Learning Accelerator Research Paper

Effects of Situational Judgment Test Format on Reliability and Validity

Michelle P. Martin-Raugh
Cristina Anguiano-Carrasco
Teresa Jackson
Meghan W. Brenneman
Lauren Carney
Patrick Barnwell
Jonathan Kochert

2018

Martin-Raugh, M.P., Anguiano-Carrasco, C., Jackson, T., Brenneman, M.W., Carney, L., Barnwell, P., & Kochert, J. (2018). Effects of SJT format on reliability and validity. *International Journal of Testing, 18*, 135-154.

This is a draft of Effects of Situational Judgment Test Format on Reliability and Validity and the copy of record is with International Journal of Testing (DOI: 10.1080/15305058.2018.1428981)

Effects of Situational Judgment Test Format on Reliability and Validity

Michelle P. Martin-Raugh, Cristina Anguiano-Carrasco, Teresa Jackson, Meghan W. Brenneman, Lauren Carney, Patrick Barnwell, and Jonathan Kochert

Single-response situational judgment tests (SRSJTs) differ from multiple-response SJTs (MRSJTS) in that they present test takers with edited critical incidents and simply ask test takers to read over the action described and evaluate it according to its effectiveness. Research comparing the reliability and validity of SRSJTs and MRSJTs is thus far extremely limited. The study reported here directly compares forms of a SRSJT and MRSJT and explores the reliability, convergent validity, and predictive validity of each format. Results from this investigation present preliminary evidence to suggest SRSJTs may produce internal consistency reliability, convergent validity, and predictive validity estimates that are comparable to those achieved with many traditional MRSJTs. We conclude by discussing practical implications for personnel selection and assessment, and future research in psychological science more broadly.

Keywords: situational judgment tests (SJTs), reliability, validity, cross-cultural competence

Situational judgment tests (SJTs) are assessments that typically present respondents with job-related situations and ask them to indicate the best and worst way to respond by selecting from a list of plausible behavioral response alternatives. Research has repeatedly demonstrated SJTs can successfully predict job performance across a variety of domains, including customer service (Motowidlo, Brownlee, & Schmit, 2008), management (Motowidlo, Dunnette, & Carter, 1990), and military settings (Waugh, Putka, & Sager, 2002). SJTs have become a popular method of assessment as a result of their moderately strong predictive validity, traditionally lower levels of adverse impact when compared to cognitive ability measures, and generally favorable applicant reactions (Ployhart & MacKensie, 2011).

A relatively new SJT format, termed the single-response SJT (SRSJT; Motowidlo, Crook, Kell, & Naemi, 2009) differs from multiple-response SJTs (MRSJTs) in that it presents test takers with edited critical incidents, which are vignettes describing behavioral episodes of critical significance within a domain, that have had the outcome of the action being described removed. SRSJTs contain no response options; they simply ask test takers to read over the action described in the incident and evaluate it according to its effectiveness using a Likert scale. Participants' effectiveness ratings can then be compared to experts' effectiveness ratings for scoring purposes. An example of an MRSJT item and SRSJT item are displayed in Figure 1.

SRSJTs have demonstrated validity in predicting performance across a few different occupations, such as those of tour guides (r=0.33 [uncorrected]; Motowidlo, Martin, & Crook, 2013), physicians (r=0.20 [uncorrected]; Kell, Motowidlo, Martin, Stotts, & Moreno, 2014; r=0.25 [uncorrected]; Ghosh, Motowidlo, & Nath, 2015), and volunteers (r=0.28 [uncorrected]; Motowidlo et al., 2009; r=0.22 [uncorrected]; Crook et al., 2011). Meta-analytic evidence suggests that traditional MRSJTs produce uncorrected correlations of 0.20 with measures of job performance (McDaniel, Hartman, Whetzel, & Grubb, 2007). Thus, although investigations examining the validity of SRSJTs are still somewhat sparse, there is reason to believe the two SJT formats may predict job performance with similar accuracy.

However, empirical research directly comparing the reliability and validity of SRSJTs to MRSJTs is thus far extremely limited (see Crook, 2015 for one notable exception). The purpose of the study reported here is to directly compare forms of a single-response and multiple-response SJT designed to measure procedural knowledge about how to behave in military contexts that require interaction with a person or persons of a different culture, and to explore the reliability, convergent validity, and predictive validity of each format. We conclude by discussing practical implications for personnel selection and assessment, and future research in psychological science more broadly.

Panel A. Traditional Multiple-Response Situational Judgment Test (MRSJT) Item.	*Panel B. Single-Response Situational Judgment Test (SRSJT) Item (Ineffective).*
You are meeting with two elders from the host nation. One of them is retiring and the other one is his replacement. The three of you are discussing improvements to schools, roads, and health care, and the new elder immediately rejects what you are saying. He says that the United States is full of empty promises and that nothing will ever be completed. For each response option for each item, please rank order each of the five behavioral responses to the situation according to their effectiveness, where 1 = most effective and 5 = least effective. A. _____ Ask for the retiring elder'sopinion on the matter. If the retiring elder agrees with his replacement, ask how things can be improved and what kind of guarantees would make them feel more confident and comfortable. B. _____ Disagree with the new elderand tell him that the United States will prove themselves to him. C. _____ Tell the new elder to do justhis job and to not worry about the end result. D. _____ Ask the new elder how hewould handle the improvements. E. _____ Assure the new elder that theUnited States military does not work individually on projects and that you will try to employ his government.	A soldier is meeting with two elders from the host nation. One of them is retiring and the other one is his replacement. The three are discussing improvements to schools, roads, and health care, and the new elder immediately rejects what the soldier is saying. He says that the United States is full of empty promises and that nothing will ever be completed. The soldier told the elder to just do his job and to not worry about the end result. How effective is the soldier's behavior in this scenario? 1 = Very Ineffective 2 = Somewhat Ineffective 3 = Slightly Ineffective 4 = Slightly Effective 5 = Somewhat Effective 6 = Very Effective

FIGURE 1

Examples of MRSJT and SRSJT items. Examples of MRSJT and SRSJT items. Panel A shows an MRSJT item from this study where option A is the best response and option C is the worst response. Panel B shows an SRSJT item that describes ineffective behavior.

