A Comparison of Scoring Strategies for Operational Situational Judgment Tests

Bryon H. Miller
Ford Motor Company

Calvin C. Hoffman
Los Angeles County Sheriff's Department

Carlos Valle
Los Angeles County Sheriff's Department

The use of situational judgment tests (SJTs) has increased recently. SJTs offer many advantages compared to other assessments, including validity, flexibility in administration, objective scoring, and reduced mean group differences. While use has increased (Roth, Bobko, & Buster, 2013), there is little consensus regarding SJT scoring (Bergman et al., 2006). This study used data from operational SJTs to evaluate and compare scoring strategies in terms of convergence, reliability, mean group differences, and candidate rank order. Results demonstrate that different scoring strategies can produce wildly different rank orders and can contribute to differences in reliability values and mean group test scores.

INTRODUCTION

The use of situational judgment tests (SJTs) in applied selection settings has increased in recent years. These assessments present job-related scenarios and candidates provide evaluative information about potential reactions to those scenarios. Although SJTs may capture responses with open-ended formats (Labrador & Christiansen, 2008), having predetermined response options is most common. These include response arrangements such as 'select the best option', 'select the best and worst option' (MacKenzie, Ployhart, Weekley, & Ehlers, 2009), rank the options, or rate the effectiveness of each option. Figure 1 provides an example of a rate the effectiveness of each response option.

SJTs also use different ways to instruct or prompt candidates to respond to items, including behavioral intention (*what would you do in this situation*) and knowledge (*what should you do in this situation*) instructions (Chan & Schmitt, 2002; McDaniel, Hartman, & Grubb, 2003). SJTs offer many advantages compared to other assessment tools, including a reasonable level of validity (*r* = .34; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001), candidate acceptance, flexibility in administration (paper and pencil, video, computer, etc.), objective scoring, and relatively smaller mean group differences compared to cognitive ability predictors (B/W *d* = 1.00; Nguyen, McDaniel, & Whetzel, 2005; Whetzel, McDaniel, & Nguyen, 2008). While there has been a tremendous increase in the number of organizations using SJTs (Campion, Ployhart, & MacKenzie, 2014), there is little consensus in the literature regarding methods for scoring them (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006).

FIGURE 1
SAMPLE SJT ITEM DESIGNED FOR FRONT-LINE MANAGERS

The IT department emailed about an employee who is online shopping at work. You are required to document this as a performance issue since it violates company policy. However, this is your best employee who delivers high quality work and never misses deadlines. Below is a list of actions that may be taken to handle this situation.

	Very Ineffective 1	Ineffective 2	Somewhat Ineffective 3	Somewhat Effective 4	Effective 5	Very Effective 6
	□	□	□	□	□	□
Ask the IT department to make an exception since it's your best employee.	□	□	□	□	□	□
Document it but let the employee know you do not agree with the policy.	□	□	□	□	□	□
Get more information from your employee before making a decision.	□	□	□	□	□	□
Ask another manager for advice on handling the situation.	□	□	□	□	□	□

There are multiple ways to score SJTs. Unlike dichotomously scored knowledge tests, SJTs do not have unambiguously or factually correct answers (Bergman, et al., 2006; Lievens, Peeters, & Schollaert, 2008). Scoring characteristics of SJTs are often comparable to those found in biodata literature in which scoring keys may be generated a priori using expert judgment, or developed on empirical or theoretical grounds (Bergman et al., 2006). The most common keying method is expert consensus, which compares candidate responses to the mean effectiveness ratings of subject matter experts. A less common alternative is peer consensus, which may be based on the responses of applicants, incumbents, supervisors of incumbents, or customers (McDaniel, Psotka, Legree, Yost, & Weekley, 2011).

