Can You Crowdsource Expertise?
Comparing Expert and Crowd-Based Scoring Keys for Three Situational Judgment Tests

Matt I. Brown
Geisinger Health System
Lewisburg, PA
mibrown9015@gmail.com

Michael A. Grossenbacher
Wonderlic, Inc.
Vernon Hills, IL
Michael.grossenbacher@wonderlic.com

Michelle P. Martin-Raugh
Educational Testing Service
Princeton, NJ
Mmartin-raugh@ets.org

Jonathan Kochert
U.S. Army Research Institute for the Behavioral and Social Sciences
jonathan.f.kochert.civ@mail.mil

Matthew S. Prewett
Central Michigan University
Mount Pleasant, MI
Prewe1ms@cmich.edu

Version 3 accepted for publication at the *International Journal of Selection and Assessment*, 9/22/2021

Running Head: CAN YOU CROWDSOURCE EXPERTISE?

Acknowledgements: We would like to thank Harrison Kell and Samuel Rikoon for their helpful feedback on a previous version of this manuscript. A portion of this research was presented as a poster at the 34th Meeting for the Society for Industrial and Organizational Psychology; National Harbor, MD.

Disclaimer
The research described herein was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army (Cooperative Agreement No. W911NF-18-2-0018. The views expressed in this presentation are those of the author and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.

Abstract

It is common practice to rely on a convenience sample of subject matter experts (SMEs) when developing scoring keys for situational judgment tests (SJTs). However, the defining characteristics of what constitutes a SME are often ambiguous and inconsistent. Sampling SMEs can also impose considerable costs. Other research fields have adopted crowdsourcing methods to replace or reproduce judgments thought to require subject matter expertise. Therefore, we conducted the current study to compare crowdsourced scoring keys to SME-based scoring keys for three SJTs designed for three different job domains: Medicine, Communication, and Military. Our results indicate that scoring keys derived from crowdsourced samples are likely to converge with keys based on SME judgment, regardless of test content (r = .88 to .94 between keys). We observed the weakest agreement among individual MTurk and SME ratings for the Medical SJT (classification consistency = 61%) relative to the Military and Communication SJTs (80% and 85%). Although general mental ability and conscientiousness were each related to greater expert similarity among MTurk raters, the average crowd rating outperformed nearly all individual MTurk raters. Using randomly-drawn bootstrapped samples of MTurk ratings in each of the three samples, we found that as few as 30 to 40 raters may provide adequate estimates of SME judgments of most SJT items. These findings suggest the potential usefulness of crowdsourcing as an alternative or supplement to SME-generated scoring keys.

Keywords: subject matter expertise; crowdsourcing; situational judgement tests; consensus-based measurement; implicit trait policies

Can You Crowdsource Expertise?
Comparing Expert and Crowd-Based Scoring Keys for Three Situational Judgment Tests

This study seeks to answer a simple question: Is it possible to develop a scoring key for a situational judgment test (SJT) without a pool of subject matter experts (SMEs)? The SJT method is widely studied and used for selection in both occupational and educational settings (Oswald et al., 2004; Lievens & Sackett, 2012). SJTs are typically designed to measure procedural knowledge about how to behave effectively in a particular job (Motowidlo et al., 2006). Along these lines, SJT items are typically scored using rational keys based on ratings gathered from SMEs (Whetzel et al., 2020). According to the U.S. Office of Personnel Management, SMEs are defined as a "person with bona-fide expert knowledge about what it takes to do a particular job" (Office of Personnel Management, 2016). Their judgments determine how each item will be scored. This poses a challenge when creating a new assessment because test developers typically need an item pool at least twice as large as the desired length of the final test (Hinkin, 1998; Murphy & Davidshofer, 1998). Each of these items needs a scoring key in order to be evaluated during an initial round of testing. This means that SMEs would likely spend most of their time rating items that will not be used in the final version of the test.

Researchers and test developers face several challenges when using SMEs. The criteria for determining who is considered a SME can be vague or inconsistent across studies. When SJTs are developed for a specific occupation, SMEs have included job incumbents, supervisors, customers, and even novices (Weekley et al., 2006). In many cases, these SMEs work in prestigious occupations (e.g., experienced physicians or medical professionals, Lievens & Sackett, 2012; Patterson et al., 2009) which can make them expensive and difficult to recruit. For more construct-driven SJTs (e.g., applied social skills), academic researchers or graduate

students have even been used as SMEs. For example, graduate students have been used to develop scoring keys for SJTs measuring personal initiative (Bledow & Frese, 2009), leadership styles (Bergman et al., 2006), and prosocial implicit trait policy (Motowidlo et al., 2016). For these reasons, researchers have indicated that using as few as 5 to 7 SMEs to generate a scoring key is acceptable (Guenole et al., 2017). Some scholars have argued that many SJTs measure some degree of general domain knowledge (Motowidlo et al., 2010), practical intelligence (Wagner & Sternberg, 1985), or common sense (Salter & Highhouse, 2009). These more general forms of expertise suggest that narrowly defined SMEs may not always be required to differentiate between effective and ineffective responses to workplace scenarios.

These challenges led us to consider whether crowdsourcing methods could be used to gather judgments that are comparable to those obtained from SMEs. Crowdsourcing has been widely investigated as a method for approximating expert judgment in a variety of fields beyond the organizational sciences. For example, crowdsourcing has been used to detect speech disorders in recorded speech (Byun et al., 2016), to identify malaria parasites in images of blood samples (Luengo-Oroz et al., 2012), and to provide valid ratings of surgical performance (Lendvay et al., 2015). In each of these applications, crowdsourced judgments were found to highly converge with judgments obtained from experts despite the fact that individual crowd members typically lacked any prior training or experience. Inspired by these findings, we sought to investigate whether crowdsourcing could provide comparable results to SME ratings for SJTs involving various forms of job contexts and psychological domains. This would greatly simplify the SJT development process and potentially provide faster and more convenient way of developing scoring keys.

Wisdom of the Crowd and Consensus-Based Measurement

As an alternative to expert judgment, some scholars have used judgments or forecasts by aggregating responses from large groups of laypeople. This practice is commonly referred to as the "wisdom of the crowd", where the mathematical aggregate of individual judgments (the crowd) tends to be more accurate than any single rater. This idea is commonly attributed to Francis Galton (1907) and has inspired the creation of the crowdsourcing concept. Crowdsourcing is defined as the use of large groups of individual contributors towards the completion of a single task. This practice has inspired several best-selling popular science books (e.g., Surowiecki, 2004) and has been used to forecast the outcomes of political elections (Berg et al., 2008) and the replication of scientific findings (Dreber et al., 2015; Hoogeveen et al., 2020). In applied psychology, one common application of the "wisdom of the crowd" is consensus-based measurement (CBM). As the name suggests, CBM relies on the aggregation of many individual judgments to determine scoring keys for individual test items. This method is particularly useful when developing assessments to measure constructs that lack an objective correct response, including emotional intelligence (MacCann & Roberts, 2008), driving knowledge (Legree et al., 2003), and leadership (Hedlund et al., 2003). Here, it is not always clearly understood which response option should be considered correct or incorrect. This approach has also been moderately successful in developing keys for items which there is an objectively correct answer (e.g., vocabulary test items; Barchard et al., 2013). In each of these applications, it is theorized that the "correct" response will emerge from aggregating judgments from a large pool of individual raters.

Not only has CBM been shown to provide adequate criterion-related validity, but scoring keys developed using CBM often correlate strongly with expert-based scoring keys (Legree et al., 2005). The rationale for this finding is based in classical test theory, where expertise (or

accurate judgment) is considered to manifest as systematic covariance in ratings across raters. On the other hand, any differences in judgment are assumed to be due to random error. Along these lines, experts are expected to have greater consistency in their individual judgments compared to novices (Einhorn, 1974). The relative lack of consistency among novice judges may be countered by simply increasing the number of judges. This is analogous to the practice of adding scale items to improve internal consistency or gathering more raters to improve interrater reliability (Schmidt & Hunter, 1996). By this logic, a larger sample of (relatively) inconsistent novices may potentially be used to obtain results similar to judgments of a smaller, more consistent sample of SMEs.

Research Question 1: How do SJT scoring keys based on crowdsourced, novice (MTurk) rater judgments compare to keys based on expert (SME) judgments?

