving Ratings Using a Generalizability Theory Approach to Reliability Estimation

Charles Scherbaum\textsuperscript{1}, Marcus Dickson\textsuperscript{2}, Elliott Larson\textsuperscript{1}, Brian Bellenger\textsuperscript{3}, Kenneth Yusko\textsuperscript{4}, and Harold Goldstein\textsuperscript{1}

1. Baruch College, City University of New York
2. Wayne State University
3. Personnel Board of Jefferson County
4. University of Maryland

ABSTRACT

The selection of a method for estimating the reliability of ratings has considerable implications for the use of assessments in personnel selection. In particular, the accuracy of corrections to validity coefficients for unreliability and test score bands is completely dependent on the correct estimation of the reliability. In this paper, we discuss how generalizability theory can be used to estimate reliability for test score bands with assessments involving ratings. Using assessment data from a municipal entity, we demonstrate the use of generalizability theory-based reliability estimates in creating score bands and compare these estimates to those obtained using the traditional approaches.

KEYWORDS

selection, generalizability theory, banding, reliability

The estimation of measurement error in personnel assessments is one of the most fundamental tasks facing users of assessments in selection contexts (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014; Society for Industrial and Organizational Psychology, 2003). In many assessment situations, the estimation of reliability is fairly straightforward. For example, reliability estimates for scores on standardized written assessments (e.g., knowledge tests, personality inventories) can be estimated using multiple strategies (e.g., internal consistency methods, test-retest methods) with little difficulty. However, there are personnel assessment situations, especially those involving ratings from multiple assessors (e.g., interviews, assessment centers, work samples), where the estimation of reliability may not be straightforward or the typical methods (e.g., interrater correlations) may not be appropriate (Putka & Sackett, 2010; Putka, Le, McCloy, & Diaz, 2008). The literature over the last 20 years has offered numerous viewpoints on how these situations should be handled (cf. Murphy & Deshon, 2000a and Schmidt, Viswesvaran, & Ones, 2000).

The selection of a method for estimating the reliability of ratings has considerable implications for the use of assessments in personnel selection. In particular, the accuracy of corrections to validity coefficients for unreliability and test score bands is completely dependent on the correct estimation of the reliability. Although the appropriate estimation of reliability for corrections to validity coefficients has been discussed extensively in the literature (e.g., Deshon, 2001; Murphy & Desh, 2000a; 2000b; Putka & Hoffman, 2014; Schmidt et al., 2000), there has been very little discussion of these issues in the context of test score banding (see Murphy, 1994 for a general discussion of the impact of reliability on test score bands).

The existing discussion of reliability in the test score banding literature generally does not consider the method of reliability estimation or assumes methods appropriate for selection processes composed entirely of written assessments (e.g., knowledge tests, personality inventories). Considering the method of reliability estimation in test score banding is important, as there is a growing consensus that the most commonly used methods are only appropriate in some assessment contexts, and those contexts are not necessarily the norm (Putka & Hoffman, 2014; Putka et
Thus, the literature provides little guidance for those who are implementing test score banding with assessments involving ratings (e.g., interviews). In this paper, we discuss how generalizability theory can be used to estimate reliability for test score bands with assessments involving ratings.

Generalizability theory is a modern psychometric approach that allows one to decompose variation in observed scores into the various sources that produce that variation (e.g., interviewees, questions asked in the interview, raters evaluating the responses, and interactions among these sources). In the following sections, we provide a brief overview of generalizability theory, discuss estimating reliability for test score banding with assessments involving ratings using generalizability theory, and provide an example of test score banding of interview ratings using generalizability theory to estimate reliability.

A Brief Overview of Generalizability Theory

Reliability is concerned with the quantification of error that exists in any score or observation that is taken as part of measurement. A given score or observation can be said to be reliable to the extent to which it is free from error. The error can be random (e.g., due to fatigue) or systematic (e.g., due to rater tendencies). The dominant model of reliability in personnel selection and assessment is the classical test theory. Classical test theory, also known as the “true score” model, is based on the notion that an observed score on a measure is composed solely of true score and random error. In classical test theory, all systematic sources of variance – including those that might be error, such as rater tendencies – are attributed to the “true score” (Putka & Sackett, 2010).