SJTs keyed based on consensus profiles may contribute to mean group differences in test scores (McDaniel, et al., 2011). Research regarding rating behavior has found that Black and Hispanic groups are more likely than White participants to engage in extreme ratings (Bachman & O'Malley, 1984; Bachman, O'Malley, & Feedman-Doan, 2010). Conversely, Asian candidates appear to engage in extreme ratings at or below the rate of White candidates (Clark, 2000; Grandy, Bachman, et al., 2010). Although the tendency to engage in extreme ratings may be impacted in the short term via coaching (Cullen, Sackett, & Lievens, 2006), these rating patterns appear to be relatively stable over time (Lau, 2007). De Leng, Stegers-Jager, Husbands, Dowell, Born, and Themmen (2016) explored several different metrics for scoring an integrity-oriented SJT. All scoring metrics, including the scoring methods introduced to control scale-use-tendencies ("systematic error"), resulted in mean group differences favoring majority candidates (Dutch applicants to medical school).

While there is agreement that SJTs typically measure multiple construct domains (Campion, Ployhart, & MacKenzie, 2014), and will generally produce low levels of internal consistency reliability when using unidimensional estimates such as coefficient alpha, there is little agreement regarding appropriate methods for estimating SJT reliability. Even with evidence of multidimensionality, it is common practice for researchers to report alpha estimates for SJTs (Campion et al., 2014), and such values are sometimes distressingly low. Bergman et al. (2006) compared 11 methods for deriving scoring keys on a leadership oriented SJT and explored the validity and incremental validity of those keys. Although Bergman et al. went into a fair amount of detail in reporting validity information, they did not address the topic of SJT reliability. De Leng et al. (2016) calculated 28 scoring metrics using data from an SJT used for medical school screening. They reported that alpha coefficients for the same SJT ranged from a low of .34 to a high of .73 depending on the scoring metric applied. They found that moving from one scoring method to another altered the acceptability of the resulting reliability estimate, which they stated added to the limited usefulness of alpha coefficients for SJTs. The topic of SJT reliability clearly deserves further research.

Legree, Kilcullen, Psotka, Putka, and Ginter (2010) explored SJT keying strategies in detail, focusing their attention on strategies where examinees rate the effectiveness of all options using a Likert scale (rate-all). Legree et al. argued that the rate-all approach greatly increases the effective length of the test without increasing reading time. For example, if an SJT has 15 scenarios and examinees select the best answer, it is a 15-item test; if the same SJT has four alternatives per scenario and examinees rate the effectiveness of each alternative, it becomes a 60-item test. While the scoring of rate-all SJTs may increase the scoring complexity relative to select-the-best response SJTs, rate-all SJTs may have the potential to provide more information about examinee capabilities (Legree et al., 2010).

As Legree et al. (2010) explain, rate-all SJTs compare the profile of each examinee's ratings to the mean profile of the SME ratings (expert consensus key) or the mean of all examinees (peer consensus key). Profile comparisons may be made on the basis of raw distance, absolute distance (D), squared absolute distance (D^2), or metrics such as the correlation between candidate profile and key profile (Legree et al., 2010). Legree et al. referred to the correlation between the candidate profile and key profile as a C-score, which is parallel to within person standardized scoring method but on a different scale. The authors believe this methodology offers advantages over purely distance-based scoring strategies, such as D or D^2. Legree et al. argued that using C-score metrics may improve the construct validity of SJTs. Legree et al. also discussed elevation distance score (Eldis^2), which is the squared difference in elevation between each respondent profile mean and the scoring profile mean.

A possible limitation of the Legree et al. study is that the assessment they investigated, the Army leadership knowledge test (LKT), better fits the definition of an adjective checklist rather than an SJT. A second possible limitation is that participants who completed the LKT in the Legree et al. study were not job applicants but were instead active members of the military. Previous SJT research has shown that subgroup test performance (Weekley, Ployhart, & Harold, 2004), criterion validity, and construct validity can differ across incumbent and applicant populations (MacKenzie, Ployhart, Weekley, & Ehlers, 2010). The current study addresses these potential limitations by applying many of the scoring protocols described by Legree et al. to data from operational SJTs used in promotional batteries for two promotional progressions in a public sector/public safety organization.