Individual Differences among Crowd Members

Although part of the allure of crowdsourcing is that the collective output is expected to be greater than the sum of its individual contributors, some researchers have investigated methods for constructing optimal crowds. One approach to improving the effectiveness of the crowd is by finding an optimal composition of members (e.g., Oliveria & Nisbitt, 2018). Another approach has been to attempt to identify the best judges within the crowd. However, this strategy has often used past performance as an indicator of ability or expertise (e.g., Mannes et al., 2014) or has relied on repeated performance on SJT items in order to quantify expertise (Zhang & Wang, 2021). For the present research, we are interested in investigating whether individual differences in general ability or disposition may be useful for identifying more accurate judges within the

crowd. Ability or dispositional constructs are ideal for screening crowd members because these characteristics can be reliably measured quickly and efficiently (e.g., Revelle et al., 2020). Although individual characteristics have been shown to be important for performance in forecasting tasks (e.g., GMA and motivation; Mellers et al., 2015), evaluations of other people (De Kock et al., 2020) or accuracy in judging another person's personality traits (Rogers & Biesanz, 2019), few studies have examined the role of individual differences in crowdsourcing tasks. Therefore, we explore whether individual differences can identify crowd members whose judgments are more similar to SMEs

One individual difference that is potentially linked to stronger convergence with expert ratings is general mental ability (GMA; Gottfredson, 1997). GMA, also commonly referred to as general intelligence or cognitive ability, is defined as a higher-order factor which is commonly found across different, more narrowly defined abilities (e.g., verbal ability or spatial reasoning). GMA is strongly linked to specific forms of job knowledge and training performance (Schmidt & Hunter, 2004). Not only is GMA linked to expertise on the job, GMA is also a predictor of performance in SJTs and other job simulations (Meriac et al., 2008). Although SJTs are sometimes considered measures of practical knowledge and distinct from GMA, SJT scores typically correlate with measures of GMA (McDaniel & Whetzel, 2005). In addition, recent research indicates that GMA scores tend to correlate equally with performance on SJTs designed to measure either job-specific or general domain knowledge (Schäpers et al., 2020). Moreover, greater GMA has also been linked to greater accuracy when judging others (De Kock et al., 2020) and with more accurate judgments on SJT items (Zhang & Wang, 2021). Thus, GMA likely plays a role in the ability to identify the correct response to SJT items and in providing ratings that are more strongly aligned with expert judgment.

Research Question 2: Do novices (MTurk raters) with higher GMA provide ratings that correlate more strongly with expert (SME) ratings?

Personality traits are also thought to be direct antecedents of the general knowledge constituted by ITPs (Motowidlo et al., 2018). In particular, conscientiousness, defined as the general tendency of an individual to be disciplined, reliable, and achievement-oriented, emerges as a construct-relevant antecedent for general job-based knowledge. Prior research has shown that conscientiousness demonstrates the highest validity in predicting performance in all job families examined, with corrected mean correlations ranging from $\rho = .19$ in skilled occupations to $\rho = .25$ in sales occupations (Barrick et al., 2001). Some researchers (Viswesvaran et al., 2005) have even gone as far as to postulate that the conscientiousness of performers is part of a general factor that contributes to all supervisory ratings of job performance. To the extent that conscientious behavior is considered truly effective for a particular job, conscientiousness should be associated with knowledge measured by an SJT and should contribute to greater scoring key validity. Conscientiousness may also promote greater scoring validity for SJT items because conscientious participants should be expected to pay closer attention to scale instructions and spend more time deliberating over responses. Previous research has found that conscientiousness is negatively correlated with an index of insufficiently effortful response patterns (Huang, Liu, & Bowling, 2015). Unlike GMA, however, conscientiousness has not been consistently found to be positively related to rater accuracy in judging others (de Kock et al., 2020; de Vries et al., 2021). However, some individual studies have found positive relationships between conscientiousness and the ability to understand human personality and more accurate in estimating their own ability

(Mayer et al., 2020). Thus, we explore whether non-SMEs who are higher in conscientiousness provide higher quality data.

Research Question 3: Do novices (MTurk raters) with higher conscientiousness provide ratings that correlate more strongly with expert (SME) ratings?

Method

We gathered samples of SME and novice ratings of response options on three different SJTs (Samples 1-3). The three SJTs vary in the extent to which they were designed to require knowledge of a specific job or work environment. We provide an example item from each SJT in Figure 1. The Medical SJT used in Sample 2 was created based on critical incidents from a specific job context (emergency medicine departments). As such, many of these SJT items use job-specific language and terminology. The remaining two SJTs were designed for a broad set of occupations within a specific setting. The Communication SJT (Sample 1) was developed to assess effective communication behaviors in a general, United States office setting. Likewise, the Military SJT (sample 3) was designed to assess cultural understanding and procedural knowledge of effective behavior within the United States military. These SJTs all target effective, interpersonal behavior in different work contexts which is representative of a large portion of SJTs used in research (Motowidlo et al., 2018) and in practice for selection or development purposes (Black et al., 2021; Lievens et al., 2016).

In the following section, we provide more detail about each of the SJTs and a description of the SME and novice raters that participated within each sample. Each sample was collected as part of separate test development projects conducted at different points in time so some of the

SJT details or SME rating tasks vary somewhat. For example, the Military SJT sample was collected as part of a large-scale test development project involving a commercial test developer and support from the U.S. Army Research Institute. This provided the resources to recruit the largest sample of SMEs (n = 66) and MTurk raters (n = 631). In contrast, data for the Medical SJT were collected by a small team of academic researchers as part of a graduate thesis project (May, 2015) and without the involvement or funding from a testing firm. As a result, fewer SMEs and MTurk raters (n = 28 and 31 respectively) were recruited in this sample. Likewise, data for the Communication SJT were collected as part of a pilot test development project at a commercial testing firm. The nature of this project provided the resources to obtain a modest sample of MTurk raters (n = 106) but we were limited by the number of available SMEs to recruit from within the organization. Although these methodological differences are not ideal for determining causality, our primary goal was to observe whether our crowdsourcing results would generalize across the three samples despite these differences. We also believe that these samples represent some of the logistical conditions and limitations faced by test developers in research and practice.

Measures

SJTs

Communication SJT. Participants in Sample 1 provided ratings for a nine-item SJT which was designed to measure effective communication in the workplace. This SJT was designed to be used as part of a soft skills training program. Each item consists of a short scenario and four or five behavioral response options. Each scenario was depicted in a comic strip format, with illustrations created by a third-party contractor. All scenarios were designed to be relevant to general office settings and were assumed to not require any specific form of procedural job

knowledge. Both SMEs and MTurk raters independently rated each of the four or five response options within each item using a 7-point Likert scale ranging from 7=*most effective* to 1=*least effective.* The Communication SJT was developed as part of a pilot project to investigate whether MTurk raters could effectively approximate SME judgments. Therefore, there is not any additional validity evidence that we can provide for these nine items. However, we attempted to replicate our findings using two other SJTs that have demonstrated evidence supporting their construct validity.

SME ratings were gathered from nine employees at a privately-owned test development company. SMEs were sampled across a variety of functions within the company, including consulting, content development, sales, and research and development. Most of the SMEs were currently working as supervisors and held titles including manager, director, and executive vice president. We did not collect any demographic information for this set of SMEs. Novice ratings were collected by recruiting 106 participants using MTurk. MTurk raters also responded to questions about their past work history. Each MTurk rater was paid $3 for completing the study. In addition, a subset of MTurk rater completed a cognitive assessment (the Wonderlic Personnel Test, n = 66) and/or a self-report personality assessment (n = 68).

Medical SJT. Participants in Sample 2 provided ratings for a 20-item SJT which was designed to assess knowledge of teamwork competencies in emergency medicine (May & Prewett, 2016). Each item was developed to measure one of five dimensions frequently used in medical teamwork assessments (Battle & King, 2010): environmental awareness, communication, leadership, team structure, and mutual support. Each item consists of a short scenario designed to assess understanding of behavioral norms for interactions between patients, nursing staff, physicians, and resident physicians. For example, items involve conflicts among specialists or between residents and physician leaders or nursing staff and attending physicians.

Both SMEs and MTurk raters ranked each of the five response options in order of their effectiveness (1=*most effective* to 5=*least effective*) within each of the 20 SJT items.

SME ratings were gathered from a sample of 28 emergency department physicians and physician assistants who were personally recruited by the researchers or contacted through the Society of Emergency Medicine Physician Assistants. This is line with past research, where incumbent medical professionals have been used as SMEs in the development of SJTs designed for use in health care settings (e.g., Buyse & Lievens, 2011; Lievens et al., 2005). Each SME received a $20 Amazon gift card for completing the SJT ratings. Most SMEs identified as female (55%) and White (77%). Novice ratings were collected by recruiting 31 participants from MTurk. Novices were paid $5 for completing the study. In addition to rating each of the SJT item responses, novices also completed a short measure of GMA, conscientiousness, and completed several demographic questions. Most novices identified as female (57%) and White (90%). General job experience ranged from 0 to 28 years, with a median of 6.5 years of experience.