The techniques for estimating the reliability of ratings in classical test theory are familiar to many assessment practitioners and researchers. For example, it is common to compute intrarater correlations or intraclass correlations using ratings that are averaged across items for each candidate by rater.

Despite the ease and prevalence of use, the appropriateness of this approach to reliability estimation in the types of rating contexts that are common in personnel assessment (e.g., rating designs where the raters do not rate all candidates) has been questioned (Murphy & DeShon, 2000a, Putka & Sackett, 2010; Putka et al., 2008). Instead, reliability estimates based on generalizability theory have been advocated.

Generalizability theory is a major extension of the assumptions about measurement error in classical measurement theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In a generalizability theory-based approach, error can be decomposed into its random and systematic factors such as rater characteristics, characteristics of the target of observation, characteristics of the observation context, and other factors (Brennan, 2001; Murphy & DeShon, 2000). For example, assessment center ratings can include estimates of error due to raters and due to exercises, as well as the interactions between these factors and the candidates. Classical test theory can be seen as a special case of generalizability theory where only two sources are believed to impact the observed scores (i.e., true score and error; Shavelson & Webb, 1991). Thus, generalizability theory approaches to reliability can appropriately estimate reliability for a variety of rating scenarios and sources of variation in ratings or scores (Brennan, 2001; Cronbach et al., 1972; Hoyt, 1941; Putka & Sackett, 2010).

The decomposition of the sources of error allows researchers to understand how observed scores generalize across these different sources of error (e.g., generalize over items, generalize over raters). In contrast, classical methods typically limit researchers to examining generalizability across one source of error such as items (e.g., internal consistency) or time (e.g., test–retest). These methods do not support generalization across the other sources of error that are simultaneously impacting an observed score (DeShon, 2001).

To understand how the sources of variance can be decomposed, consider a scenario of a structured interview with multiple raters, multiple items, and multiple candidates. In this scenario, there are several sources that could contribute to the variance in the observed scores, including raters (job candidates), raters (interviewers/assessors), items (individual structured interview questions or assessment center exercises), and the various interactions between these sources. In this scenario, the generalizability theory model for the variance in ratings across raters, items, and candidates can be shown in Equation 1:

\[
\sigma^2_{pri} = \sigma^2_p + \sigma^2_i + \sigma^2_r + \sigma^2_{pr} + \sigma^2_{pi} + \sigma^2_{ri} + \sigma^2_{pri.e}
\]

where \(\sigma^2_{pri} \) is the total variation in the observed scores, \(\sigma^2_p \) is ratee main effect; \(\sigma^2_i \) is the rater main effect; \(\sigma^2_r \) is the item main effect; \(\sigma^2_{pr} \) is the ratee × rater interaction; \(\sigma^2_{pi} \) is the ratee × item interaction; \(\sigma^2_{ri} \) is the rater × item interaction; and \(\sigma^2_{pri.e} \) is residual variations after accounting for the other sources. These specific components are estimated using a variance components analysis.\(^1\)

The specific components contributing to a reliability estimate will depend on the rating design and the specific sources of variation across which one wishes to generalize (Cronbach et al., 1972; Putka & Hoffman, 2014; Putka et al., 2008). Three common rating designs in assessment contexts are fully crossed, partially nested, and ill structured. The structure of each rating design is shown in Figure 1. A fully crossed rating design involves all raters rating all candidates on all items or exercises. A partially nested rating design involves different sets of raters rating different sets of candidates or may rate different sets of items/exercises.