ADVANTAGES OF SINGLE-RESPONSE SJTS

One feature of SRSJTs that differentiates them from typical MRSJT formats is that each item can be characterized as being either particularly effective or particularly ineffective according to the mean of experts' effectiveness ratings of the items. This allows for participants' scores to be computed for effective items and ineffective items separately (see Crook et al., 2011; Martin-Raugh, Kell, & Motowidlo, 2016). There has been some evidence, although sparse, to suggest that these two classes of knowledge are not strongly correlated (Crook et al., 2011; Motowidlo et al., 2013), and consequently, may be considered distinct constructs

as opposed to two facets of one construct (Crook et al., 2011). For example, some research has shown that subscores for knowledge of effective behavior and ineffective behavior measured using an SRSJT were weakly correlated (r D 0.11 to r D 0.30; Crook et al., 2011), which is surprising given that the two scores were derived using the same methodology and scoring approach, a factor that has been known to artificially inflate correlations (Maul, 2013). Given that in many MRSJTs respondents are asked to evaluate both effective and ineffective response options (e.g., select the "best" and/or "worst" response) this notion runs counter to the assumption that evaluations of both effective and ineffective behaviors are indicators of the same underlying knowledge construct (Nguyen, Biderman, & McDaniel, 2005). Although a person may know that, for example, treating religious artifacts with irreverence when in a new culture is ineffective, that knowledge may not necessarily translate to a commensurate understanding of how to interact with people from a new culture in way that is effective, and vice versa. Perhaps because MRSJTs typically incorporate both effective and ineffective response options into the same items, the items are considered multidimensional, at least partially contributing to the generally low internal consistency reliability estimates often associated with many MRSJTs (McDaniel & Whetzel, 2007). In contrast, SRSJT alpha reliability estimates have been relatively high, ranging from 0.69 to 0.92 in the limited number studies conducted using SRSJTs (Crook et al., 2011; Kell et al., 2014; Martin-Raugh et al., 2016; Motowidlo et al., 2013).

Another advantage of SRSJTs compared to MRSJTs is the reduced number of steps required for item development. Typically, MRSJT development involves three major steps (see Motowidlo, 1990). First, critical incidents are collected from experts in order to build problem situations that form the basis of the SJT stems. Second, response options are developed by either a second group of experts, or alternatively, by test developers. Third, a final group of experts review the response options, often rating each one for effectiveness, to develop the scoring key. The construction of SRSJTs eliminates the second step described here by simply having a second group of experts evaluate edited critical incidents collected in the first step order to devise a scoring key (see Motowidlo, Crook, Kell, & Naemi, 2009). The omission of this step makes SRSJT development less labor intensive, which is likely to result in cost-savings for assessment developers.

MRSJTs typically require test takers to read over multiple response options in addition to the item stem, and demand mental effort from test takers to maintain various response options in working memory as they are compared. Meta-analytic evidence suggests MRSJTs with knowledge instructions and with behavioral tendency instructions yield corrected population correlation estimates of 0.35 and 0.19, respectively, with cognitive ability measures (McDaniel et al., 2007). Consequently, many MRSJTs can be considered cognitively loaded (Marentette, Meyers, Hurtz, & Kuang, 2012; McDaniel et al., 2007). SRSJTs, however, pair a situational stem with only one behavioral response option to form a brief, discrete

behavioral episode that must simply be rated for effectiveness. As there is only one focal behavior to review within a given item, SRSJTs both require less reading on the part of test takers and eschew the need to relatively compare multiple response options, which should make them less cognitively loaded than most MRSJTs (Arthur et al., 2014). Additionally, because each SRSJT item contains only one behavioral response option to evaluate as opposed to several per situational stem for a typical MRSJT item, SRSJTs should take less time for respondents to complete, which could result in more favorable applicant reactions.

Crook (2015) previously conducted a study comparing a multiple and singleresponse SJT measuring interpersonal skills regarding their relationships with contextual performance and antecedents such as personality traits and emotional intelligence in a sample of 220 undergraduate business school students. The study employed a between-subjects design, in which two different groups of participants each completed the SRSJT or MRSJT. Findings revealed that in that study the SRSJT used was more reliable than the MRSJT used, producing alphas of 0.85 and 0.61, respectively. Results showed both formats yielded similar relationships with contextual performance. The MRSJT correlated 0.24 ($p < 0.05$) with a measure of contextual performance, while the SRJT correlated 0.19 ($p < 0.05$) with a measure of contextual performance.

THE CURRENT STUDY

To create an SRSJT equivalent to its MRSJT counterpart, Crook (2015) employed a battery of SRSJT items created by combining situational stems with each of the response options for that stem, so that each MRSJT item became four SRSJT items (one for each response option). Thus, each cluster of singleresponse items shared a situational stem. Although this approach allowed for exactly the same content to appear in both SJT formats, it also introduced testlets within the measure, and consequently, may have introduced local item dependence that can artificially inflate reliabilities and validity coefficients (Zenisky, Hambleton, & Sireci, 2001). As each MRSJT item became four SRSJT items, this resulted in the SRSJT having four times as many items as the MRSJT. Given alpha is, in part, a function of the number of items within a measure (Cronbach, 1951), this increase in the number of SRSJT items compared to MRSJT items may have also artificially inflated the reported estimates of internal consistency for each measure.

