PERSONNEL ASSESSMENT AND DECISIONS

Personnel Assessment and Decisions

Volume 4 | Issue 1 Article 1

4-13-2018

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, Charles.Scherbaum@baruch.cuny.edu

Marcus Dickson
Wayne State University, ad4795@wayne.edu

Elliott Larson
Baruch College and Graduate Center, City University of New York, elliott.c.larson@gmail.com

Brian Bellenger
Personnel Board of Jefferson County, Brian.Bellenger@pbjcal.org

Kenneth Yusko
University of Maryland, kyusko@umd.edu

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*: Vol. 4 : Iss. 1 , Article 1.
DOI: 10.25035/pad.2018.001
Available at: https://scholarworks.bgsu.edu/pad/vol4/iss1/1

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[1], Marcus Dickson[2], Elliott Larson[1], Brian Bellenger[3], Kenneth Yusko[4], and Harold Goldstein[1]

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

KEYWORDS

selection, generalizability theory, banding, reliability

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.

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 are 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 & DeShon, 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

Corresponding author:
Charles Scherbaum
Email: Charles.Scherbaum@baruch.cuny.edu

al., 2008). 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 interrater 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 ratees (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 ratees, raters, and items is presented in Equation 1:

$$\sigma^2_{pri} = \sigma^2_p + \sigma^2_r + \sigma^2_i + \sigma^2_{pr} + \sigma^2_{pi} + \sigma^2_{ri} + \sigma^2_{pri.e} \qquad (1)$$

where σ^2_{pri} is the total variation in the observed scores, σ^2_p is ratee main effect; σ^2_r is the rater main effect; σ^2_i is the item main effect; σ_{pr} is the ratee × rater interaction; σ^2_{pi} is the ratee × item interaction; σ^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:

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \left[\frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pri,e}^2}{n_i n_r}\right]} \quad (2)$$

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:

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \left[\frac{\sigma_{r:p}^2}{n_{r:p}} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{ri,pri,e}^2}{n_i n_{r:p}}\right]} \quad (3)$$

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).

FIGURE 1.

Examples of Different Rating Designs

Fully crossed rating design

Candidate	Rater A	Rater B
1	X	X
2	X	X
3	X	X
4	X	X
5	X	X
6	X	X
7	X	X

Partially nested rating design

Candidate	Rater A	Rater B	Rater C	Rater D
1	X	X		
2	X	X		
3	X	X		
4			X	X
5			X	X
6			X	X
7			X	X

Ill-structured rating design

Candidate	Rater A	Rater B	Rater C	Rater D
1	X	X		
2		X	X	
3	X		X	
4	X		X	
5			X	X
6		X		X
7	X			X

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \left[q\sigma_r^2 + q\frac{\sigma_{ri}^2}{n_i} + \frac{\sigma_{pr}^2}{n_{r:p}} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pri.e}^2}{n_i n_{r:p}} \right]} \quad (4)$$

In Equation 4, q is a multiplier to scale the rater variance and is calculated as follows:

$$q = \frac{1}{\hat{k}} - \frac{\sum_i \sum_{i'} \frac{c_{i,i'}}{k_i k_{i'}}}{n_t(n_t - 1)} \quad (5)$$

where $\hat{k}$ is average number of raters per candidate, n_t is the number of candidates, $c_{i,i'}$ is the number of raters that each pair of candidates share, k_i is the number of raters rating candidates one of a pair, and $k_{i'}$ is the number of raters rating candidate two of a pair.

Generalizability theory is by no means new (e.g., Cronbach et al., 1972), but its application in personnel assessment has been limited (DeShon, 2001, LoPilato, Carter & Wang, 2015; Putka et al., 2008). A major advantage of generalizability theory is that it yields a clear articulation of the sources of variance that impact a set of ratings. Thus, hypotheses about which sources of variance have the largest relative contribution can be examined. For example, in circumstances where there is a concern that raters may be systematically and differentially rating candidates (e.g., systematically rating minority or female candidates lower), the main effect of raters and the interaction between raters and candidates can be examined to determine if these types of rating patterns are having an impact that is relatively larger than other potential sources of variance.