Reliability and validity evidence for the Medical SJT were documented by May and Prewett (2016) based on samples of medical professionals and medical students. The Medical SJT displayed good internal consistency ($\alpha = .82$) as well as strong parallel form reliability from a single testing session ($r = .88$). The authors found that professionals scored significantly higher on the SJT compared to students ($d = 1.54$, $t(56) = p < .01$), which was interpreted to indicate a difference in job-related knowledge between groups. Individuals who performed better on the SJT also tended to report more positive attitudes towards teamwork as measured by the TeamSTEPPS teamwork attitudes questionnaire ($r = .30$; Baker et al., 2010).

Military SJT. Participants in Sample 3 provided ratings for a 30-item SJT which was designed to assess procedural knowledge about how to behave in military contexts that require interactions with person(s) of a different culture. Participants were instructed to rank each of the response options in terms of their effectiveness by dragging and dropping them in order of most effective to least. SMEs consisted of incumbent U.S. Army personnel (n = 66) with overseas deployment experience that had frequent interactions with foreign populations during deployment. Past research has relied on incumbent military personnel as SMEs for other instruments designed for use in similar settings (e.g., Hedlund et al., 2003). Data were collected as part of a larger on-going effort to develop and validate a cross-cultural competence assessment test battery for the U.S. Army. SMEs rated response options using a six-point response scale and were not forced to rank each response within each item.

A total of 631 novice (MTurk) participants[1] completed the SJT rating task along with a battery of 24 assessments over the course of five, one-hour blocks. The majority of the sample was female (~53%), White (~76%), and between 20 and 40 years of age (~73%). Approximately 98% of participants were native English speakers. Unlike SMEs, MTurk raters ranked each of the five response options in order of their effectiveness (1=*most effective* to 5=*least effective*) for all 30 SJT items. In order to compare SME and MTurk ratings, we converted SME ratings into within-item ranks after removing any instances where a SME did not provide a rating for all five options within a single item.

Reliability and validity evidence for the Military SJT used in this study were previously reported by Martin-Raugh et al. (2018). This SJT was designed based on 203 critical incidents obtained from U.S. soldiers and response options were created by an independent sample of

[1] This sample of novices represents a subset of the sample of individuals that completed the SJTs in a study conducted by Martin-Raugh et al. (2018).

SMEs. Martin-Raugh et al. observed strong internal consistency for the Military SJT (α = .94, ω = .82). The authors also found that scores on the Military SJT were positively related to performance on an interpersonal skills SJT (r = .54) and an empathic concern measure (r = .34), and that SJT scores were correlated with self-reported cultural competence (r = .23), supporting the construct-related validity of the measure. Results from an on-going study (Kochert, 2021) using the Military SJT to predict cross-cultural performance in Soldiers stationed overseas has found evidence supporting the relationship between the Military SJT and supervisor ratings of cross-cultural performance. Specifically, Military SJT scores were positively related to the cross-cultural performance dimensions of: Applying Cultural Knowledge (r = .19, p = .02, n = 146), Strengthening Cross-Cultural Relationships (r = .17, p = .05, n = 145), and Influencing Across Cultures (r = .18, p = .03, n = 144).

Individual Difference Measures

GMA. In Sample 1, we measured GMA using the Wonderlic Personnel Test (WPT; Wonderlic, 2002). The WPT is a speeded, 50-item test which is widely used as a measure of GMA in both research and practice (Schmidt & Hunter, 2004). Test-retest reliability estimates for the WPT are reported to range from 0.82 to 0.94 (Wonderlic, 2002). In Sample 2, we administered a 30-item version of the WPT (WPT-Q; Wonderlic, 2003). In Sample 3, we used both self-reported college GPA and educational attainment as proxies for GMA. Both measures have been suggested as suitable alternatives for assessing GMA in place of a traditional GMA test (e.g., Ployhart & Holtz, 2008; Wai et al., 2018). Educational attainment was reported on a nine-point scale. Most participants held at least a college degree (~66%). Participants reported their college GPA using a six-point scale ranging from 1 = “Below 1.5” to 6 = “3.5-4.0 or greater”.

Personality. We used different self-report measures of personality traits within each sample. Although we focus on the trait of conscientiousness to test our RQ 2, we were also able to collect data on the remaining personality traits from the Five-Factor and HEXACO models in Samples 1 and 3. A subset of participants in Sample 1 completed the IPIP version of the six-factor HEXACO personality scales (Ashton et al., 2007; Honesty-Humility $\alpha = .91$, Emotionality $\alpha = .91$, Extraversion $\alpha = .95$, Agreeableness $\alpha = .96$, Conscientiousness $\alpha = .95$, and Openness $\alpha = .94$). In Sample 2, we measured Conscientiousness using a 10-item scale developed by Goldberg (1992). In Sample 3, personality traits were measured using the Big Five Inventory (BFI-44; John et al., 2008). The BFI-44 generates trait scores for each of the primary Big Five dimensions (Agreeableness $\alpha = .84$, Conscientiousness $\alpha = .89$, Extraversion $\alpha = .89$, Neuroticism $\alpha = .91$, and Openness $\alpha = .85$).

Expert similarity. We measured an individual's similarity to expert ratings by calculating the mean absolute deviation (AD_M) between each individual's rating and the average SME rating across all items. AD_M is a commonly used method of assessing interrater agreement (Burke et al., 1999). According to Burke and Dunlap (2002), an $AD_M < 0.80$ is considered to indicate high agreement when using a five-point scale. When testing whether any of our hypothesized individual difference variables were correlated with expert similarity (RQs 2 and 3), we transformed expert similarity scores by multiplying them by –1 (e.g., $AD_M * -1$) so that greater values represented stronger similarity.

Results

Comparing Response Option Ranking between Expert and Novice Raters

We first observed agreement between MTurker and SME raters by determining the classification consistency of the within-item rankings of response options between groups (Table

1). Here, we found a strong correlation between the within-item ranks across all response options (Sample 1 $r = .94$; Sample 2 $r = .88$, $p < .001$; Sample 3 $r = .96$, $p < .001$). Across the three SJTS, between 63 – 85% of response options received identical, within-item rankings between SMEs and MTurk raters. In Sample 1, we observed perfect agreement between MTurk raters and SMEs for response options that were considered the most effective and second most effective. However, we observed weaker agreement among the remaining response options (64%). In comparison, in Samples 2 and 3, rankings were most consistent for options considered least effective (Sample 2 = 88%; Sample 3 = 100%) or most effective (2 = 74%; Sample 3 = 87%). In comparison, we observed the weakest agreement for the second (Sample 2 = 37%; Sample 3 = 73%) and third-best response options (Sample 2 = 46%; Sample 3 = 77%).

In all three samples, we observed a strong correlation between mean response option ratings between SMEs and MTurk raters (Sample 1 $r = .94$; Sample 2 $r = .91$, $p < .001$; Sample 3 $r = .88$, $p < .001$). These correlations were equally strong despite differences in MTurk rater sample size and SJT content between each sample. We also compared AD_M relative to the average SME rating for SME and MTurk raters in each our samples (Table 2). We found greater agreement among SMEs compared to MTurk rater agreement with SME judgments in Sample 1 ($d = –1.39$, $F(1,113) = 9.22$, $p = .003$), Sample 2 ($d = –1.46$, $F(1,56) = 30.89$, $p < .001$) and Sample 3 ($d = –0.42$, $F(1,694) = 9.18$, $p = 0.003$). In particular, we observed less variability among Medical SMEs (SD = 0.05) compared to MTurk raters (SD = 0.20). We also observed this difference for SME (SD = 0.04) and MTurk raters (SD = 0.24) for the Communication SJT. This suggests that there was stronger consensus among SMEs compared to MTurk raters for the Medical and Communication SJTs. However, the standard deviations for Military SMEs (SD =

0.19) and MTurk raters (SD = 0.20) were roughly equivalent which suggests a similar rate of agreement was among SME and MTurk raters despite different sample sizes.

We also estimated within-group agreement for SME and MTurk raters separately by calculating the mean absolute deviation in each sample (Table 3). We found the strongest degree of mean within-group agreement among SMEs for the Communication (AD_M = .45) and Medical SJT (AD_M = .62). In both cases, we also observed greater within-group agreement among SME raters compared to MTurk raters. In contrast, we found no difference in within-group agreement among SMEs and MTurk raters for the Military SJT (d = 0.14, $F(1,298)$ = 1.53, p = .22). These results further indicate that SME raters often demonstrate stronger agreement, potentially due to shared job knowledge or norms, but that large samples of crowdsourced raters may achieve similar levels of agreement based on the wisdom of the crowd.