If researchers use D^2 and Eldis^2 as defined, candidates who perform better will have lower scores (Legree et al., 2010). For example, the best score a candidate could obtain under D or D^2 is a score of 0, and the same is true of Eldis^2. At the same time, the C-score and dichotomous consensus score metrics produce a high score as the better score. Rather than keep track of negative and positive signs across the analyses reported here, particularly when referring to analyses of mean group differences, we reflected scores on distance based metrics, such as D^2 and Eldis^2, so a high value represents a better score. See Legree et al. (2010) for a discussion of the steps and rationale involved in reflecting SJT scoring metrics such as D and D^2.

The current study reanalyzed data from Miller (2015), who discussed and compared multiple strategies for creating SJT scoring keys, including standardized distance, dichotomous consensus (McDaniel et al., 2011), and a hybrid method Miller proposed combining standardized distance and dichotomous consensus. McDaniel et al. dichotomized 6- and 9-point scales and reported that this strategy was successful in reducing mean subgroup differences. Those researchers referred to this method as dichotomous consensus. Although the dichotomous consensus method reduced subgroup differences somewhat, it also reduced the granularity of the scale by converting either 6- or 9-point scales to a dichotomous scale. De Leng et al. (2016), reported that dichotomous consensus scoring produced the lowest reliability estimates (.34) and also reduced ethnic group differences.

Miller (2015) based his hybrid scoring approach on an expert scoring key that provided candidates credit for response pattern as well as distance. Candidates were awarded points for response pattern by comparing their responses with an expert key which was dichotomized from a 6-point scale into effective or ineffective categories. If candidates' effectiveness ratings were within the keyed category, they received credit.

In addition to the hybrid metric from Miller (2015), for this study we added metrics of D^2, C-score, and $Eldis^2$ as discussed and defined by Legree et al. (2010). A description of each scoring metric and its corresponding formula can be found in the appendix. We also examined reliability estimates resulting from the different scoring strategies and explored mean group differences for each scoring strategy. Because previous research has suggested that SJTs scored with Likert scales can contribute to mean differences for Black and Hispanic candidate groups, we also explore how using each of these metrics might impact effect sizes (*d*-values) for race, ethnic and sex group comparisons.

METHOD

Study participants were law enforcement promotional job candidates who had successfully completed a pass-fail job knowledge test (JKT) and subsequently completed an SJT. The JKT sampled approximately 25 job knowledge domains identified via a multi-method job analysis as important for successful performance as a sergeant in this agency. Candidates were required to pass the JKT before participating in the SJT, which was administered at a later date. This study examined data from two separate promotional paths for entry level supervisor positions. These include patrol and custody/corrections tracks labeled as Path I and Path II, respectively.

We designed unique SJTs for the two career tracks based on critical incidents written by SMEs. The item generation process produced over 70 scenarios with four to nine response options each. Although the objective was to develop eight scenarios with four response options per SJT, having an abundance of content provided the opportunity to capture the most valid content containing variance among response options and competencies. The SJTs targeted multiple competencies including planning and organizing, judgment and decision-making, initiative, stress tolerance, interpersonal relations, adaptability, integrity, and leadership. After editing of question stem (scenario) and response options, SMEs content validated each scenario and rated all response options in terms of relative effectiveness using a 6-point scale (where 1 = Very Ineffective and 6 = Very Effective). Options were evaluated in terms of descriptive statistics including mean and SD, and r_{wg} was used to guide retention or deletion of response options. The SME group rating means were used to establish the scoring key and to determine which response options to retain; response options with low levels of interrater agreement (large SD and hence low r_{wg}) were dropped from the final version of the two SJTs. Response scenarios having acceptable r_{wg} but little variation in mean effectiveness of response options, such as all options being rated a '4' on a 6-point scale, were dropped from the test's item pool.