An additional benefit concerns the accuracy of the reliability estimates. The research comparing reliability estimates computed from ratings data generally finds that estimates from generalizability theory methods are more accurate than traditional methods, especially when there are a smaller number of raters or candidates and the rating designs are not fully crossed (e.g., Putka et al., 2008). These situations are often encountered in selection processes for many government and corporate jobs. Recent work has developed Bayesian approaches to estimating generalizability theory coefficients that hold promise to improve estimation in a variety of situations (LoPilato et al., 2015). The major disadvantages include the complexity of the terminology and accurately estimating the coefficients for the type of data (incomplete, unbalanced) that assessment practitioners and researchers often encounter (DeShon, 2001; Putka & Hoffman, 2014; Putka et al., 2008).

Estimating Reliability for Test Score Banding With Assessments Involving Ratings Using Generalizability Theory

Thus far, we have advanced a generalized argument for the merits of generalizability theory in the estimation of reliability in personnel assessments involving ratings. To further the argument, we now turn to a specific example of the application of this approach: test score banding. Test score banding is a method of using scores or ratings that creates groups of scores that are not statistically significantly different from one another (Cascio, Outtz, Zedeck, & Goldstein, 1991). The groups of scores are created using a confidence interval anchored on the highest score. Candidates within the confidence interval are treated as statistically equivalent.

The literature on banding has identified several approaches to create test score bands (Aguinis, Cortina, & Goldberg, 1998; Hanges & Gettman, 2004; Hanges, Grojean, & Smith, 2000). In our work, as is the case with many assessment professionals, we have used what is often referred to as the "classic" or "traditional" approach, which is based on the reliability of the test scores and the standard error of the difference (SED; Cascio et al., 1991). This approach is represented mathematically in Equations 6 and 7:

$$\text{Band Width} = C * \text{SED} \quad (6)$$

$$\text{SED} = S_x * \sqrt{1 - r_{XX}} * \sqrt{2} \quad (7)$$

where C is the z-score associated with the desired confidence level (used to set the number of SEDs included in the band width), S_x is the standard deviation of the test scores, and r_{XX} is the reliability of the scores.

A key challenge, however, is that the majority of the research literature on test score banding either implicitly or explicitly assumes a reliability model appropriate for written tests such as job knowledge tests (e.g., Alpha, KR-20). In the case of ratings, this model of reliability is not appropriate. Models of reliability based on multiple raters, such as generalizability theory are needed.

It is important to note that the values of the reliability estimates for ratings data (e.g., scores on interviews) can be considerably lower than the reliability estimates typically seen and expected of data from written tests. Although the appropriateness of interrater correlations is heavily debated (e.g., Putka et al., 2008; Murphy & DeShon, 2000a; 2000b; Schmidt et al., 2000), the meta-analytic estimate of these estimates of interrater reliability for structured interviews with independent ratings is 0.61 or less (Huffcutt, Culbertson & Weyhrauch, 2013). These numbers are far less than the values of coefficient Alpha or KR-20 on written tests that are typically greater than 0.70.

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[2].

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%

2 These jobs were a sample of all of the interview processes administered over a 3-year period. Jobs for which there were less than four candidates or jobs that used multiple selection components are not included in this sample.

TABLE 1.