\(^1\) A variance components analysis is a standard routine in most statistics platforms (e.g., SPSS, SAS, R).
One such situation could be when half of the candidates completing a structured interview are assessed by the same two raters but the other half of the candidates are assessed by two different raters. An ill-structured rating design is one where the ratings are neither nested nor crossed. A common case in which this approach is employed is a rotational ratings panel where different pairings or combinations of raters rate the candidates or items/exercises. For example, Raters A and B might assess Candidates 1 and 2 on Exercise Z, then Raters B and C assess Candidates 2 and 3 on Exercise Y, and Raters C and D assess Candidates 3 and 4 on Exercise X. Unlike a partially nested rating design, the sets of raters are not unique and can overlap over candidates or items/exercises.

Depending on the rating design and desired generalizations, specific components in the formula above are modified, included, or excluded to compute the reliability estimate. Using the generalizability theory approach, one computes a generalizability coefficient as an index of the generalizability of the scores on a measure across the specified sources of error. In a structured interview, for example, the reliability coefficient for a fully crossed rating design (i.e., all raters rate all candidates on all items) when one wishes to generalize across raters and items would be estimated using the following formula:

$$r^{2} = \frac{\sigma_{p}^{2}}{\sigma_{p}^{2} + \sigma_{r}^{2} + \sigma_{i}^{2} + \sigma_{r,p}^{2}}$$  \hspace{1cm} (2)

where $n_r$ is the total number of raters and $n_i$ is the number of items. This reliability coefficient is in the form of an intraclass correlation (ICC) for consistency in ratings and can be compared to an interrater correlation.

Continuing with an example of a structured interview, a variation of Equation 2 can be used to estimate reliability when the rating design is partially nested, as shown in Equation 3:

$$r^{2} = \frac{\sigma_{p}^{2}}{\sigma_{p}^{2} + \sigma_{r}^{2} + \sigma_{r,p}^{2}}$$  \hspace{1cm} (3)

where $n_{r,p}$ is the number of raters rating each candidate and $\sigma_{r,p}^{2}$ is variance associated with raters nested in candidates.

In the case of a structured interview with an ill-structured rating design, reliability can be estimated using Equation 4 (see Putka et al., 2008).
Invited Article

The image contains a page of text discussing reliability estimates in personnel assessments, focusing on generalizability theory and its application. The text includes mathematical equations and formulas, with a specific equation for estimating reliability. The page references multiple studies and authors, such as LoPilato et al., Putka et al., and others, and discusses the limitations and advantages of using generalizability theory in various contexts, including rating candidate pairs and assessing inter-rater reliability. The page aims to provide a clear articulation of the contributions of reliability estimates in personnel assessments.
Creating Test Score Bands

This difference in reliability estimates can serve to increase the width of the test score band. Further, when reliability is estimated using a generalizability theory approach, the values could be even lower, as more of the systematic sources contributing to the variation in the ratings are statistically modeled. Thus, it is possible that the range of scores inside a band would be larger than what is typically seen with written tests.

For example, common practice when using traditional approaches to reliability estimation would be to create bands of 2 standard errors of the difference (SEDs). If the reliability estimate was 0.45 and the standard deviation of the test was 2.30 standard deviations, the SED is 2.41 when rounded to two decimal places. If the highest candidate score in an assessment was a z-score of 2.50, a 2 SED band includes z-scores from 2.50 to -2.32 in the first band. In other words, a 2 SED band would include scores that range from two and a half standard deviations above the mean to two and one-third standard deviations below the mean. This band is also likely to include a large percentage of the candidates completing the tests.

Thus, it may be necessary to consider other possible strategies for determining band widths when using banding with ratings data. A 1 SED band in this example includes only z-scores from 2.50 to 0.09 in the first band. Consistent with best practice to construct bands that are not too wide, and to avoid equating candidates who are differentiated in terms of their scores, we advocate the use of 1 SED bands with ratings data (Guion, 2004). The use of 1 SED bands, given the likely lower reliability estimates that result from this more appropriate approach to estimating reliability, will allow for the creation of bands that include similar scores and do not include the majority of the candidates in the first band. Of course, other testing contexts could yield different strategies for calculating band widths—in our work with structured interviews with multiple raters, 1 SED bands have proven to yield useful information in selecting candidates.