Although the actual SJTs in this promotional process used several different item types, including 'choose the best answer', place the alternatives in the proper sequence, and 'rate-all' options, in this study we examined only the data for rate-all items. The SJTs used in these analyses were based on eight scenarios with four alternatives per scenario (equivalent to 32 questions). Following a process similar to that described by Chan and Schmitt (2002), candidates rated all response options on their relative effectiveness.

RESULTS

Table 1 provides descriptive statistics for all scoring metrics for each career path. Similar to the findings of Legree et al. (2010), D^2 was more reliable than the C-score, and both metrics produced higher reliability estimates than did the hybrid strategy. The lower reliability found for the hybrid strategy, which incorporates dichotomous consensus scoring (McDaniel et al., 2011), is consistent with the findings of De Leng et al. (2016), where dichotomous consensus scoring produced the lowest reliability estimates among the scoring methods those authors studied.

TABLE 1
DESCRIPTIVE STATISTICS FOR SJT SCORING METRICS

	Promotional Path I (N = 419)			Promotional Path II (N = 260)		
Scoring Metric[a]	M	SD	Reliability[b]	M	SD	Reliability[b]
Hybrid	42.01	11.37	.31	57.82	10.01	.14
C-score	0.64	0.13	.50	0.72	0.10	.36
D^2	5.15	0.83	.71	2.65	0.55	.59
$Eldis^{2b}$	1.81	0.26	.26	1.85	0.21	.50

Notes: a – see appendix for descriptions; b – reliability for $Eldis^2$ based on split-half corrected with Spearman-Brown formula; all other reliability estimates are based on Cronbach's alpha.

In absolute terms, the reliability estimates we found (see Table 1) using data from operational SJTs were substantially lower than Legree et al.'s results based on an adjective checklist (the Army LKT). Legree et al. reported reliability estimates ranging from .67 to .97 depending on the scoring metric in question and the particular data set (military rank of participant), while our estimates ranged from a low of .14 to a high of .71 depending on the SJT, scoring metric, and career track in question. There was also a substantial disparity in the magnitude of our reliability estimates when compared to the results reported by Legree et al. (2010), with reliability estimates for operational SJTs substantially lower than the values reported by those authors. Our results were consistent with those reported by De Leng et al. (2016), who reported alphas ranging from .34 to .73 depending on scoring metric using applicant data from an SJT used for screening candidates to medical school. Given these findings, it may be assumed the higher reliability estimates found by Legree et al. were a result of the instrument's design characteristics.

Table 2 reports intercorrelations among scores for the various metrics applied to the two operational SJTs. Intercorrelations between the hybrid, D^2, $Eldis^2$, and C-score metrics ranged from .18 to .80. Legree et al. reported intercorrelations between the various metrics (based on expert keys) of .35 to .91 for the LKT Trait scale, and between -.10 and .94 for the LKT Skill scale. In our operational SJTs, we found lower convergence between the various profile scoring metrics. This low level of convergence has implications regarding how the various scoring metrics would rank order candidates, and illustrates why choice of SJT scoring metric is an important decision in applied settings.

TABLE 2
INTERCORRELATIONS OF SJT SCORING METRICS: PROMOTIONAL PATH I AND II

Scoring Metric[b]	Hybrid	C-score	D^2	$Eldis^2$
Hybrid	--	.70	.59	.28
C-score	.74	--	.67	.20
D^2	.64	.80	--	.43
$Eldis^2$	.40	.18	.36	--

Notes: a – Promotional Path I candidates (N = 419) *below* diagonal; Promotional Path II candidates (N = 260) *above* diagonal; b – see appendix for descriptions.