Summary of Results by Job and Rating Design

Job	Design	G-theory		Traditional ICC		Pearson correlation	
		Reliability	SED	Reliability	SED	Reliability	SED
Job 1	Fully crossed	.869	3.078	.962	1.667	.958	1.745
Job 2	Fully crossed	.284	4.947	.751	2.912	.681	3.299
Job 3	Fully crossed	.261	4.058	.835	1.917	.742	2.398
Job 4	Fully crossed	.910	3.303	.944	2.620	.985	1.351
Job 5	Fully crossed	.336	4.533	.908	1.682	.833	2.271
Job 6	Fully crossed	.574	4.223	.948	1.480	.910	1.939
Job 7	Fully crossed	.619	3.984	.960	1.286	.930	1.706
Job 8	Fully crossed	.777	4.027	.949	1.926	.912	2.530
Job 9	Fully crossed	.417	5.022	.941	1.598	.894	2.141
Job 10	Fully crossed	.780	4.000	.954	1.871	.913	2.572
Job 11	Fully crossed	.781	3.395	.958	1.484	.932	1.892
Job 12	Fully crossed	.619	3.984	.991	0.612	.995	0.456
Job 13	Ill structured	.560	5.227	.875	2.783	.779	3.703
Job 14	Ill structured	.815	3.250	.947	1.745	.900	2.393
Job 15	Ill structured	.683	3.180	.886	1.908	.802	2.517
Job 16	Ill structured	.683	4.683	.967	1.501	.938	2.071
Job 17	Ill structured	.826	3.020	.951	1.606	.906	2.220
Job 18	Ill structured	.933	1.073	.982	0.557	.965	0.775
Job 19	Partially nested	.800	4.127	.939	2.269	.893	3.016
Job 20	Partially nested	.853	3.374	.939	2.168	.893	2.882
Job 21	Partially nested	.766	3.341	.889	2.301	.809	3.016
Job 22	Partially nested	.798	3.738	.978	1.236	.960	1.666
Job 23	Partially nested	.811	3.306	.936	1.920	.898	2.430
Job 24	Partially nested	.755	3.936	.960	1.586	.925	2.180
Job 25	Partially nested	.804	3.558	.940	1.967	.898	2.566
Job 26	Partially nested	.780	3.206	.947	1.579	.914	2.003
Job 27	Partially nested	.553	5.741	.651	5.071	.483	6.172
Job 28	Partially nested	.671	4.098	.952	1.572	.931	1.876
Job 29	Partially nested	.784	3.720	.855	3.046	.827	3.329
Job 30	Partially nested	.780	0.151	.938	0.077	.885	0.106
Job 31	Partially nested	.640	2.700	.977	0.679	.958	0.925
Job 32	Partially nested	.830	3.794	.939	2.266	.887	3.090
Job 33	Partially nested	.383	4.614	.952	1.291	.908	1.780
Job 34	Partially nested	.791	2.979	.972	1.088	.953	1.413
Job 35	Partially nested	.616	3.610	.977	0.874	.956	1.222
Job 36	Partially nested	.870	3.253	.939	2.225	.889	3.006
Job 37	Partially nested	.727	3.38	.963	1.242	.930	1.710

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

Aguinis, H., Cortina, J. M., & Goldberg, E. (1998). A new procedure for computing equivalence bands in personnel selection. Human Performance, 11, 351-365.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

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

Cascio, W., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. Human Performance, 8, 233-264.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York, NY: John Wiley.

DeShon, R. P. (2001). Generalizability theory. In F. Drasgow & N. Schmitt (Eds.), Advances in measurement and data analysis. San Francisco, CA: Jossey-Bass.

Guion, R. (2004). Banding: Background and general management purpose. In H. Aguinis (Ed.) Test score banding in human resource selection: Technical, legal, and societal issues. Westport, CT: Praeger Publishers.

Hanges, P. J., & Gettman, H. J. (2004). A comparison of test-focused and criterion-focused banding methods: Back to the future? In H. Aguinis (Ed.) Test score banding in human resource selection: Technical, legal, and societal issues. Westport, CT: Praeger Publishers.

Hanges, P. J., Grojean, M. W., & Smith, D. B. (2000). Bounding the concept of test banding: Reaffirming the traditional approach. Human Performance, 13, 181-198.

Hoyt, C. (1941). Test reliability obtained by analysis of variance. Psychometrika, 6, 153-160.

Huffcutt, A. I., Culbertson, S., & Weyhrauch, W. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. International Journal of Selection and Assessment, 21, 264-276.

LoPilato, Carter, & Wang (2015). Updating Generalizability Theory in Management Research: Bayesian Estimation of Variance Components. Journal of Management, 41, 692 – 717.

Murphy, K. R. (1994). Potential effects of banding as function of test reliability. Personnel Psychology, 47, 477-495.

Murphy, K. R., & De Shon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. Personnel Psychology, 53(4), 873-900.

Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? Personnel Psychology, 53(4), 913-924.

Putka, D. J., & Hoffman, B. J. (2014). "The" reliability of job performance ratings equals 0.52. In C. E. Lance, & R. J. Vandenberg (Eds.), More statistical and methodological myths and urban legends. New York, NY: Taylor & Francis.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins, (Eds.), Handbook of employee selection (pp. 9-49). New York, NY: Routledge.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. Journal of Applied Psychology, 93, 959-981.

Schmidt F. L. (1991). Why all banding procedures are logically flawed. Human Performance, 4, 265-277.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. Personnel Psychology, 53(4), 901-912.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Thousand Oaks, CA: Sage.

Society for Industrial and Organizational Psychology (2003). Principles for the validation and use of personnel selection procedures. Bowling Green, OH: Author.

Zedeck, S., Outtz, J., Cascio, W. F., & Goldstein, I. L. (1991). Why do "testing experts" have such limited vision? Human Performance, 4, 297-308.

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