Example of Test Score Banding of Interview Ratings Using Generalizability Theory

We now offer a demonstration of constructing test score bands for structured interview ratings using generalizability theory estimates of reliability and compare these results to those that would have been obtained using traditional methods of estimating reliability. The data in this example come from the structured interviews for 37 different jobs that were part of the hiring process of a municipal government. The 37 jobs varied in their level (e.g., entry-level jobs with minimal qualifications, senior managerial positions) and their functions (e.g., jobs involving the paving and maintenance of roads, engineering jobs, trades jobs, IT jobs). Across all of the jobs, fully crossed, partially nested, and ill-structured rating designs were used in the assessment of the interviews.

The interviews were highly structured (e.g., highly structured rating benchmarks, computer delivery of interview questions, extensive rater training) and were developed from a structured job analysis and content validation strategy. Thus, a caveat with this example is that the high level of structure in the interview process and assessments could have impacted the observed reliabilities and amount of variance attributed to each source. Research applying the approach used here with less structured interview data would be a valuable replication.

The number of candidates per job ranged from 4 to 46. The number of items per interview ranged from 7 to 13. For all of the jobs, each candidate was assessed by two raters. To compute the traditional interrater reliability estimates, the ratings were averaged across items for each candidate for each rater. These averages were used to compute the Pearson correlation and the intraclass correlation between the two raters’ average ratings on each candidate. Although many have argued that these estimates are not appropriate to use with partially nested and ill structured rating designs (e.g., Putka et al., 2008), we include them as a point of comparison.

The results of these analyses are shown in Table 1. As can be seen in the table, the reliability estimates from the traditional methods are consistently higher than the generalizability theory estimates, with the traditional ICC generally showing the highest reliability estimate. In this sample of jobs, there is only one instance where a traditional reliability estimate is lower than the generalizability theory estimate (job 27). Although there may be many reasons why the reliability estimates for the traditional methods are generally higher, including that the traditional estimates are not appropriate for the rating designed used with many of these jobs (see DeShon, 2001; Murphy & Deshon, 2000a; Putka & Hoffman, 2014; Putka et al., 2008), it is worth highlighting that the traditional methods fail to account for the candidate by item interactions that were a considerable source of variance in the rating for this sample. In other words, the candidates completing these interviews displayed differential performance across the items (i.e., scoring high on some items, but low on others). This pattern was likely a result of the broad KSA coverage of the interview items.

The implication of the reliability differences for the width of the test score bands is considerable. Generally, the SED (which in part determines score bands width) in this sample produced by the traditional methods were 40%–50%
### TABLE 1.
Summary of Results by Job and Rating Design