In the current study we replicated and extended Crook's (2015) study by comparing two analogous multiple and single-response SJTs, each containing the exact same number of items, designed to measure procedural knowledge about how to behave in military contexts that require interaction with a person or persons of a different culture. We used a within-subjects design, in which all participants completed both the MRSJT and SRSJT on separate occasions. As

there is concern in the psychological community surrounding the replicability of many of its results and effects (e.g., Pashler & Wagenmakers, 2012), our findings will provide important information about the merits of SRSJTs, on which research is still somewhat limited. In contrast to Crook's (2015) study, we created a singleresponse version of an MRSJT by combining each situational stem with only one of its associated behavioral response options, balancing the number of effective and ineffective items presented. Although this single-response version does not contain the exact same information presented in the MRSJT version, its items are fully independent from one another and testlets within the measure do not exist, producing what may be a more accurate estimate of each format's reliability and validity. In contrast, the SRSJT measure contains less information than the MRSJT, as only one of the five responses presented in the MRSJT is featured in each SRSJT item. Thus, when comparing both measures we expect to obtain more conservative estimates of the SRJT's merits than those found by Crook (2015).

To compare convergent validity evidence for the SRSJT and MRSJT formats, we plan to also collect information from participants regarding individual difference constructs closely tied to cross-cultural competence. Past research has shown that personality traits are related to cross-cultural competence (Caligiuri, 2000). People who are more open, agreeable, conscientious, extraverted, and emotionally stable should have greater cross-cultural competence than those who are not. Another construct thought to be related to cultural competence, cultural awareness, involves the recognition that culture shapes a person's beliefs, values, and behavior, as well as those of others (Abbe, Gulik, & Herman, 2008). Individuals who have greater cultural awareness and are able to effectively identify and analyze cultural contrasts in order to avoid cultural conflicts (Bhawuk & Brislin, 2000) should be more cross-culturally competent. Empathy, defined as the ability to "put oneself in another's shoes" (Ruben, 1976) is particularly relevant to cross cultural contexts where it is important for people to understand the emotions that members of other cultures express in response to a given situation. Similarly, research suggests that perspective taking, which "allows individuals to think about the world from another person's point of view" (Reid, Kaloydis, Sudduth, & Greene-Sands, 2012, p. 10), is also crucial to the development of cross-cultural competence (Abbe et al., 2008; McCloskey, Behymer, Papautsky, Ross, & Abbe, 2010). Finally, as knowledge about how to interact effectively with individuals from a different culture in social situations is important in predicting cross-cultural competence (Yamazaki & Kayes, 2004), interpersonal skills are another important individual difference to consider when examining the convergent validity of SJTs designed to measure knowledge about effective interaction with a person or persons of a different culture.

METHOD

Sample

The sample comprised 659 adults, 49% of which (*n*=323) were male and 51% (*n*=336) were female. The age of the sample ranged from 18 to 70 years of age (*M*=35.26, *SD*=10.50). The self-reported ethnicities of the sample were 6.7% African American (*n*=44), 3.3% Asian American (e.g., Japanese, Chinese, Korean; *n*=22), 6.1% South Asian (e.g., Indian, Pakistani, Afghani, Bangladeshi; *n*=40), .5% Southeast Asian American/Southeast Asian (e.g., Cambodian, Hmong, Khmer, Laotian, Vietnamese; *n*=3), 2.1% Mexican, Mexican American, or Chicano (*n*=14), .6% Puerto Rican (*n*=4), 1.4% other Hispanic, Latino, or Latin American (*n*=9), 75% Caucasian (*n*=494), 2.6% identifying as two or more ethnicities (*n*=17), and .2% (*n*=1) identifying simply as "other." The portion of the sample reporting being either currently or formerly in the military was 5.3% (*n*=35). The highest level of education of the sample was: 1.4% earning a GED (*n*=9), 11.2% earning a high school diploma (*n*=74), 21.4% completing some college but not obtaining a degree (*n*=141), 5.5% earning an occupational Associate's degree (n D 36), 7.4% earning an academic Associate's degree (*n*=49), 40.1% earning a Bachelor's degree (*n*= 264), 9.7% earning a Master's degree (*n*=64), 2.3% earning a professional degree (*n*=15), 0.6% earning a doctoral degree (*n*=3), and 0.5% indicating "other" (*n*=3). Most participants (*n*=576) also self-reported their cumulative grade point average (GPA) in college. A total of 38.9% (=224) reported a GPA of 3.5-4.0, 45% (n=259) reported a GPA of 3.49-4.0, 13.4% (*n*=77) reported a GPA of 2.50–2.99, 2.4% (*n*=14) reported a GPA of 2.0–2.49, and .3% (*n*=2) reported a GPA of 1.50–1.99 or lower.

Procedure

After receiving approval from an institutional review board, we administered both our multiple and single-response SJT items, along with other measures potentially associated with cross-cultural competence (personality, cultural openness, empathy, perspective taking, interpersonal skills) and a measure of cross-cultural competence (an outcome variable) to the sample over four different online assessment blocks over a three week period, to reduce test taker fatigue. Participants were recruited via a Human Intelligence Task (HIT) posted on Amazon Mechanical Turk (MTurk), an online recruitment platform. Assessment blocks occurred at least three or more days apart from one another. Respondents were paid $6 for completing each assessment block and voluntarily consented to participate in the study prior to participation. As each block took approximately one hour to complete, participants were paid about $30 for the participation in the study. All items within each measure were randomized and assessment exposure was also random, such that each participant was presented with assessments displayed in a unique order.

