Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 107–131). Mahwah, NJ: Erlbaum.

Whitenack, D. A., & Harvey, R. J. (2015, April). Within-title heterogeneity in rationally derived target profiles for jobs. In R. J. Harvey (Chair), *Examining alternatives to criterion-related validity studies when setting worker requirements*. Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Why Some Situational Judgment Tests Fail To Predict Job Performance (and Others Succeed)

Deborah L. Whetzel and Matthew C. Reeder
Human Resources Research Organization

Situational judgment tests (SJTs) occasionally fail to predict job performance in criterion-related validation studies, often despite much effort to follow scholarly recipes for their development. This commentary provides some plausible explanations for why this may occur as well as some tips for SJT development. In most cases, we frame the issue from an implicit trait policy (ITP) perspective (Motowidlo, Hooper, & Jackson, 2006a, 2006b) and the measurement of general domain knowledge. In other instances, we believe that the issue does not have a direct tie to the ITP concept, but our experience suggests that the issue is of sufficient importance to include in this response. The first two issues involve challenges gathering validity evidence to support the use of SJTs, and the remaining issues deal more directly with SJT design considerations.

Challenges Gathering Validity Evidence To Support the Use of SJTs

1. If an SJT assesses attributes that are required for critical, though narrow, aspects of job performance, validity estimates may be low or statistically insignificant. This may occur in spite of a well-conducted job analysis showing that a job requires knowledge, skills, and abilities (KSAs) often measured by SJTs. As an example, one may be interested in predicting performance in a job that consists primarily of cognitively loaded activities, such as monitoring displays on a computer screen for anomalies or

Deborah L. Whetzel and Matthew C. Reeder, Human Resources Research Organization, Alexandria, Virginia.

Correspondence concerning this article should be addressed to Deborah L. Whetzel, Human Resources Research Organization, 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314. E-mail: dwhetzel@humrro.org

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press

interpreting data from multiple sources (e.g., air traffic controller, systems operator, or energy dispatcher jobs). In addition to attributes such as vigilance, attention to detail, and skill in interpreting incoming data, a job analysis may reveal that interpersonal skills are also needed for successful job performance under certain circumstances. For instance, perhaps subject matter experts (SMEs) identify that interacting with others is critical in high-stress crisis situations and, consequently, rate such tasks as being high in importance and low in frequency. According to Guion (2011), "some tasks are rarely done but are immensely important when needed. Their importance, as well as their rarity, needs to be known" (p. 43). In order to avoid omitting important tasks or KSAs from consideration when developing predictors, one could use two indicators: one for criticality (operationalized as importance × frequency) and one for importance alone. In our example above, interpersonal skills in crisis situations may receive low criticality yet high importance ratings, which may warrant the development of an SJT that measures interpersonal skills based on importance ratings alone.

The failure of SJTs to predict job performance in such situations can be explained with reference to ITP (Motowidlo et al., 2006a, 2006b), emphasized in the focal article. When conducting a criterion-related validation study, incumbents in technically oriented jobs who perform well in domains related to core task performance observed on a regular basis (e.g., they are attentive to anomalies on computer screens) may or may not have sufficient interpersonal procedural knowledge to know the circumstances under which expressing certain traits (e.g., agreeableness or extroversion) leads to higher levels of job performance. Thus, high performing employees in such jobs may not perform particularly well on a carefully constructed SJT designed to measure traits that underlie interpersonal skills, leading to low validity coefficients.

As a potential remedy, if "interpersonal relations in stressful situations" truly represent a key performance area, the criterion needs to be designed such that it is sensitive to individual differences in this domain. In other words, there should be a performance dimension reflecting interpersonal relations during crisis situations on which supervisor ratings or other data are collected, even though behaviors relevant to this dimension may occur infrequently. Further, because of the "low base-rate" nature of the dimension (i.e., true emergencies do not happen on a daily basis, at least in most jobs), additional care should be taken to ensure that supervisors have had sufficient opportunity to observe incumbents so that they can provide accurate dimension ratings. In either case, the concern may not reflect a problem with the SJT, per se, but rather a deficiency with the criterion measure.