<table>
<thead>
<tr>
<th>Job</th>
<th>Design</th>
<th>G-theory</th>
<th>Traditional ICC</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reliability</td>
<td>SED</td>
<td>Reliability</td>
<td>SED</td>
</tr>
<tr>
<td>Job 1</td>
<td>Fully crossed</td>
<td>.869</td>
<td>3.078</td>
<td>.962</td>
</tr>
<tr>
<td>Job 2</td>
<td>Fully crossed</td>
<td>.284</td>
<td>4.947</td>
<td>.751</td>
</tr>
<tr>
<td>Job 3</td>
<td>Fully crossed</td>
<td>.261</td>
<td>4.058</td>
<td>.835</td>
</tr>
<tr>
<td>Job 4</td>
<td>Fully crossed</td>
<td>.910</td>
<td>3.303</td>
<td>.944</td>
</tr>
<tr>
<td>Job 5</td>
<td>Fully crossed</td>
<td>.336</td>
<td>4.533</td>
<td>.908</td>
</tr>
<tr>
<td>Job 6</td>
<td>Fully crossed</td>
<td>.574</td>
<td>4.223</td>
<td>.948</td>
</tr>
<tr>
<td>Job 7</td>
<td>Fully crossed</td>
<td>.619</td>
<td>3.984</td>
<td>.960</td>
</tr>
<tr>
<td>Job 8</td>
<td>Fully crossed</td>
<td>.777</td>
<td>4.027</td>
<td>.949</td>
</tr>
<tr>
<td>Job 9</td>
<td>Fully crossed</td>
<td>.417</td>
<td>5.022</td>
<td>.941</td>
</tr>
<tr>
<td>Job 10</td>
<td>Fully crossed</td>
<td>.780</td>
<td>4.000</td>
<td>.954</td>
</tr>
<tr>
<td>Job 11</td>
<td>Fully crossed</td>
<td>.781</td>
<td>3.395</td>
<td>.958</td>
</tr>
<tr>
<td>Job 12</td>
<td>Fully crossed</td>
<td>.619</td>
<td>3.984</td>
<td>.991</td>
</tr>
<tr>
<td>Job 13</td>
<td>Ill structured</td>
<td>.560</td>
<td>5.227</td>
<td>.875</td>
</tr>
<tr>
<td>Job 14</td>
<td>Ill structured</td>
<td>.815</td>
<td>3.250</td>
<td>.947</td>
</tr>
<tr>
<td>Job 15</td>
<td>Ill structured</td>
<td>.683</td>
<td>3.180</td>
<td>.886</td>
</tr>
<tr>
<td>Job 16</td>
<td>Ill structured</td>
<td>.683</td>
<td>4.683</td>
<td>.967</td>
</tr>
<tr>
<td>Job 17</td>
<td>Ill structured</td>
<td>.826</td>
<td>3.020</td>
<td>.951</td>
</tr>
<tr>
<td>Job 18</td>
<td>Ill structured</td>
<td>.933</td>
<td>1.073</td>
<td>.982</td>
</tr>
<tr>
<td>Job 19</td>
<td>Partially nested</td>
<td>.800</td>
<td>4.127</td>
<td>.939</td>
</tr>
<tr>
<td>Job 20</td>
<td>Partially nested</td>
<td>.853</td>
<td>3.374</td>
<td>.939</td>
</tr>
<tr>
<td>Job 21</td>
<td>Partially nested</td>
<td>.766</td>
<td>3.341</td>
<td>.889</td>
</tr>
<tr>
<td>Job 22</td>
<td>Partially nested</td>
<td>.798</td>
<td>3.738</td>
<td>.978</td>
</tr>
<tr>
<td>Job 23</td>
<td>Partially nested</td>
<td>.811</td>
<td>3.306</td>
<td>.936</td>
</tr>
<tr>
<td>Job 24</td>
<td>Partially nested</td>
<td>.755</td>
<td>3.936</td>
<td>.960</td>
</tr>
<tr>
<td>Job 25</td>
<td>Partially nested</td>
<td>.804</td>
<td>3.558</td>
<td>.940</td>
</tr>
<tr>
<td>Job 26</td>
<td>Partially nested</td>
<td>.780</td>
<td>3.206</td>
<td>.947</td>
</tr>
<tr>
<td>Job 27</td>
<td>Partially nested</td>
<td>.553</td>
<td>5.741</td>
<td>.651</td>
</tr>
<tr>
<td>Job 28</td>
<td>Partially nested</td>
<td>.671</td>
<td>4.098</td>
<td>.952</td>
</tr>
<tr>
<td>Job 29</td>
<td>Partially nested</td>
<td>.784</td>
<td>3.720</td>
<td>.855</td>
</tr>
<tr>
<td>Job 30</td>
<td>Partially nested</td>
<td>.780</td>
<td>0.151</td>
<td>.938</td>
</tr>
<tr>
<td>Job 31</td>
<td>Partially nested</td>
<td>.640</td>
<td>2.700</td>
<td>.977</td>
</tr>
<tr>
<td>Job 32</td>
<td>Partially nested</td>
<td>.830</td>
<td>3.794</td>
<td>.939</td>
</tr>
<tr>
<td>Job 33</td>
<td>Partially nested</td>
<td>.383</td>
<td>4.614</td>
<td>.952</td>
</tr>
<tr>
<td>Job 34</td>
<td>Partially nested</td>
<td>.791</td>
<td>2.979</td>
<td>.972</td>
</tr>
<tr>
<td>Job 35</td>
<td>Partially nested</td>
<td>.616</td>
<td>3.610</td>
<td>.977</td>
</tr>
<tr>
<td>Job 37</td>
<td>Partially nested</td>
<td>.727</td>
<td>3.38</td>
<td>.963</td>
</tr>
</tbody>
</table>
smaller than those produced by the generalizability theory method. Thus, score bands based on these SEDs would produce differences in the number of candidates in each score band and thus differences in the band into which the candidates would fall. Depending on the perspective one takes on the differing opinions of test score bands (cf. Schmidt, 1991; Zedeck, Outtz, Cascio, & Goldstein, 1991), this may be seen as a positive or negative. Regardless of viewpoint, if the reliability estimates from the traditional methods are not appropriate for many rating designs and they are less accurate (especially when there are a small number of raters), then the SED and the ultimate score bands in this sample produced by the traditional methods are too small. Thus, candidates could be placed in bands that are lower than where they should be placed. In turn, there would be candidates that should have been considered for employment but would not have been under the traditional approach.