Individual Differences in Expert Similarity

Next, we examined the correlation between expert similarity and GMA within each sample of MTurk raters. All correlational results for each sample are reported in Table 4. GMA as measured by the WPT was strongly correlated with rater accuracy in both Samples 1 (r = .52, $p < .05$) and 2 (r = .40, p = .03). In Sample 3, only one of our two GMA proxy measures (self-reported college GPA) was positively related to rater accuracy (r = .16, $p < .001$). Regarding personality, we found that Conscientiousness was positively related to expert similarity within each of our three samples (Sample 1 r = .37, $p < .01$; Sample 2 r = .41, $p < .05$; Sample 3 r = .27, $p < .001$). Moreover, we further explored the relationships between the remaining personality dimensions and expert similarity in Samples 1 and 3. In Sample 1, only Honesty-Humility (r = .26, $p < .05$) and Openness to Experience (r = .34, $p < .05$) were found to be positively related to expert similarity. Likewise, in Sample 3, greater expert similarity scores were found for MTurk

raters high in Agreeableness (r = .27, p < .001) and Openness to Experience (r = .09, p < .05) but low in Extraversion (r = –.20, p < .001). Thus, the personality traits of Agreeableness, Openness, or Honesty-Humility may help identify MTurk raters who could provide judgments that are more aligned with SMEs, but the utility of these traits appears to vary somewhat based on the content of the SJT. In comparison, we observed that MTurk raters with greater GMA or Conscientiousness were generally more similar to expert raters, regardless of SJT content.

Even though several of our individual difference measures were correlated with rater accuracy, we also found that the aggregate MTurk rating was more similar to SME ratings than almost all individual MTurk raters in each sample (Sample 1: aggregate AD_M = 0.32, Sample 2: AD_M = 0.40, Sample 3: AD_M = 0.45). This suggests that the accuracy of crowd estimates may be affected more by the sheer number of raters rather than the ability or characteristics of the individual raters themselves. To this end, we conducted a bootstrapping analysis to estimate the number of individual MTurk raters that may be needed to provide a reliable estimate of SME-derived scoring keys within each of our three samples. This analysis was conducted by estimating the average absolute rater error (AD_M) for 1,000 randomly drawn sample sizes ranging from n = 5 to n = 100 with replacement. Due to a smaller total sample size, we could only estimate the average error for sample sizes up to n = 30 for the Medical SJT in Sample 2. This procedure was conducted for each of the SJTs and the results are reported in Figure 2. Across all three samples, the largest decreases in rater error were found when increasing the number of MTurk raters from 5 to 20. These results were consistent regardless of the content of the SJTs themselves. For both the Communication and Military SJTs, obtaining more than 40 individual raters appears to provide diminishing returns in reducing rater error. The decrease in rater error from adding additional raters appears to reach an asymptote when the total sample

size approaches between n = 25 to 30 raters across all three samples. Based on these results, a crowd size of roughly n = 30-40 raters appears to be sufficient to obtain results that approximate SME ratings. Gathering additional raters beyond this point appears to yield relatively little further improvement in accuracy.

Discussion

Our goal in the present study was to determine whether we could replicate SME ratings of SJT items using crowdsourced samples of novice raters recruited from MTurk. In order to observe whether the usefulness of this method varied based on SJT content, we tested this approach using three different SJTs. A consistent finding across our three samples is that crowdsourcing appears to be a viable alternative or complement to SME scoring for many SJTs regardless of the content of the test. These findings replicate the results of past research where CBM was found to produce scoring keys that were highly correlated with expert judgment (Legree et al., 2005). As individuals, novice raters were less able to reliably identify the effectiveness of SJT response options compared to SMEs. Yet, aggregated MTurk ratings generally produced a similar rank order of response options within each SJT item compared to SME ratings (correlations ranged from r = .88 to .94). Thus, it appears that SJT scoring keys derived from crowdsourced samples seem likely to converge with keys based on SME judgment.

Our findings are far from revolutionary. The past work of Motowidlo and colleagues (2010, 2018) has shown that much of the variance in SJT performance can often be attributed to general domain knowledge or implicit trait theories. Yet, convenience samples of SMEs, typically simply a group of incumbents, are often used to generate SJT scoring keys without a more thorough vetting of their expertise. We do not recommend eliminating SMEs from SJT scoring entirely, but we believe that our work shows that crowdsourcing can be a useful

alternative or supplement to SME-based methods. This may be especially true for situations where SMEs may be difficult to identify or costly to obtain. In medical contexts, where SJTs are increasingly used for selection into training programs (Lievens, 2013), SMEs may be easy to identify (based on licensure or educational attainment) but very expensive to recruit. Crowdsourced samples may lack face validity but can be used to approximate SME judgment in aggregate. In a rare example, Teng and colleagues (2020) used crowdsourcing to develop the scoring key for their SJT measuring resilience. Moreover, SME and crowdsourced ratings could be compared in order to identify items that discriminate the most between experts and novices. For instance, in the example Medical SJT item (shown in Figure 1, 56% of SMEs identified the consensus, least effective option. Although this was also chosen as the least effective option by a consensus of MTurk raters, this only amounted to 29% of responses. However, MTurk raters were more likely to identify the most effective response (54%) compared to SMEs (41%). Future research is needed to help determine when crowdsourcing is best suited for constructing SJTs.

The crowdsourcing methods that we use in the present study are especially well-suited for designing domain-general SJTs. Although SJTs have traditionally been designed to assess job-specific knowledge or expertise (Motowidlo et al, 1990), recent research has demonstrated that the SJT method can be used to measure more general psychological constructs (Guenole et al., 2017; Tiffin et al., 2020). These domain-general SJTs include validated measures of personality (Oostrom et al., 2019), integrity (Becker, 2005), and social or emotional intelligence (MacCann & Roberts, 2008; Speer et al., 2019). In these contexts, it is more useful to validate SJT content based on empirical prediction with criterion measures (e.g., the psychological construct of interest) rather than job-specific expertise. Therefore, large samples of crowd workers can be used to test SJT item pools and construct forms which most strongly predict

alternate measures of the criterion construct. We hope that our findings inspire future researchers to use crowdsourcing methods to design more reliable and valid construct-focused SJTs.

Among the three tests that we examined, we generally found the greatest disparities between SMEs and novice MTurk raters on the Medical SJT. Although aggregated MTurk ratings still corresponded with SME rankings of response options 61% of the time, this classification consistency was the lowest of the three tests and the average individual MTurker was less accurate relative to the SME sample. This SJT was created based on critical incidents from a specific job context (emergency medicine departments). As such, many of these SJT items use job-specific language and terminology. Moreover, May and Prewett (2016) previously observed a significant advantage for medical professionals compared to students in their validation research. These results suggest that the Medical SJT is the most job-related among the three tests in our study. However, due to our study design we are not able to determine whether this difference should be attributed to the job-relatedness of the SJT items or the expertise of the SMEs.

In the future, comparisons between SME and crowdsourced raters may help provide further construct validity for the job-relatedness of SJTs. Currently, SME judgment is considered by many as evidence for construct validity, especially when the test is designed for a specific occupation or work context. Content validity is useful for justifying the use of a test in practice, even if it does not necessarily indicate criterion-related validity (e.g., Guion, 1978; Murphy, 2009). In comparison, it seems unlikely that crowdsourced ratings would be perceived as equally content valid. Yet, our results suggest that crowdsourcing may yield similar rank order results as those obtained from SMEs. Therefore, we recommend that crowdsourcing would be most useful early in test development process when large pools of items are being tested. The crowdsourced

keys could be used to identify the best performing items to be included in the final version of the test. Afterwards, SMEs could rate and review this final subset of items. This would reduce the amount of time and effort for SMEs in the test development process and act as a way of providing construct validity for the scoring of the items.

Crowdsourcing and Individual Differences

In all three samples, individual differences in cognitive ability and conscientiousness were each related to stronger similarity with expert (SME) ratings. These findings suggest that these individual difference variables are generally useful for identifying the most desirable responses to SJT items. This conclusion is supported by past research which has generally reported similar correlations between individual differences in personality and GMA with SJT performance (McDaniel et al., 2001). Our findings also suggest that performance on SJTs often reflects individual differences beyond procedural knowledge about a specific job or task. Motowidlo and Beier (2010) proposed that people higher in mental ability tend to better learn what trait expressions are likely to be effective in situations described in SJT items. Consequently, they argue that more intelligent individuals develop ITPs that are more closely aligned with the true relations between trait expression and effectiveness in a particular domain. ITP theory also posits that people tend to think that actions expressing their own personality traits are more generally effective (Motowidlo, 2003). Our findings provide further support for these aspects of ITP theory and suggest that performance on SJTs may often depend less on procedural knowledge gained from specific, work-related experiences and more on basic individual differences.