Measures

MRSJT. A total of 203 critical incidents obtained from SMEs (US soldiers) who had previously deployed to another country were used to build the two SJTs. These critical incidents were used to craft situational prompts that would form each MRSJT item stem. The focal behavior carried out by the actor in response to the situation described in each critical incident was retained as one of the possible response options for each respective item stem. A separate, independent sample of 35 SMEs who were current or former members of the US military were asked to describe how they would handle the situations described in the prompts. Using their responses, a list of five to eleven unique behavioral response options were crafted for each prompt. Prompts that did not receive at least five unique behavioral response options were discarded. A total of 57 remaining items were then administered to a third group of 134 SMEs (current or former veterans with international deployment experience) who rated the response options for subsets of items according to the effectiveness of each. Each response alternative was rated by at least 34 SMEs. Items for which SMEs did not adequately agree in their evaluations of response effectiveness (e.g., standard deviation of 1.7 or over on a 6-point effectiveness scale; Schwab, Heneman, & DeCotiis, 1975) or for which response alternatives did not differ sufficiently in mean effectiveness such that a best or a worst response for each situational prompt could be identified, were dropped, resulting in a final test battery of 30 items. The mean of SME effectiveness ratings for each response option was used to determine the SME rankings that would be used for scoring purposes.

Each of the 30 items in the MRSJT presented respondents with five plausible behavioral response options designed to measure procedural knowledge about how to behave in military contexts that require interaction with a person or persons of a different culture. Respondents were asked to rank order each of the five response options presented for each item according to their effectiveness, where 1 D most effective and 5 D least effective. An example item from the MRSJT is displayed in Panel A of Figure 1. Participants completed this measure as part of the fourth study session. Scores were computed by computing a Spearman rankorder correlation between participants' rankings and SME rankings for each item (Weekley, Ployhart, & Holtz, 2006), eliminating any local dependency or testlet effects that could be observed with other scoring approaches. Each participant's final score is the average Spearman rank-order correlation across all 30 items. Internal consistency reliability (Cronbach's alpha) for the measure was 0.94.

SRSJT. Each SRSJT item was built by combining the situational prompt from the MRSJT with either the most effective, best response option or the most ineffective, worst response option such that the final battery contained 15 effective

items and 15 ineffective items. Participants were instructed to evaluate the effectiveness of the behavior presented in each item using a six-point scale lacking a true midpoint, where 1 D very ineffective and 6 D very effective. Participants completed this measure as part of the second study session. An example of an ineffective item from the SRSJT is displayed in Panel B of Figure 1. After reverse-scoring ineffective items, respondents' ratings were averaged to form the total score (cf., Crook et al., 2011, Martin-Raugh et al., 2016; Motowidlo et al., 2013). Internal consistency reliability (Cronbach's alpha) for the overall measure was .91. We also computed subscores for the 15 effective and 15 ineffective items. Alpha is 0.87 for both subscales.

Personality. Respondents' Big Five personality traits were measured using the Big Five Inventory (BFI; John, Naumann, & Soto, 2008), a brief (44 items), widely used, multidimensional self-report personality inventory. Participants were asked to indicate their agreement with each statement using a five-point Likert scale with anchors ranging from 1 D disagree strongly to 5 D agree strongly. Participants completed this measure as part of the first study session. Internal consistency reliability estimates (Cronbach's alpha) for each trait were: 0.85 for Openness to Experience, 0.89 for Conscientiousness, 0.89 for Extraversion, 0.84 for Agreeableness, and 0.91 for Neuroticism.

Cultural Openness. Cultural openness, defined as the level of enjoyment in engaging with activities related to other cultures, was measured using a sixitem biodata measure (Brenneman et al., 2016). Participants responded to items using a six-point scale. Participants completed this measure as part of the first study session. The first 5 anchors range from 1 D disliked very much to 5 D liked very much; there is also a sixth option to select not applicable (N/A), which is scored as missing data. An example item is "In the past, how much have you liked interacting with people from different backgrounds?" Cronbach's alpha for this measure was 0.79.

Empathetic Concern Scale. Empathy was measured using a 34-item measure (Brenneman et al., 2016). Respondents were asked to evaluate how well each statement describes themselves using a five-point Likert-scale with anchors ranging from 1 D does not describe me very well to 5 D describes me very well. Participants completed this measure as part of the second study session. An example item is "Feels or shows sympathy and concern for others." Cronbach's alpha for this measure was 0.96.

Perspective Taking. The ability to take the perspective of another person was measured using a five-item assessment (Brenneman et al., 2016) where participants were instructed to evaluate the extent to which each statement

describes themselves. Respondents used a five-point Likert scale, where 1 D does not describe me very well and 5 D describes me very well. Participants completed this measure as part of the first study session. An example item is "I find it difficult to put myself in someone else's shoes." Cronbach's alpha for this measure was 0.65.

Interpersonal Skills. Interpersonal skills were measured with a 23-item single-response SJT (Brenneman et al., 2016). Each item depicts a scenario from daily life and the behavior an actor displays to attempt to resolve the situation. Items evenly represent effective and ineffective behaviors. Participants were instructed to evaluate the effectiveness of each behavior using a sevenpoint Likert scale, where 1 D very ineffective and 7 D very effective. Participants completed this measure as part of the first study session. An example of an effective item is "You are at a professional conference where you do not know anybody. In the middle of the conference session you are attending, the speaker asks everyone to find a partner for an activity. You turn to the person nearest you and suggest you two partner for the activity." Internal consistency reliability (Cronbach's alpha) for the measure was 0.65.

Cultural Competence Self-Assessment Checklist. Cross-cultural competence, an outcome variable in this study, was measured using a 19-item measure (Chen & Androsiglio, 2010). Participants were asked to respond to a series of statements using a four-point scale where 1 D never and 4 D always/very well. Participants completed this measure as part of the fourth study session. An example item is "I am able to adapt my communication style to effectively communicate with people who communicate in ways that are different from my own." Internal consistency reliability (Cronbach's alpha) for the measure was 0.93.