Table 3 reports mean group differences (*d*) resulting from the use of different SJT scoring strategies. Values of *d* range between .01 to -.70 depending on the career tracks and scoring metrics being compared. Of potential concern in this organization is the finding that the $Eldis^2$ metric produced relatively large Female/Male group differences favoring males for Promotional Path I (*d* = -.70) and Promotional Path II (*d* = -.44) candidates. These differences suggest that Females are using the rating scales differently (larger observed mean differences) than males when compared to the SME key. The Female/Male differences in scale use were greater than Black/White or Hispanic/White differences. Whereas Female pattern/shape is consistent with Black candidates, their ratings occupy a different location on the rating scale which may be contributing to differences in overall measures of agreement (Legree et al., 2010). Females may be an under-represented group in many contexts, and metrics that may be associated with increased mean differences between males and females are potentially problematic in applied selection research.

TABLE 3
MEAN GROUP DIFFERENCES (*d*) RESULTING FROM SJT SCORING METRICS

Groups	Career Track	Hybrid	C-score	D^2	$Eldis^2$
Black/White	Path I	-.013	-0.21	-0.29	-0.36
	Path II	-0.31	-0.44	-0.28	-0.15
Hispanic/White	Path I	-0.13	-.023	-0.21	-0.28
	Path II	0.01	-0.34	-0.10	-0.03
Asian/White	Path I	-0.25	-0.39	-0.27	0.06
	Path II	-0.29	-0.44	-0.25	0.07
Female/Male	Path I	-0.26	-0.23	-0.41	-0.70
	Path II	-0.20	-0.18	-0.28	-0.44

Notes: Promotional Path I candidates (N = 419); Promotional Path II candidates (N = 260); a negative sign indicates White or Male subgroup outperformed comparison group.

As a means of illustrating how differently candidates were ranked using the various scoring metrics, Table 4 compares the magnitude of rank changes for both groups of candidates. We performed the rank order analysis by obtaining the rank of each case for each scoring method. Candidates with identical scores within a scoring method received the same rank. The absolute change in rank order of each candidate by each scoring comparison was then calculated. Scoring comparisons were done in pairs of scoring methods (e.g., C-score versus D^2). Finally, the average absolute rank order change across all candidates for each pair of metrics was calculated and is presented in Table 4. For Promotional Path I (*N* = 419) candidates, rank order changes varied from a low of 56 to a high of 123. For Promotional Path II (*N* = 260) candidates, rank order changes varied from a low of 49 to a high of 139. These substantial shifts in the rank order of candidates have important implications in applied settings, particularly for organizations using top-down ranking to guide decision making. In settings where scores are banded, these shifts in rank order would be somewhat less problematic but are still of concern.

TABLE 4
RANK ORDER CHANGES RESULTING FROM DIFFERENT SJT SCORING METRICS

Promotional Path I				
Scoring Metric[a]	Hybrid	C-score	D^2	$Eldis^2$
Hybrid	--			
C-score	68	--		
D^2	71	56	--	
$Eldis^2$	122	123	111	--
Promotional Path II				
Scoring Metric[a]	Hybrid	C-score	D^2	$Eldis^2$
Hybrid	--			
C-score	137	--		
D^2	139	49	--	
$Eldis^2$	122	80	66	--

Notes: Promotional Path I candidates (N = 419); Promotional Path II candidates (N = 260); a – see appendix for descriptions.

DISCUSSION

Legree et al. (2010) identified a number of advantages of SJTs that use the rate-all format relative to other formats. In particular, such an approach allows researchers and practitioners to make far better use of SJTs, and allows one to collect much more data from a specific number of SJT scenarios compared to the more commonly used "pick the best answer" approach. Legree et al. reported that correlations between the LKT score and variables such as soldier rank or personality inventory scores were higher using their proposed metrics compared to more commonly used metrics like D or D^2. The criteria used in Legree et al. reflect on the construct validity of the LKT, but do not speak to its predictive validity.

Although content validation methods are appropriate for developing SJT content, such methods do not provide unambiguous guidance in choosing between competing scoring rules. Miller, Hoffman, Fiorentino and Lopez (2016) argued that it would be preferable if SJT scoring strategies were based on some type of external criterion. Given the fact that different scoring metrics can result in substantial changes in candidate rank order, researchers and practitioners do not have sufficient guidance in choosing between SJT scoring strategies. We replicated the analysis used by Miller et al. (2016) and also found large disparities in how candidates were rank ordered (see Table 4) depending on the SJT scoring metric used.