We also observed that MTurker conscientiousness was positively related to expert similarity scores across all three samples. These results suggest that the trait of conscientiousness

is relevant to a variety of interpersonal situations or norms for behavior in different workplace settings (Martin-Raugh & Kell, 2021). This trait is likely also important for the task of reading and rating SJT content itself. Not only are highly conscientious individuals expected to outperform others in most job roles (Judge et al., 2001), but they are also characterized as being highly attentive to detail. According to trait activation theory (Tett & Burnett, 2003), this tendency may be especially useful when reading and rating SJT items. In contrast, other traits (e.g., extraversion or agreeableness) may not be as directly related to the task of completing SJT items which would explain some of the variability in correlations observed in Samples 1 and 3.

Despite the observed effects of GMA and personality, it is important to note that we also found that nearly no individual MTurk rater could achieve or exceed the accuracy observed for the aggregate MTurker in all three samples. Individual MTurk raters generally performed worse compared to individual SMEs, especially in Samples 2 and 3. This consistent finding is a replication of the "wisdom of the crowd" effect where the performance of the group is found to exceed the performance of any one individual group member. Instead of using individual difference measures in order to find the best judges, the size of the crowd appears to be more important for obtaining judgments which best align with the experts. This is similar to past crowdsourcing research which has reported that screening crowd members yields little, if any, increase in accuracy (e.g., de Oliveria & Nisbitt, 2018) and that individual crowd member performance is often inconsistent (e.g., Bayus, 2013). Similar findings have also been found in other crowdsourcing studies where small groups of randomly selected MTurk workers were found to outperform individual workers (Vercammen et al., 2019). However, future research is still needed in order to observe how strongly these findings generalize to other SJTs requiring a greater degree of procedural job knowledge or technical expertise.

Implications for Research and Practice

Our study has several implications for future research and development of SJTs in practice. In particular, our findings call into question some long held assumptions about the necessity of subject matter expertise. We do not believe that these results indicate that SMEs do not possess expert judgment or that expertise is not necessary for scoring SJTs. Instead, we find that crowdsourcing can be used to estimate expert judgment even if individual crowd members may not be experts themselves. Individual MTurk raters may not be as knowledgeable as individual SMEs, but our results suggest that crowdsourcing methods (e.g., CBM) are capable of approximating expert knowledge using aggregation. We found that the largest gains in predicting SME ratings could be obtained with as few as 30-40 randomly sampled MTurk raters. In practice, these crowdsourcing methods are highly cost-effective compared to relying on SME judgment. Instead of evaluating crowd workers in order to find those who have greater expertise (e.g., Zhang & Wang, 2021), our results suggest that aggregate results may be just as effective without screening. This may enable the development of larger SJT item pools which could help solve longstanding problems of poor internal consistency and overall construct measurement (McDaniel et al., 2016).

Moreover, prior research has suggested that the average cost of developing a job-specific text-based SJT for a particular organization from scratch ranges from $60,000 to $120,000 (Lievens et al., 2008). It may also be nearly as expensive to conduct a full-scale validation study to generate criterion-based scoring keys. As our findings suggest that crowdsourced methods of developing SJTs may produce scoring keys that are similar to those generated using more costly and difficult to source SMEs, it may be viable to produce SJTs in a more cost-effective manner with greater facility. Although we expect that SMEs will continue to play an important role in

SJT development, we believe that their time and effort is best spent providing critical incidents and possible behavioral responses to help generate an initial item pool. These items can then be evaluated using relatively cheaper and convenient crowdsourced samples. Results from the crowdsourced sample would be used to refine the test and remove any poorly-performing or redundant items. If the SJT is meant to assess specific job knowledge, the shortened version to a small sample of SMEs in order to establish whether the items can distinguish between novice crowd workers and SMEs. We believe that crowdsourcing will help alleviate some of the demand on SMEs in SJT development while also providing greater statistical power for conducting item analyses. This would also ensure that developers can establish validity by testing for differences between novices and experts. However, future research should investigate whether crowdsourced keys yield criterion validities similar to those produced by SME-generated keys.

Although our study specifically focused on using crowdsourcing methods for a specific task (scoring SJT items), we believe that our findings suggest that crowdsourcing could be a useful alternative to SME judgment in other tasks. Crowdsourcing methods have been used in other aspects of test development including item writing (Sadler et al., 2016). A growing number of organizations also use crowdsourced ratings, gathered by third-party companies such as Glassdoor, in order to estimate employee job satisfaction without administering an internal organizational survey (Landers et al., 2019). Despite the popularity of crowdsourcing methods in other areas of research and practice, many aspects of industrial and organizational psychology still rely heavily on SME judgment (e.g., job or task analysis). Yet, the criteria for what actually constitutes subject matter expertise tends to be subjective and is rarely quantified. Zhang and Wang (2021) have recently proposed a promising new method for quantifying expertise in SJT

performance but our three studies did not use a repeated-measures design which is necessary to use their methodology. Further research is needed to determine how this approach may help inform the use of crowdsourcing methods to estimate expert judgment and whether it can be used to differentiate between SMEs and crowd workers. We also hope that the results from our present study help to promote greater use of crowdsourcing methods for future organizational research and practice more generally.

Directions for Future Research

SJTs are often assumed to be job-specific due in part to the face validity of item content based on actual critical incidents. However, this does not necessarily mean that the test only measures procedural knowledge (Krumm et al., 2015). Some scholars have argued that interpersonally-focused SJTs mostly measure general knowledge or norms regarding acceptable workplace behavior (e.g., Motowidlo et al., 2018; Salter & Highhouse, 2009). Across our three samples, we found that MTurk raters could be used to identify effective and ineffective behaviors carried out in response to social situations, even in specific occupational contexts (e.g., the U.S. military or emergency medicine departments). Although these contexts may each have their own specific norms or culture regarding how people are expected to interact with one another, these expectations may not be truly unique. To this end, past research suggests that some SJT content may even be equally valid across different national cultures (Lievens et al., 2015). However, Schmitt and colleagues (2019) found some differences in SJT responding between American and Chinese test takers based on cultural content. Thus, future research is needed to help empirically evaluate the extent to which SJT content potentially requires job knowledge, cultural or organizational norms, and other specific forms of knowledge.

Even though our crowdsourced scoring keys correlated strongly with expert-based keys, it is unclear whether there are any differences in criterion-related validity between the two methods. As noted by McCornack (1956), substantial differences in validity are still possible as long as the correlation between scoring keys is less than unity ($r < 1.0$). Past work has focused on comparing different methods of scoring SJT responses (e.g., McDaniel et al., 2011; Weng et al., 2018) but rarely have researchers studied differences in scoring keys between different rating sources. In a rare exception, Bergman and colleagues (2006) found that the SME-based key accounted for incremental prediction of leader effectiveness ratings beyond other keying methods. However, this study did not directly compare the SME key to a scoring key derived only from novice ratings. Motowidlo and Beier (2010) found that both expert-based keys produced significant validities in predicting supervisory ratings of job performance but only explained a small amount of incremental variable compared to a key based on undergraduate student judgments. These results suggest that inexperienced novices can be used to develop knowledge predictive of job performance even if they have had no prior experience in a particular job, providing further support for the notion that inexperienced, crowd-sourced samples may be able to generate SJT scoring keys that have predictive validity. Yet, further research is needed to determine the boundary conditions where aggregate novice ratings fail to correspond with expert judgment. For example, crowdsourcing may not be suitable for SJTs meant for licensure or credentialing purposes.

Likewise, future research is also needed to examine how crowdsourcing performs compared to criterion-based methods, such as empirical keying. Generally speaking, empirical keying tends to provide stronger criterion-related validity for various types of assessments, including biodata scales (Cucina et al., 2012) and self-report personality measures (Cucina et al.,

2019). Along these lines, Guenole and colleagues (2017) recommend using criterion-keying when scoring SJTs. However, practitioners often lack the necessary time, samples, or resources to collect validation data and tend to rely on expert judgment or content validation approaches instead (Tan, 2009). Especially in the early stages of test development, it may not be feasible to conduct a suitable validation study in order to create criterion keys. Instead, crowdsourcing could potentially be used early in the development process while more costly validation studies could be conducted after initial rounds of item analysis and the removal of poorly performing items. Thus, it would be highly valuable to determine the extent to which crowdsourced keys could be expected to converge with criterion-based keys. Future research should continue to investigate the utility of deriving crowdsourced scoring keys by comparing their criterion-related validity with empirical keys. This work may also help identify whether job-specific SJT content provides any additional criterion-related validity compared to SJT content that is more general in nature.