RESULTS

The multiple and single-response SJTs produced a raw correlation of 0.82 ($p < 0.001$) with one another, which is not surprising given that the SRSJT is comprised entirely of information captured in the MRSJT. Unexpectedly, the MRSJT demonstrated a significantly higher alpha reliability estimate (0.94) than the SRSJT (0.91; $x^2[1, N = 659] = 75.24$, $p < 0.001$), despite both measures having the same number of items. However, there is an alternative method for estimating internal consistency reliability within a latent variable modeling framework via McDonald's omega hierarchical (v_h) estimate, which is less likely to over or underestimate reliability (Dunn, Baguley, & Brunsden, 2014), and is superior to Cronbach's alpha for estimating the degree to which scale items measure the same latent variable (Zinbarg, Revelle, Yovel, & Li, 2005).
Omega hierarchial (v_h) calculated using Revelle and Zinbarg's (2009) method is

0.76 (x^2 [348, N = 659] = 595.60, p < 0.001, RMSEA D 0.03) for the SRSJT and 0.82 for the MRSJT (x^2 [348, N = 659] = 545.12, p < 0.001, RMSEA = 0.03), providing a strong indication that there is a unidimensional structure and general hierarchical factor saturating both SJTs, given that omega values are > 0.75 for both measures (Reise, Bonifay, & Haviland, 2013).

The MRSJT and SRSJT's raw correlations with antecedents of cross-cultural competence and the measure of cross-cultural competence are displayed in Table 1 along with operational validities for the predictors. Convergent validity estimates for both SJT formats are, for the most part, similar across constructs. However, some differences were observed; these correlations were tested to see if they were significantly different from one another using equation 14 from Steiger's (1980) tests for comparing dependent correlation coefficients.

The SRSJT correlated more strongly with the measure of perspective taking (r = 0.12, p < 0.01) than the MRSJT did (r = 0.07, p = 0.075) as evidenced by the significant difference in the two correlations (Z_H= 2.15, p = 0.03). Also, as expected, the SRSJT measuring knowledge about cross-cultural competence correlated significantly more strongly with the measure of interpersonal skills, which was also a single-response SJT, as evidenced by the significant difference in the two correlations (Z_H= 4.86, p < 0.001). The SRSJT correlated 0.63 (p < 0.001) with interpersonal skills, compared with 0.54 (p < 0.001) for the MRSJT.

Although we did not directly measure cognitive ability in this study, we can examine the relationship between SJT scores and self-reported GPA and highest level of education as proxy variables for cognitive ability to get an indirect indication of the cognitive load associated with each SJT format. Counter to expectations, level of education was less strongly related to MRSJT scores (r = –0.08, p = 0.053) than SRSJT scores (r = –0.13, p < 0.05), and the difference between the two was significant (Z_H= 2.15, p < 0.05). Despite the correlation between self-reported college GPA and MRSJT scores (r = 0.15, p < 0.001) being slightly stronger than GPA and SRSJT scores (r = 0.14, p < 0.05) the difference between the two was not statistically significant (Z_H= –0.432, p = 0.33).

Knowledge of effective behavior measured by the SRSJT was strongly correlated with knowledge of ineffective behavior measured by the same instrument (r = 0.51, p < 0.001). The reliability of each knowledge subscale measured by the SRSJT yielded a lower reliability (0.87) compared to when the two subscales were combined (0.91), perhaps as a result of fewer items contributing to each of the subscores. Knowledge of effective behavior was more strongly correlated with cross-cultural competence (r = 0.32, p < 0.001) than knowledge of ineffective behavior was (r = 0.17, p < 0.001) and this difference was statistically significant (Z_H= 4.04, p < 0.001).

TABLE 1

MRSJT and SRSJT's Correlations with Antecedents of Cross-Cultural Competence and the Measure of Cross-Cultural Competence (N = 659)

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. MRSJT	.65	.23	.94													
2. Overall SRSJT	4.80	.62	.82	.91												
3. Effective SRSJT Items	5.05	.63	.65	.83	.87											
4. Ineffective SRSJT Items	4.56	.80	.76	.90	.51	.87										
5. Neuroticism	2.47	.92	−.04	−.03	−.07	.02	.91									
6. Agreeableness	3.87	.65	.28	.26	.30	.17	−.45	.84								
7. Conscientiousness	4.00	.69	.28	.26	.27	.19	−.50	.51	.89							
8. Extraversion	2.83	.92	−.21	−.21	−.13	−.22	−.45	.24	.29	.89						
9. Openness	3.46	.73	.08*	.11	.13	.07	−.20	.20	.24	.31	.85					
10. Cultural Openness	4.15	.61	.10	.10	.12	.07	−.06	.26	.19	.16	.36	.79				
11. Empathy	3.99	.61	.36	.40	.41	.29	−.24	.66	.50	.25	.32	.30	.96			
12. Perspective Taking	3.69	.69	.07	.12	.16	.05	−.31	.57	.45	.39	.44	.39	.65	.65		
13. Interpersonal Skills	4.15	.42	.54	.63	.51	.57	−.14	.29	.29	−.05	.06	.15	.32	.19	.65	
14. Cultural Competence	4.47	.72	.23	.27	.32	.17	−.20	.44	.39	.22	.41	.43	.53	.55	.26	.93
			(.24)	(.28)	(.33)	(.18)	(-.21)	(.46)	(.40)	(.23)	(.43)	(.45)	(.55)	(.57)	(.27)	

*$p < 0.05$; **$p < 0.01$ (two-tailed). Reliability estimates (Cronbach's alpha) appear on the diagonal. Operational validities corrected for measurement error in cultural competence appear in parentheses.