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press

2. As mentioned by Lievens and Motowidlo (2016), SJTs measure procedural knowledge, part of which is general domain knowledge acquired through fundamental socialization processes and personal dispositions. Furthermore, they indicate that this type of knowledge is malleable and can be taught, suggesting that training and development can influence SJT scores. One implication of this is that the predictive validity of an SJT for an outcome such as leadership performance may be diminished if incumbents have been exposed to developmental activities (e.g., leadership training courses, job assignments, action learning) and other forms of on-the-job experience that affect attributes assessed by the SJT. For incumbents, this would be evidenced by higher mean SJT scores, as well as range restriction, and lower validity, relative to those observed in applicant samples.

   The limited available evidence pertaining to this conjecture is mixed. Research on incumbent/applicant comparisons on mean SJT scores (e.g., MacKenzie, Ployhart, Weekley, & Ehlers, 2010; Weekley, Ployhart, & Harold, 2004) suggests that incumbent samples often obtain higher mean scores than applicant samples. Findings pertaining to range restriction and validity are less conclusive, however. For instance, incumbent standard deviations (*SD*s) were actually somewhat larger than applicant *SD*s in five of six organizations examined by MacKenzie and colleagues (2010; see Tables 2–7). Similarly, Weekley and colleagues (2004) found that SJT criterion-related validity estimates were comparable across incumbent and applicant samples. Additional research, particularly on applicant/incumbent differences in validity, is needed to obtain more conclusive evidence.

Challenges Designing and Developing SJTs

3. During the development of SJTs, developers make decisions that may affect predictive validity. For example, an SJT developer must choose between response instruction formats (i.e., knowledge vs. behavioral tendency). Knowledge instructions ask respondents to evaluate the effectiveness of possible responses to a given situation; behavioral tendency instructions ask respondents to identify how they would likely behave in a given situation. Research suggests that the use of knowledge instructions results in higher levels of validity and higher levels of subgroup differences than behavioral tendency instructions (McDaniel, Hartman, Whetzel, & Grubb, 2007). Consequently, if one selects behavioral tendency instructions with the goal of minimizing subgroup differences, there may be a risk of lower validity. However, recent research using one large sample of medical school applicants showed that there was no difference in validity between the two instruction types

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press

in high-stakes settings (Lievens, Sackett, & Buyse, 2009). One could argue that, given applicants' likely propensity to respond in a socially desirable manner in high-stakes settings, both kinds of tests become "knowledge" tests (e.g., McDaniel et al., 2007; Weekley, Ployhart, & Holtz, 2006). Indeed, when an SJT with behavioral tendency instructions is administered, applicants with a stronger ITP toward the trait being assessed (e.g., agreeableness) may be better able to fake the instrument than applicants with less of an ITP toward the trait. Yet high-stakes settings are where concerns regarding the effect of instructions on SJTs are most prominent. Thus, our recommendation is to use knowledge instructions in high-stakes settings.

4. SJT developers' choices among response methods also may affect validity. These choices include having respondents rank order the response options, select the best and/or worst option(s), or rate all responses options on an attribute, such as effectiveness. Research shows that SJTs in which respondents rank order the options or select the best/worst options result in larger subgroup differences than SJTs where respondents are instructed to rate all response options (Arthur et al., 2014). Arthur and colleagues suggested that larger group differences might be attributable to the higher level of cognitive complexity required in the ranking and selecting of best/worst tasks than in the rating task. A tradeoff exists such that if the activities needed to perform a job are cognitively loaded, then having respondents rate all response options may result in lower levels of criterion-related validity than if another response method is used. However, there are advantages to having respondents rate all response options. One such benefit is the ability to obtain more information from a single scenario because response data are collected on all options within a stem as opposed to a subset of the options. In addition, rating all response options may reduce construct contamination if the SJT was not designed to assess attributes in the traditional cognitive domain (e.g., verbal abilities). Finally, some approaches to scoring SJTs using Likert scales result in increased validity and lower mean subgroup differences compared with traditional scoring methods (McDaniel, Psotka, Legree, Yost, & Weekley, 2011).
5. To establish or maintain the credibility of an SJT, developers may include an identifiably "correct" answer among the response options, particularly for domains where the correct option is frequently very transparent (e.g., integrity or ethics). When the instructions ask examinees to rate all of the response options, the effect of this practice may be to increase the effectiveness ratings for the transparently correct option and reduce the effectiveness ratings of the other options due to a contrast effect (i.e., the perceived effectiveness of the "noncorrect"