Our results indicate that novice, crowdsourced workers are most effective at identifying the best and worst behavioral responses to SJT items but it is not clear whether similar results would be found if each response option was evaluated independently. This is important to consider given that some researchers have introduced single-response SJTs, where only one behavioral response is presented at a time, as an alternative to those using the traditional, multiple-choice format (Crook et al., 2011; Martin-Raugh et al., 2018). It may be the case that novices are more able to identify the best and worst options when given all options to evaluate and compare, as suggested by Krumm et al (2015) and as observed by Robertson et al. (2020) when conducting open and closed cart sort tasks. To explore this question, we conducted secondary analyses of the single response SJT rating data from Martin-Raugh et al (2018). In this study, 659 novice MTurk raters completed a 30-item, single-response SJT which was a subset of

the multiple-response Military SJT used in our study. The average deviation between individual MTurk and SME raters for these single-response items was slightly greater (AD_M = 1.07, SD = 0.24) than what we observed among MTurk raters who completed the same items as part of a multiple-choice SJT (AD_M = 0.84, SD = 0.26). However, we found a similar degree of agreement between the aggregate MTurk and SME ratings for the single-response (AD_M = 0.41, SD = 0.25) and multiple-response SJTs (AD_M = 0.48, SD = 0.32). Based on this post-hoc analysis, the single-response format may demonstrate stronger differences between SMEs and individual novice raters but the overall crowd estimates appear equally effective in estimating the effectiveness of SJT response options across both SJT formats. Yet, future research is still needed to determine whether these initial findings will generalize to other tests or domains.

Another direction for future research is applicant or stakeholder reactions to crowdsourcing methods of scoring SJTs. Even though crowdsourcing methods are widely adopted in other fields and our results suggest that they may yield useful estimates of expert judgment, it is unclear how they would be perceived in practice. Past research suggests that individuals often tend to place more trust in expert judgment when evaluating job candidates compared to standardized assessments or algorithmic decision making (Highhouse, 2008). Even though using selection criteria for crowd members may not noticeably improve the accuracy of crowd-based judgments, past research suggests that these criteria can make crowds appear more trustworthy (e.g., Mannes et al., 2014). Thus, organizational stakeholders may perceive the use of selection criteria as more favorable even if it may not greatly improve rating quality. Moreover, it may be best to use crowdsourcing in tandem with more traditional, SME-based judgments in order to ensure construct validity while also minimizing development time and costs.

Study Limitations

Despite the relevance of our findings for SJT research and practice, it is important to acknowledge some of the limitations of the present study. One limitation is that each of the SJTs that were used in the present study were all newly developed and had little pre-existing criterion-related validation data. Although the Communication SJT was only developed as part of a pilot project, we report construct validity evidence for the Medical and Military SJTs which have been presented as part of conference presentations (May & Prewett, 2016) or published in peer-reviewed journals (Martin-Raugh et al., 2018). It would be especially insightful to observe whether the within-item agreement between SMEs and crowd estimates has any relationship with the validity of the item (e.g., a point-biserial correlation with a criterion measure). However, the SJTs that we used in this study are representative of many existing tests that are used in practice. Not only are SJTs commonly used in military and medical professions (e.g., Lievens & Sackett, 2012), many SJTs are designed to measure applied social or interpersonal skills (Christian et al., 2010). In addition, all three SJTs were created based on SME reports of critical workplace incidents relevant to the construct of interest. This follows common practice for the development of job-specific and domain-general SJTs in academic and commercial settings (e.g., Christian et al, 2010; Guencole et al., 2017).

Another limitation of our study is that the specific methodological design for each sample was not fully consistent. Across the three samples, we obtained varying numbers of SME and MTurk raters, used different defining characteristics when identifying SME raters, used different measures of individual differences among the MTurk raters, and administered three different SJTs. These differences are due to differences in resources for each of the practical SJT test development projects for which these data were collected. It is possible that using a more

narrowly defined set of SMEs may result in greater consistency in judgments due to greater similarity in perspective, past experience, or training. Likewise, the gap between the average SME and crowdsourced rater may be larger for SJTs which require more job-specific expertise. Future experimental research is needed to help answer these questions, however, our results still suggest that crowdsourcing via MTurk yielded similar results to our most specific set of SMEs. Moreover, we argue that our ability to replicate many of our findings despite these differences in research design suggest that these effects are generalizable to different types of SJTs, different conceptualizations of SMEs, and across different practical conditions faced by researchers and practitioners.

Lastly, it is not clear whether our results should generalize to more domain-specific SJTs. However, SJTs are often designed to assess the understanding of effective responses to interpersonal interactions in work settings and it is not always clear whether these tests distinctly measure domain-specific or domain-general knowledge (Motowidlo et al., 2018). For example, several of the Medical SJT questions focus on managing interpersonal conflict in an emergency medicine setting (see sample item in Figure 1). The effectiveness of each response may be determined by specific rules and behavioral norms in this line of work, as identified by the SME raters, but these may not be truly unique to emergency medicine. These norms may be shared across similar occupations (e.g., other health care or service professions) or informed more broadly by cultural expectations for behavior at work (e.g., favoring prosocial or agreeable behavior).

Conclusions

Overall, we found that crowdsourcing could be used to create scoring keys that strongly converged with keys based on SME judgment. We found that aggregate crowdsourced SJT item

ratings were strongly correlated to those from three different samples of SMEs, even though individual MTurk raters were generally less accurate than individual SMEs. Moreover, we were able to replicate these results across three different SJTs varying in job-relatedness. Although our results highlight one specific application of crowdsourcing in the development of SJTs, we believe that our findings may also generalize to similar tasks where SME judgments are often used. Crowdsourcing is often a quicker and cheaper alternative to recruiting SMEs and can potentially be a useful tool in test development or in other tasks traditionally requiring subject matter expertise. We hope that our findings inspire other researchers and practitioners to use crowdsourcing methods in future organizational research.

Disclaimer

The research described herein was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army (Cooperative Agreement No. W911NF-18-2-0018. The views expressed in this presentation are those of the author and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request**.**

References

Ashton, M. C., Lee, K., & Goldberg, L. R. (2007). The IPIP-HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences*, *42*, 1515–1526.

Baker, D. P., Amodeo, A. M., Krokos, K. J., Slonim, A., & Herrera, H. (2010). Assessing teamwork attitudes in healthcare: Development of the TeamSTEPPS teamwork attitudes questionnaire. *Quality and Safety in Health Care*, *19*, e49. https://doi.org/10.1136/qshc.2009.036129

Barchard, K. A., Hensley, S., & Anderson, E. (2013). When proportion consensus scoring works. *Personality and Individual Differences, 55*, 14–18.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go from here. *International Journal of Selection and Assessment, 9*, 9–30.

Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell Ideastorm community. *Management Science*, *59*, 226–244.

Becker, T. E. (2005). Development and validation of a situational judgement test of employee integrity. *International Journal of Selection and Assessment*, *13*, 225–232.

Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, *24*, 285–300.

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*, 223–235.

Black, E. W., Schrock, B., Prewett, M. S., & Blue, A. V. (2021). Design of a situational judgment test for preclinical interprofessional collaboration. *Academic Medicine*, *96*, 992–996.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, *62*, 229–258.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, *10*, 99–109.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods*, *5*, 159–172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*, 49–68.

Byun, T. M., Harel, D., Halpin, P. F., & Szeredi, D. (2016). Deriving gradient measures of child speech from crowdsourced ratings. *Journal of Communication Disorders*, *64*, 91–102.

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

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*, 347–361.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83–117.

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.

Cucina, J. M., Caputo, P. M., Thibodeaux, H. F., & MacLane, C. N. (2012). Unlocking the key to biodata scoring: A comparison of empirical, rational, and hybrid approaches at different sample sizes. *Personnel Psychology*, *65*, 385–428.

Cucina, J. M., Vasilopoulous, N. L., Su, C., Busciglio, H. H., Cozma, I., DeCostanza, A. H., Martin, N. R., & Shaw, M. N. (2019). The effects of empirical keying of personality measures on faking and criterion-related validity. *Journal of Business and Psychology*, *34*, 337–356.

Davis-Stober, C. P., Budescu, D. V., Broomwell, S. B., Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*, 130–143.

De Oliveria, S., & Nisbitt, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogenous crowds. *Proceedings of the National Academy of Sciences*, *115*, 2066–2071.

de Kock, F. S., Lievens, F., & Born, M. P. (2020). The profile of the 'Good Judge' in HRM: A systematic review and agenda for future research. *Human Resource Management Review*, *30*, 100667.

de Vries, R. E., Barends, A. J., & de Kock, F. S. (2021). Dispositional insight: Its relations with HEXACO personality and cognitive ability. *Personality and Individual Differences*, *173*, 110644. https://doi.org/10.1016/j.paid.2021.110644

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., & Nosek, B. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*, 15343–15347.

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*, 562–571.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26–42.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79–132.

Guenole, N., Chernyshenko, O. S., Weekley, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*, 234–252.

Guion, R. (1978). "Content validity" in moderation. *Personnel Psychology*, *31*, 205–213.

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S., & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *Leadership Quarterly*, *14*, 117–140.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*, 333–342.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*, 104–121.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*, 40-46.