The difference in how MRST and SRSJT scores correlated with cross-cultural competence was not statistically significant ($p = 0.09$). However, we carried out a series of multiple regressions where all the predictor measures included in this study were entered simultaneously to predict cross-cultural competence scores, with results shown in Table 2. Five regressions were performed, with each including either the MRSJT scores, overall SRSJT scores, both the effective and ineffective SRSJT scores, both the overall MRSJT scores and overall SRSJT scores, or both the effective and ineffective SRSJT scores and overall MRSJT scores. When all of the other predictors were entered into the model along with only one of the SJT format scores, MRSJT scores did not produce a significant standardized beta weight ($b = 0.05$, $p = 0.183$) while SRSJT scores did ($b= 0.09$, $p < 0.05$). When SRSJT scores were broken down into knowledge of effective and ineffective behavior, knowledge of effective behavior was the only significant SJT-based predictor of cross-cultural competence ($b = 0.13$, $p < .01$). When both MRSJT scores and overall SRSJTs were included in the model along with all of the other predictor variables, neither score accounted for enough variance for its beta weight to be considered statistically significant. Finally, when cross-cultural competence was regressed on both the effective and ineffective SRSJT scores and overall MRSJT scores, knowledge of effective behavior was the only significant SJT-based predictor of cross-cultural competence ($b= 0.13$, $p < .01$).

We also conducted two hierarchical regressions to examine the respective amounts of incremental variance in cross-cultural competence accounted for by each SJT. We first regressed cross-cultural competence on MRSJT scores and the other predictor variables in Step 1, and then added the SRSJT scores as a predictor in Step 2. The R^2 at Step 1 was 0.410 ($p < 0.001$). Once the SRSJT scores were added at Step 2, R^2 was .003 ($p = 0.096$) and the total R^2 for the model was 0.412 ($p < 0.001$). We then repeated the hierarchical regression but reversed the order of SJT score entries. When SRSJT scores were entered at Step 1, the R^2 was 0.408 ($p < 0.001$). When the MRSJT scores were added at Step 2, ΔR^2 was .004 ($p < 0.05$). Thus, although this change is statistically significant, the amount of incremental validity accounted for by both the MRSJT and SRSJT is very small.

TABLE 2
Standardized Beta Weights Predicting Cross-Cultural Competence (N = 659)

	Model 1		Model 2		Model 3		Model 4		Model 5	
	β	SE	β	SE	β	SE	β	SE	β	SE
MRSJT	.05	.13	—	—	—	—	−.01	.17	.01	.17
Overall SRSJT	—	—	.09	.05	—	—	.09	.07	—	—
Effective SRSJT Items	—	—	—	—	.13	.04	—	—	.13	.05
Ineffective SRSJT Items	—	—	—	—	−.02	.04	—	—	−.03	.04
Neuroticism	.03	.03	.03	.03	.04	.03	.03	.03	.04	.03
Agreeableness	.07	.05	.07	.05	.07	.05	.07	.05	.07	.05
Conscientiousness	.08	.04	.08	.04	.08	.04	.08	.04	.08	.04
Extraversion	.01	.03	.02	.03	.02	.03	.02	.03	.02	.03
Openness	.16	.03	.15	.03	.15*	.03	.15	.03	.15*	.03
Cultural Openness	.20	.04	.20	.04	.20	.04	.20	.04	.20	.04
Empathy	.17	.06	.16	.06	.15	.06	.16	.06	.15	.06
Perspective Taking	.21	.05	.21	.05	.21	.05	.21	.05	.21	.05
Interpersonal Skills	.07	.06	.04	.07	.05	.07	.04	.07	.05	.07
R^2	.44		.44		.45		.44		.45	

*$p < .05$; **$p < .01$ (two-tailed).

DISCUSSION

The results of this study provide an important replication of prior research (Crook, 2015) comparing the reliabilities, convergent, and predictive validities of a more recently introduced SRSJT and a traditional MRSJT. Moreover, this study extends Crook's (2015) prior findings by conducting a more conservative test of validity and reliability estimates for the SRSJT, as each item in the SRSJT contains a unique situational stem and behavioral response, eliminating local item dependency that may have artificially inflated the reliability estimates for the SRSJT used in prior research on this topic.

Although the MRSJT was more reliable ($\alpha = 0.94$, $\nu_h = 0.82$) when compared to the SRSJT ($\alpha = 0.91$, $\nu_h = 0.76$) in this investigation, the reliabilities of both measures are comparable, falling in what can be considered the high range (George & Mallery, 2003). The unusually high reliability of the MRSJT observed in this investigation is surprising given that internal consistency reliability estimates for many MRSJTs tend to be low (Lievens et al., 2008; Whetzel & McDaniel, 2009). The unidimensional factor structure of both SJT scales, which is especially unusual for the MRSJT (McDaniel & Whetzel, 2005), may be what is driving the unusually high internal consistency estimates. Replication in other contexts is warranted, as findings from this study contrast with those reported by Crook (2015) where the SRSJT produced a higher reliability estimate than the MRSJT used in that investigation.

The two halves of the SRSJT, each measuring knowledge of effective behavior and knowledge of ineffective behavior, were strongly correlated in this study. This finding contrasts with those of prior research (e.g., Crook et al., 2011; Motowidlo et al., 2013) showing the correlation between the two subscales to be weak or even negative. The strong correlation between knowledge of effective behavior and knowledge of ineffective behavior in the SJTs used in this study may have contributed to the observed high reliability of the MRSJT. The MRSJT response options in this study may be unidimensional rather than multidimensional, which is often the case with many MRSJTs. In this study, knowledge of effective behavior was more strongly correlated with cross-cultural competence than knowledge of ineffective behavior was. Moreover, results of the regression analyses showed that when SRSJT scores are broken down into knowledge of effective and ineffective behavior, knowledge of effective behavior was the only significant SJT-based predictor to account for unique variance in cross-cultural competence scores. This finding holds true even when overall MRSJT scores were included in the model. Findings reported in the literature regarding the differential validity of knowledge of effective versus ineffective behavior in predicting criteria have been mixed (see Crook et al., 2011; Motowidlo et al., 2013). Thus, the predictive validity of each construct may vary depending on the domains and outcome variables. Additional research should compare the predictive validity of

knowledge of effective and ineffective behavior using other outcomes and domains.