Although this example application shows how the use of generalizability theory to estimate reliability can increase the utility of test score bands, it is also important to recognize that there may be situations where it decreases the utility of score bands or even makes them useless (Putka et al., 2008). For example, in situations where there are large item effects leading to low reliability estimates or a large standard deviation for the assessment, the score bands could become so wide that most candidates end up in the first band. We have advocated the use of 1 SED bands to address this possibility. However, it is still possible that a 1 SED band may be very large and include most candidates, which would limit the usefulness of the banding procedure. In a situation like this, one may need to choose a different band width (e.g., 0.5 band width) or choose not to use banding. Thus, the use of score banding with generalizability theory estimates of reliability should be used thoughtfully and judiciously to ensure that the improved accuracy in reliability does not negate the utility.

Conclusion

In this paper, we consider the role of reliability estimation in constructing test score bands for assessments involving ratings. Drawing on the findings and best practices from the literature, we advocate a generalizability theory approach to estimating reliability of ratings. Using a sample of data from structured interviews for municipal government jobs, we find that there are considerable differences in the SED depending the method used to estimate reliability.

Throughout this paper, we have primarily advanced an argument based on accuracy and appropriateness of the reliability estimate. There are, of course, other reasons to consider the proposed generalizability theory-based approach, including utility. As noted previously, in the traditional approach it is possible for assessor bias to be hidden – and in fact to yield enhanced estimates of interrater reliability – if the assessors share a bias. For example, two assessors who each believe that a particular job is not appropriate for women will likely each assign lower than merited scores to female candidates, and that bias will not be identifiable in the traditional approach. It is identifiable in the current approach, however, allowing decisions to made about who participates in ratings or whether additional training is needed. Further, by being able to break sources of unreliability into multiple components, problematic items (e.g., items that are more difficult than intended or that are ambiguous to candidates) can also be identified and addressed.

Though we have couched our argument in the context of structured interviews, the same arguments hold true for assessment centers in which various exercises are assessed by multiple assessors. The same concerns about accuracy of band widths, identifying rater tendencies, and identifying item/exercise issues are each relevant in many assessment contexts.

We fully recognize that there is controversy around the use of banding procedures. However, banding is in fact a frequently used approach, especially in municipal/governmental selection settings. As such, we believe that if banding is being employed, it should be employed as accurately as possible. Given our findings, we believe that the generalizability theory-based approach presented here provides the best opportunity to do so.

REFERENCES


RECEIVED 11/30/17 ACCEPTED 12/20/17