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E. J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*, 267–285.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*, 828–845.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology*, *13*, 373–386.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk (Technical Report CeDER-10-01). Retrieved from http://www.ipeirotis.com/?publication=demographics-of-mechanical-turk.

John, O. P., Naumann, L., & Soto, C. J. (2008). The Big Five trait taxonomy: Discovery, measurement, and theoretical issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed.). New York, NY: Guilford Press.

Kaminski, K., Felfe, J., Schapers, P., & Krumm, S. (2019). A closer look at response options: Is judgment in situational judgment tests a function of the desirability of response options? *International Journal of Selection and Assessment*, *27*, 72–82.

Kochert, J. (2021). [Unpublished raw data of cross-cultural competencies]. U.S. Army Research Institute for the Behavioral and Social Sciences.

Krumm, S., Lievens, F., Huffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology*, *100*, 399–416.

Landers, R. N., Brusso, R. C., & Auer, E. M. (2019). Crowdsourcing job satisfaction data: Examining the construct validity of Glassdoor.com ratings. *Personnel Assessment and Decisions*, *5*, doi: 10.25035/pad.2019.03.006

Landy, J. F., Jia, M. (L.), Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., . . . Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin, 146*, 451–479.

Lee, K., Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, *39*, 329–358.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence*, *21*, 247–266.

Legree, P. J., Heffner, T. S., Psotka, J., Medsker, G. J., & Martin, D. E. (2003). Traffic crash involvement: Experiential driving knowledge and stressful contextual antecedents. *Journal of Applied Psychology*, *88*, 15–26.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 155–179). Cambridge, MA: Hogrefe & Huber.

Lendvay, T. S., White, L., & Kowalewski, T. (2015). Crowdsourcing to assess surgical skill. *JAMA Surgery*, *150*, 1086–1087.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of

matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90*, 442–452.

Lievens, F., Corstjens, J., Sorrel, M. A., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain? *International Journal of Selection and Assessment*, *23*, 361–372.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, *9*, 3–22.

Lievens, F., Patterson, F., Corstjens, J., Martin, S., & Nicholson, S. (2016). Widening access in selection using situational judgement tests: Evidence from the UKCAT. *Medical Education*, *50*, 624–636.

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

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology*, *97*, 460–468.

Luengo-Oroz, M. A., Arranz, A., & Frean, J. (2012). Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research, 14*, e167.

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*, 540-551.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*, 276–299.

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

Martin-Raugh, M. P., & Kell, H. J. (2021). A process model of situational judgment test responding. *Human Resource Management Review*, *31*, 10073.

May, T. A. (2015). *Development of a situational judgment test for teamwork in medicine*. [Master's Thesis, Central Michigan University]. https://scholarly.cmich.edu/cgi-bin/cmuscholarly?a=d&d=CMUGR2015-61&_ga=2.184915055.791516209.1626887241-735048095.1626887241

May, T. A., & Prewett, M. S. (2016). *Development of a situational judgment test for teamwork in medicine*. Poster presented at the 31st Meeting for the Society for Industrial and Organizational Psychology; Anaheim, CA.

Mayer, J. D., Panter, A. T., & Caruso, D. R. (2020). When people estimate their personal intelligence who is overconfident? Who is accurate? *Journal of Personality*, *88*, 1129–1144. https://doi.org/10.1111/jopy.12561

McCornack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology*, *40*, 343–344.

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*, 509–516.

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.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730–740.

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*, 327–336.

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

Motowidlo, S. J. (2003). Job performance. In W. C. Borman, D. R. Ilgen, and R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 39–53). New York, NY: Wiley.

Motowidlo, S. J. & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, *95*, 321–333.

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

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182*).* Mahwah, NJ: Erlbaum.

Motowidlo, S. J., Lievens, F., & Ghosh, K. (2018). Prosocial implicit trait policies underlie performance on different situational judgment tests with interpersonal content. *Human Performance*, *31*, 238–254.

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology*, *2*, 453–464.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Oostrom, J. K., de Vries, R. E., & De Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*, 1–29.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*, 187–207.

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Medical Education*, *43*, 50–57.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*, 153–172.

Revelle, W., Dworak, E. M., & Condon, D. (2020). Cognitive ability in everyday life: The utility of open-source measures. *Current Directions in Psychological Science*, *29*, 358–363.

Robertson, I., Kortum, P., Oswald, F. L., & Acemyan, C. Z. (2020). Novices Perform Like Experts on a Closed Card Sort but Not an Open Card Sort. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64, 1249–1253.

Rogers, K. H., & Biesanz, J. C. (2019). Reassessing the good judge of personality. *Journal of Personality and Social Psychology*, *117*, 186–200.

Sadler, P. M., Sonnert, G., Coyle, H. P., & Miller, K. A. (2016). Identifying promising items: The use of crowdsourcing in the development of assessment instruments. *Educational Assessment*, *21*, 196–214.

Salter, N. P., & Highhouse, S. (2009). Assessing managers' common sense using situational judgment tests. *Management Decision*, *47*, 392–398.

Schäpers, P., Mussel, P., Lievens, F., Konig, C. J., Freudenstein, J. P., & Krumm, S. (2020). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology*, *105*, 800–818.

Schmidt, F. L., & Hunter, J. (1992). Development of a causal model of processes determining job performance. *Current Directions in Psychological Science*, *1*, 89–92.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*, 162–173.

Schmitt, N., Prasad, J. J., Ryan, A. M., Bradburn, J. C., & Nye, C. D. (2019). Culture as a determinant of option choice in a situational judgment test: A new look. *Journal of Occupational and Organizational Psychology*, *92*, 330–351.

Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O., & Kiessling, C. (2008). A situational judgement test of professional behaviour: Development and validation. *Medical Teacher*, *30*, 528–533.

Speer, A. B., Christiansen, N. D., & Laginess, A. J. (2019). Social intelligence and interview accuracy: Individual differences in the ability to construct interviews and rate accurately. *International Journal of Selection and Assessment*, *27*, 104-128.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY: Doubleday.

Tan, J. A. (2009). Babies, bathwater, and validity: Content validity is useful in the validation process. *Industrial and Organizational Psychology*, *2*, 513–515.

Teng, Y., Brannick, M. T., & Borman, W. C. (2020) Capturing resilience in context: Development and validation of a situational judgment test of resilience. *Human Performance*, *33*, 74–103.

Tenopyr, M. L. (1969). The comparative validity of selected leadership scales relative to success in production management. *Personnel Psychology*, *22*, 77–85.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–517.

Tiffin, P. A. Paton, L. W., O'Mara, D., MacCann, C., & Lang, J. W. B. (2020). Situational judgment tests for selection: Traditional versus construct-driven approaches. *Medical Education*, *54*, 105–110.

Vercammen, A., Ji, Y., & Burgman, M. (2019). The collective intelligence of random small crowds: A partial replication of Kosinski et al. (2012). *Judgment and Decision Making*, *14*, 91–98.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology*, *49*, 436–458.

Wai, J., Brown, M. I., & Chabris, C. F. (2018). Using standardized test scores to include general cognitive ability in education research and policy. *Journal of Intelligence*, *6*, 37; doi: 10.3390/jintelligence6030037.

Weekley, J. A., Labrador, J. R., Campion, M. A., & Frye, K. (2019). Job analysis ratings and criterion-related validity: Are they related and can validity be used as a measure of accuracy? *Journal of Occupational and Organizational Psychology, 92*, 764–786.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application (*pp. 157–182*).* Mahwah, NJ: Erlbaum.

Weiss, D. J., & Shanteau, J. (2012). Decloaking the privileged expert. *Journal of Management & Organization*, *18*, 300–310.

Weng, Q. D., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior*, *104*, 199–209.

Whetzel, D. L., Sullivan, T. S., & McCloy, R. A. (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, *6*(1). https://doi.org/10.25035/pad.2020.01.001

Wonderlic (2002). *Wonderlic personnel test manual*. Libertyville, IL: Author.

Wonderlic (2003). *Wonderlic personnel quick-test (WPT-Q) user's guide*. Libertyville, IL: Author.