Our results suggest convergent validity estimates for the SRSJT and MRSJT used in this study are comparable. In the instances where convergent validity coefficients significantly differed across SJT formats, relationships with perspective taking and interpersonal skills were stronger for the SRSJT. The stronger correlation between the SRSJT measuring procedural knowledge and our measure of interpersonal skills, which were also measured using an SRSJT, were unsurprising given the common method variance of the two measures (Arthur & Villado, 2008).

Although we were only able to indirectly explore the cognitive load associated with each SJT format in this study, we were surprised to find that level of education was more strongly related to SRSJT scores than MRSJT scores. Moreover, differences in how strongly related self-reported college GPA was to SJT scores were not statistically significant in this investigation. These findings were counter to our expectations given prior research indicates many MRSJTs may be cognitively loaded (Marentette, et al., 2012; McDaniel et al., 2007) and that SRSJTs should be less so because they do not require respondents to simply make evaluative judgments using a rating scale rather than making comparative judgments across response options (Arthur et al., 2014). Future research should compare the cognitive load of MRSJTs and SRSJTs explicitly using direct measures of cognitive ability to allow for stronger inferences to be made.

In this study, we found that the difference in how well each SJT predicted cross-cultural competence was not statistically significant, replicating results reported by Crook (2015) showing the two formats produce comparable predictive validity estimates. Multiple regression analyses further support the comparability of both SJT formats by showing that MRSJT scores and SRSJT scores yielded very similar, very small estimates of incremental variance in cross-cultural competence accounted for when other predictors were included in the model. Thus, both formats appear to perform similarly in terms of predictive validity. Nonetheless, replication in other contexts with other criteria is needed.

As both this investigation and that conducted by Crook (2015) were carried out in low-stakes settings and employed an outcome variable measured via selfreport, replication in other contexts using behavioral measures or others' ratings is warranted. Future research should compare the predictive validity of SRSJTs and MRSJTs in high-stakes settings, and in other domains, perhaps using supervisory ratings of job performance or behaviors displayed in simulations or roleplays as outcome variables.

Our findings concerning the use of SRSJTs as developmental or selection tools are promising. Results suggest SRSJTs can produce internal consistency reliability, convergent, and predictive validity estimates that are comparable to those achieved with many traditional MRSJTs. However, as SRSJTs eliminate the step in item development in which response options are generated by an additional

group of subject matter experts, they are simpler and less time-intensive to generate. As a result, SRSJTs should be less expensive to develop than many MRSJTs. However, scientists and practitioners may wish to test this prediction empirically by placing a concrete monetary value on any cost-savings afforded by SRSJTs.

Our study is not without limitations. It should be noted that in this study we used one particular variant of the many MRSJT configurations represented in the literature, which may vary according to instruction type, number of response options, and scoring approaches. Thus, further research should compare the SRSJT methodology to different types of MRSJTs. Crook (2015) created an SRSJT by combining situational stems with each of the response options for that stem such that each MRSJT item became four SRSJT items, while our study combined only one response option with a stem to form each SRSJT item. However, those are not the only available study designs for comparing SJT formats. The construction of the SRSJT format used in this investigation represents a single combination of MRSJT stems and response options, and different combinations could have been produced that may have resulted in different reliability and validity estimates than those reported in this study. Future research may develop and compare multiple different versions of an SRSJT such that each response option from its analogous MRSJT is represented as a single-response item. Such an approach would simultaneously avoid local item dependence and ensure that all response option content in the MRSJT would be captured by the SRSJT, although it would be considerably more resource intensive, time-consuming for participants, and would require careful administration to avoid learning effects that may become pronounced after several SRSJT administrations. Finally, another caveat worth mentioning is that our use of a within-subjects design does not allow us to completely rule out the effect of learning on observed score differences.

In sum, the study reported here addresses a current need in psychological science for the replication of effects demonstrated by previous studies. Psychology has recently been viewed as facing a "replication crisis" (Maxwell, Lau, & Howard, 2015), although this issue has been seen as a larger problem for social psychology than the study of individual differences (see Baumeister, 2016). Our findings generally confirm those reported by Crook (2015) and further demonstrate the validity of SRSJTs using a much larger sample (N D 659 compared to N D 220), an SRSJT that is designed to eliminate local dependence, a different focal construct, and a within-subjects design. However, multiple replication studies across a variety of domains and settings are called for to more firmly establish the validity of SRSJTs.

FUNDING

This work was supported by the Army Research Institute for the Behavioral and Social Sciences (W5J9CQ-12-C-0039).

REFERENCES

Abbe, A., Gulick, L. M. V., & Herman, J. L. (2008). Cross-cultural competence in army leaders: A conceptual and empirical foundation (Study Report 2008–01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Arthur Jr., W. Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. Journal of Applied Psychology, 99, 535–545. doi:10.1037/a0035788

Arthur Jr., W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. Journal of Applied Psychology, 93, 435–442. doi:10.1037/0021-9010.93.2.435

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. Journal of Experimental Social Psychology, 66, 153–158. doi:10.1016/j.jesp.2016.02.003

Bhawuk, D., & Brislin, R. (2000). Cross cultural training: A review. Applied Psychology: An International Review, 49, 162–192. doi:10.1111/1464-0597.00009

Brenneman, M. W., Barnwell, P., Anguiano-Carrasco, C., Carney, L., Ezzo, C., Golubovich, J., ... Kochert, J. (2016). Development of an assessment of cross cultural competence (3C): Assessment guide (W5J9CQ-12-C-0039). Fort Belvoir, VA: Army Research Institute Technical Report.