The findings from the present research support Miller et al.'s reasoning and demonstrate that not only do candidate rank orders differ substantially across scoring metrics, but test score reliability estimates and mean subgroup differences can vary substantially depending on the scoring strategy used. Additional research into SJT scoring metrics is needed. As matters currently stand, there does not seem to be a consensus regarding appropriate SJT scoring metrics (Bergman et al., 2006; De Leng et al., 2016; Legree et al., 2010; Miller et al., 2016). Researchers and practitioners should exercise caution when choosing which scoring metric(s) to use in a given situation because of the differences we observed in reliability and mean group differences, the differences reported by De Leng et al. (2016), and because of the differences in empirical validity reported by Bergman et al. (2006).

The hybrid metric investigated here was included in this research as a means of attempting to reduce mean group differences in SJT scores. The hybrid metric tended to produce *d*-values slightly smaller than analogous values for the D^2 and C-score metrics. At the same time, the eldis^2 metric produced the largest male/female mean group differences observed in this study. While the hybrid approach may have had modest success in reducing mean group differences, it appears this metric also produced substantially lower reliability estimates than did the other metrics explored here. In all comparisons between scoring metrics for the two different samples, reliability was lowest for the hybrid approach. In reviewing Table 4, it also appears that the hybrid metric tends to rank order candidates much differently than the other metrics. It may be that the hybrid model sacrifices too much information about candidates in its quest to reduce mean group differences.

STRENGTHS AND LIMITATIONS

This study highlighted differences in reliability estimates across the scoring metrics explored, mean group differences and candidate ranking that vary based on the scoring metric used. This study provides additional support to challenge the use of internal consistency reliability estimates for SJTs when test-retest coefficients are unavailable for comparison. Test-retest strategies are often impractical in applied settings given that most candidates will not take a given test more than once. Although there have been efforts to improve the integrity of alpha by using a stratified approach (Catano, Brochu, & Lamerson, 2012), evidence to support the stratified approach is not substantial at this time. Another strength of the current study is that we replicated many of the profile metrics described in Legree et al.'s (2010) study, along with the inclusion of sex and ethnic group comparisons using data from operational SJTs completed by job applicants. The discrepancies in findings between these studies highlight a need to investigate how contextual factors may impact the operation of SJTs.

Despite its contributions to SJT research, the current study would have benefited from the addition of criterion-related validation techniques using actual job performance measures. Given the foundational limitations of content validation strategies for SJTs, additional analyses with outcome measures could provide better evidence of test validity and would have aided with interpretation. Lastly, having a test-retest research design would have increased inference strength by allowing the researchers to compare the stability of internal consistency estimates.

DIRECTIONS FOR FUTURE RESEARCH

The results of this study provide opportunities for expansion, redesign, and replication. Where possible, investigators should include criterion measures, which would allow for baseline validity comparisons across different scoring approaches. We also recommend designing and evaluating SJTs based on factor structure. Particular focus should be placed on the application of dimension-free internal consistency estimates, including factor analysis (Revelle & Zinbarg, 2009), nonlinear structural equation modeling (Green & Yang, 2009), and coefficient omega (ω; McDonald, 1999). Dimension-free estimates of internal consistency have been available for quite some time (Bentler, 1972) but are overshadowed by the acceptance and popularity of alpha, and the ease of use and general availability of SPSS compared to SEM software.

REFERENCES

Bachman, J. G., & O'Malley, P. M. (1984). Black-White differences in self-esteem: Are they affected by response styles? *American Journal of Sociology*, 624-639.