Zhang, D. C., & Wang, Y. (2021). An empirical approach to identifying subject matter experts for the development of situational judgment tests. *Journal of Personnel Psychology*. Advance online publication. https://doi.org/10.1027/1866-5888/a000279

Table 1.
Classification Consistency of Within-Item Rankings between SMEs and MTurk Raters

		MTurk Ranks					
Communication SJT		**1**	**2**	**3**	**4**	**5**	**CC**
SME Ranks	Most Effective = 1	**9**					**100%**
	2		**9**				**100%**
	3			**6**	3		**67%**
	4			3	**5**	1	**56%**
	Least Effective = 5				1	**3**	**75%**
Medical SJT		**1**	**2**	**3**	**4**	**5**	**CC**
SME Ranks	Most Effective = 1	**17**	4				**74%**
	2	3	**7**	8	1		**28%**
	3		8	**8**	4		**46%**
	4		2	3	**12**	3	**62%**
	Least Effective = 5				3	**17**	**88%**
Military SJT		**1**	**2**	**3**	**4**	**5**	**CC**
SME Ranks	Most Effective = 1	**26**	4				**87%**
	2	4	**22**	3	1		**73%**
	3		4	**23**	3		**77%**
	4			4	**26**		**87%**
	Least Effective = 5					**30**	**100%**

Note. CC = classification consistency; SME = subject matter expert; Overall agreement in Sample 1 = 80%, Sample 2 = 61%, Sample 3 = 85%

Table 2.
Mean Expert Similarity Scores by Sample

	MTurk Raters			SME Raters			
	M	SD	*n*	M	SD	*n*	Cohen's *d*
S1: Communication SJT	0.69	0.24	106	0.45	0.04	9	–1.39*
S2: Medical SJT	0.84	0.20	31	0.62	0.05	28	–1.46*
S3: Military SJT	0.90	0.19	631	0.83	0.20	66	–0.36*

Note. Lower expert similarity scores indicate greater agreement (AD_M) with the average SME rating; ANOVA results for Communication SJT: $F(1,113) = 9.22$, $p = .003$. ; ANOVA results for Medical SJT: $F(1,56) = 30.89$, $p < .001$; ANOVA results for Military SJT: $F(1,694) = 9.18$, $p = 0.003$.
* $p < .05$

Table 3.
Within-Group Mean Rater Agreement among SME and MTurk Raters

	MTurk Raters		SME raters		
	M	SD	M	SD	Cohen's *d*
S1: Communication SJT	0.67	0.20	0.45	0.26	–0.95*
S2: Medical SJT	0.78	0.21	0.62	0.28	–0.69*
S3: Military SJT	0.78	0.18	0.81	0.24	0.14

Note. Average mean deviation for each group is calculated by determining the average deviation across all response options across all SJT items; S1 $n = 40$ total response options; S2 $n = 100$ total response options; S3 $n = 150$ total response options; Lower accuracy scores indicate greater within-group agreement; ANOVA results for Communication SJT: $F(1,77) = 19.24$, $p < .001$; ANOVA results for Medical SJT: $F(1,198) = 21.02$, $p < .001$; ANOVA results for Military SJT: $F(1,298) = 1.53$, $p = .22$.
* $p < .05$

Table 4.
Correlations between Expert Similarity Scores and Individual Difference Variables

	1	*2*	*3*	*4*	*5*	*6*	*7*
Sample 1 (*n* = 56–104)							
1. WPT [a]							
2. Honesty-Humility [b]	.10						
3. Emotionality [b]	–.01	.21					
4. Extraversion [b]	–.27	–.35*	–.52*				
5. Agreeableness [b]	–.11	.25*	–.42*	.40*			
6. Conscientiousness [b]	–.07	.11	–.20	.41*	.23*		
7. Openness [c]	.27	–.24*	–.12	.45*	.17	.41*	
8. Expert Similarity	.52*	.26*	.04	.01	.02	.37*	.34*
Sample 2 (n = 31)							
1. WPT							
2. Conscientiousness	.10						
3. Expert Similarity	.40*	.41*					
Sample 3 (*n* = 631)							
1. Educational Attainment							
2. GPA [d]	.10*						
3. Conscientiousness	.00	.20*					
4. Agreeableness	–.01	.09*	.50*				
5. Extraversion	.09*	.03	.29*	.24*			
6. Neuroticism	–.08*	–.04	–.50*	–.46*	–.44*		
7. Openness	.09*	.14*	.22*	.18*	.30*	–.18*	
8. Expert Similarity	–.07	.16*	.27*	.27*	–.20*	–.04	.09*

Note. We reverse-scored expert similarity scores (–1 * AD_M) to make higher values represent greater similarity; WPT = Wonderlic Personnel Test; GPA = grade point average; [a] *n* = 66; [b] *n* = 68; [c] *n* = 56; [d] *n* = 551.
* $p < .05$

Sample Item: Communication SJT

Although you've put in extra hours, you haven't met your sales quota since you started selling to larger companies. You've found that larger clients take more time to make buying decisions. Your supervisor calls you to his office to ask why you're not meeting your sales goals.

- "I'm making progress. I just needed to learn how larger companies make buying decisions."
- "Maybe I just don't understand how to make sales to larger companies."
- "I'm sorry. These larger companies take too long to make buying decisions."
- "I'm already putting in extra hours. What else do you want me to do?"

Sample Item: Medical SJT

You are one of two attending emergency physicians tonight. The workload is light, until you receive word of a patient arriving in critical condition in the ambulance bay. The other attending physician is "next up" on taking charge of the patient, but this patient's condition is more serious than that of the other patients in the ER.

- Treat the next patient to arrive; protocol only requires one attending physician for an emergency and the other patients have been waiting longer.
- Ask the EMTs for their opinion about who to treat next.
- Assist the other physician in treating the patient who just arrived in critical condition.
- Assume control over treatment of the patient who just arrived in critical condition.
- Call for a meeting to discuss treatment priorities as a team.

Sample Item: Military SJT

Early in the host county's conflict you have had both the National Army and Army Services Forces living on the base. During operations they work together well, but in their free time they fight and argue.

- Speak with some of them to find out what issues they are fighting about and try to help resolve them.
- Let them handle it themselves.
- Organize sports activities during free time to allow both groups the opportunity to release some steam.
- Try to increase the amount of physical separation between where the two groups live on the forward operating base (FOB)
- Insist they live on separate sides of the village and build a wall to divide the living quarters.

Figure 1. Sample SJT Items. For each SJT, test takers are instructed to read the scenario and to rate the effectiveness of each of the behavioral response options.

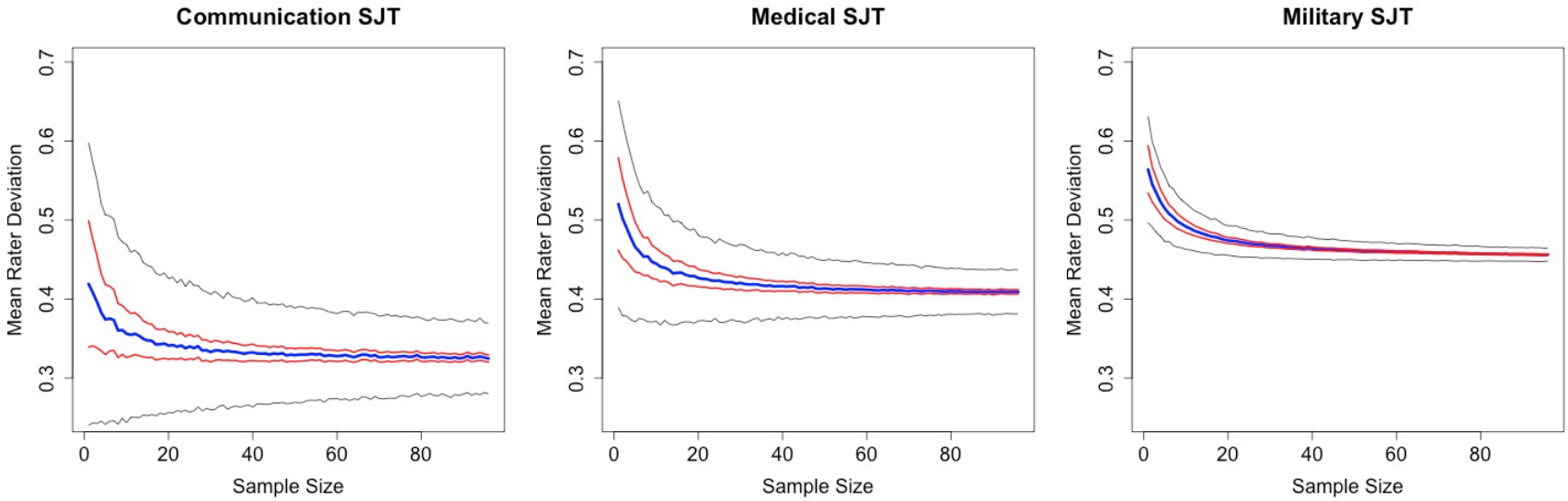

Figure 2. Estimated Average Rater Deviation based on Crowd Size . The blue lines indicate the bootstrapped average rater deviation achieved using 1,000 randomly drawn sets for a given sample size. Mean rater absolute deviation represents the similarity between crowd-based judgments and SME judgments (smaller values indicate greater similarity). Red lines represent the upper and lower 95% confidence intervals for the average rater deviation. Black lines indicate the maximum and minimum rater deviation estimates for each level of sample size.