Caligiuri, P. M. (2000). Selecting expatriates for personality characteristics: A moderating effect of personality on the relationship between host national contact and cross-cultural adjustment. MIR: Management International Review, 61–80.

Chen, E. C., & Androsiglio, R. (2010). Cultural competency self-assessment checklist. In C. S. Clauss-Ehlers (Ed.), Encyclopedia of cross-cultural school psychology (pp. 300–302). Boston, MA: Springer.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334. doi:10.1007/BF02310555

Crook, A. E. (2015). Comparing single-response and multiple-response SJTs. Paper presented at the 29th annual meeting of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single response situational judgment tests. International Journal of Selection and Assessment, 19, 363–373. doi:10.1111/j.1468-2389.2011.00565.x

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105, 399– 412. doi:10.1111/bjop.12046

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed). Boston, MA: Allyn & Bacon.

Ghosh, K., Motowidlo, S.J., & Nath, S. (2015). Technical knowledge, prosocial knowledge, and clinical performance of Indian medical students. International Journal of Selection and Assessment, 23, 59–70. doi:10.1111/ijsa.12095

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five Trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (3rd ed., pp. 114–158). New York, NY: Guilford Press.

Kell, H. J., Motowidlo, S. J., Martin, M. P., Stotts, A. L., & Moreno, C. A. (2014). Testing for independent effects of prosocial knowledge and technical knowledge on skill and performance. Human Performance, 27, 311–327. doi:10.1080/08959285.2014.929692

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. Personnel Review, 37, 426–441. doi:10.1108/00483480810877598

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment test items: A case of construct irrelevant difficulty. International Journal of Selection and Assessment, 20, 319–332. doi:10.1111/j.1468-2389.2012.00603.x

Martin-Raugh, M. P., Kell, H. J., & Motowidlo, S. J. (2016). Prosocial knowledge mediates effects of agreeableness and emotional intelligence on prosocial behavior. Personality and Individual Differences, 90, 41–49. doi:10.1016/j.paid.2015.10.024

Maul, A. (2013). Method effect and the meaning of measurement. Frontiers in Psychology, 4, 1–13. doi:10.3389/fpsyg.2013.00169

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? American Psychologist, 70, 487–498. doi:10.1037/a0039400

McCloskey, M. J., Behymer, K. J., Papautsky, E. L., Ross, K. G., & Abbe, A. (2010). A developmental model of cross-cultural competence at the tactical level (Technical Report 1278). Alexandria, VA: U.S. Army Research Institute.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L, III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology, 60, 63–91. doi:10.1111/j.1744-6570.2007.00065.x

McDaniel, M.A., & Whetzel, D.L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. Intelligence, 33, 515–525. doi:10.1016/j.intell.2005.02.001

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), Applied measurement: Industrial psychology in human resources management (pp. 235–257). New York, NY: Erlbaum.

Motowidlo, S. J., Brownlee, A. L., & Schmit, M. J. (2008). Effects of personality characteristics on knowledge, skill, and performance in servicing retail customers. International Journal of Selection and Assessment, 16, 272–281. doi:10.1111/j.1468-2389.2008.00433.x

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. Journal of Business and Psychology, 24, 281–288. doi:10.1007/s10869-009-9106-4

Motowidlo, S., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. Journal of Applied Psychology, 75, 640–647. doi:10.1037/00219010.75.6.640.

Motowidlo, S. J., Martin, M. P., & Crook, A. E. (2013). Relations between personality, knowledge, and behavior in professional service encounters. Journal of Applied Social Psychology, 43, 1851–1861. doi:10.1111/jasp.12137

Nguyen, N.T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. International Journal of Selection and Assessment, 13, 250–260. doi:10.1111/j.1468-2389.2005.00322.x.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science, 7, 528–530. doi:10.1177/1745691612465253

Ployhart, R. E., & MacKenzie Jr., W. I. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), APA handbook of industrial and organizational psychology (Vol. 2, pp. 237–252). Washington, DC: American Psychological Association.

Reid, P., Kaloydis, F. O., Sudduth, M. M., & Greene-Sands, A. (2012). Executive summary: A framework for understanding cross-cultural competence in the Department of Defense (Technical Report No. 15-12). Washington, D.C.: DEOMI.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95, 129–140. doi:10.1080/00223891.2012.725437

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika, 74, 145–154. doi:10.1007/s11336-008-9102-z

Ruben, B. D. (1976). Assessing communication competency for intercultural adaptation. Group & Organization Management, 1, 334–354.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549–562. doi:10.1111/j.1744-6570.1975.tb01392.x

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245–251. doi:10.1037/0033-2909.87.2.245

Waugh, G. W., Putka, D. J., & Sager, C. E. (2002, October). Development and validation of a U.S. Army situational judgment test. In G.W. Waugh (Chair), Tailoring a situational judgment test to different pay grades. Symposium conducted at the conference of the International Military Testing Association, Ottawa, Canada.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. Situational Judgment Tests: Theory, Measurement, and Application, 26, 157–182.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. Human Resource Management Review, 19, 188–202. doi:10.1016/j.hrmr.2009.03.007

Yamazaki, Y., & Kayes, D. C. (2004). An experiential approach to cross-cultural learning: A review and integration of competencies for successful expatriate adaptation. Academy of Management Learning & Education, 3, 362–379. doi:10.5465/AMLE.2004.15112543

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test, and ability statistics. MCAT Monograph.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's a, Revelle's b, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70, 123–133. doi:10.1007/s11336-003-0974-7