Bachman, J.G., O'Malley, P.M., & Freedman-Doan, P. (2010). *Response styles revisited: Racial/ethnic and gender differences in extreme responding* (Monitoring the Future Occasional Paper No. 72). Ann Arbor, MI: Institute of Social Research.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14(3), 223-235.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, 27(4), 283-310

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, 20(3), 333-346.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, 15(3), 233-254.

Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality*, 5(1), 137-152.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment*, 14(2), 142-155.

De Leng, W.E., Stegers-Jager, K.M., Husbands, A., Dowell, J.S., Born, M.P., & Themmen, A.P.N. (2016). Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Science Education*, 22, 243-265.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74(1), 121-135.

Lau, M.Y. (2007). Extreme response style: An empirical investigation of the effects of scale response format and fatigue (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3299156)

Legree, P. J., Kilcullen, R., Psotka, J., Putka, D., & Ginter, R. N. (2010). *Scoring situational judgment tests using profile similarity metrics* (No. ARI-TR-1272). Army Research Institute for the Behavioral and Social Sciences Arlington, VA.

MacKenzie Jr, W. I., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2010). Contextual effects on SJT responses: An examination of construct validity and mean differences across applicant and incumbent contexts. *Human Performance*, 23(1), 1-21.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology,* 60(1), 63-91.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*(2), 327-336. doi: 10.1037/a0021983

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

Miller, B.H. (2015). *An exploration of reliability estimates and scoring models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3712696)

Miller, B.H, Hoffman, C.C., Lopez, P.D., & Fiorentino, D. (2016). *Scoring situational judgment tests (SJTs): A demonstrated need for criterion data*. Presented as part of a symposium at the annual conference of the Society for Industrial and Organizational Psychology, Anaheim: CA.

Nguyen, N. T., & McDaniel, M. A. (2003). Response instructions and racial differences in a situational judgment test. *Applied HRM Research*, 8(1), 33-44.

Nguyen, N. T., McDaniel, M. A., & Whetzel, D. L. (2005). Subgroup differences in situational judgment test performance: A meta-analysis. In *20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles*.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

Roth, P. L., Bobko, P., & Buster, M. A. (2013). Situational judgment tests: The influence and importance of applicant status and targeted constructs on estimates of Black–White subgroup differences. *Journal of occupational and organizational psychology*, 86(3), 394-409.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance*, 17(4), 433-461.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21(3), 291-309.

APPENDIX

SJT SCORING METRICS

Measure	Description	Formula
Correlation score (C-score)	Product moment correlation between values in candidate response profile (X) and expert-based scoring rubric (K; Legree et al., 2010). Scoring rubric established using expert-based scoring key.	$r\ = 1 - \sum(Z_x - Z_k)^2\ /2(n-1)$
Mean squared distance (D^2)	Mean squared distance between scoring rubric (K) and each candidates' scores (X). Note: values were reflected so a higher score indicates better correspondence with scoring key.	$D^2 = \sum(x_i -\ k_i)^2\ /n$
Elevation distance score (Eldis^2)	Distance of the mean of the items for a candidate from the scoring rubric (K). Note: values were reflected so a higher score indicates better correspondence with scoring key.	$Eldif\ = \vert X_{mean} - K_{mean}\vert$
Hybrid	Hybrid method (Miller, 2015) combines McDaniel et al.'s (2011) dichotomous consensus (DC) and mean standardized distance. DC uses the raw item mean across respondents to determine if an item is effective or ineffective. Using a 6-point Likert scale, endorsements and group means above 3.5 were recoded as *Effective*; all lower values recoded as *Ineffective*. Candidates who correctly endorse effectiveness category receive +2 points; incorrect responses receive -2 points. Mean standardized distance *(D*$_{\text{stand}}$) is the distance between z-score conversions of scoring rubric (K) and candidates' scores (X).	*Hybrid* = *Dichotomous consensus*+ D_{stand} Dichotomous consensus scoring: Correct endorsement = +2 points Incorrect endorsement = -2 points $D_{\text{stand}} = \sum(x_i -\ k_i)^2\ /n$