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press

options is greatly diminished when paired with a "correct" option that is highly transparent). Consequently, the lack of variation among respondents' effectiveness ratings may depress the validity of the SJT if it includes many items where this occurs. Reduced score variability may also occur when response options are included that are relatively improbable given the demands of the situation (e.g., the stem describes an emergency situation and a response option suggests gathering opinions on future action steps). A possible solution is to collect information (e.g., from SMEs) on the transparency and/or feasibility of each option in light of the situation. Options could then be screened on these characteristics, removing options that clearly no one would choose or where the keyed response is high in transparency. Another solution to the problem of transparency may be to provide dilemmas that do not clearly implicate one correct answer and where the demands in the situation do not overly restrict the range of behaviors that one might choose to enact. If response options to a dilemma are equal in social desirability, then the respondent must rely on his/her ITPs to answer the item. For example, if one must choose between two desirable response options, such as balancing a budget and achieving consensus on an issue, someone with an ITP regarding the effectiveness of agreeableness may be more likely to choose the option about achieving consensus on an issue.

6. Developers also must choose among SJT delivery methods. The use of written SJTs, as opposed to video or avatar-based SJTs, could explain the failure of some SJTs to predict job performance. Research shows that, keeping verbal content constant, video-based SJTs tend to have lower correlations with cognitive ability than written SJTs (Chan & Schmitt, 1997; Lievens & Sackett, 2006). This is not surprising given the reading requirements involved in written SJTs. Further, the video-based SJT had higher predictive and incremental validity for predicting interpersonally oriented criteria than the written version. This does not seem surprising, either, given the likelihood that video-based SJTs can include more nuanced and nonverbal cues (especially important for making ratings regarding interpersonal skills) than a written SJT. Using a video-based format, the SJT developer should be able to more clearly convey nuances and subtleties underlying the situation (e.g., the frustration level of a protagonist). Respondents with an ITP for a trait relevant to that measured by an SJT should be able to use the richer information available in a video-based format to better discriminate among response options, which should lead to more accurate measurement of that trait than would be obtained with a test using written scenarios.

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press

In sum, we describe issues to consider when developing a situational judgment test. The issues are framed with referenced to constructs described in the focal article (e.g., ITP and general domain knowledge). The suggestions in this response are not intended to be used as a checklist nor are they intended to be comprehensive. The purpose of this commentary is to extend the focal article to provide practical guidance to SJT developers.

References

Arthur, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of 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.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). New York, NY: Taylor & Francis Group.

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

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*, 1181–1188.

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94*, 1096–1101.

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

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., 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.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*, 749–761.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application* (pp. 57–82). Mahwah, NJ: Erlbaum.

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

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 &

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press

R. R. Ployhart (Eds.), *Situational judgment tests: Theory, management, and application* (pp. 157–182). Mahwah, NJ: Erlbaum.

SJTs as Measures of General Domain Knowledge for Multimedia Formats: Do Actions Speak Louder Than Words?

Bobby Naemi, Michelle Martin-Raugh, and Harrison Kell
Educational Testing Service

Lievens and Motowidlo (2016) present a case for situational judgment tests (SJTs) to be conceptualized as measures of general domain knowledge, which the authors define as knowledge of the effectiveness of general domains such as integrity, conscientiousness, and prosocial behaviors in different jobs. This argument comes from work rooted in the use of SJTs as measures of implicit trait policies (Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006), measured with a format described as a "single response SJT" (Kell, Motowidlo, Martin, Stotts, & Moreno, 2014; Motowidlo, Crook, Kell, & Naemi, 2009). Given evidence that SJTs can be used as measures of general domain knowledge, the focal article concludes with a suggestion that general knowledge can be measured not only by traditional text-based or paper-and-pencil SJTs but also through varying alternate formats, including multimedia SJTs and interactive SJTs.

We extend this point by exploring several ways this conceptualization of SJTs as measures of general domain knowledge might interact with different formats, pointing out issues and concerns across differing format types and presenting areas in need of further research.

Alternate Formats: Video and Virtual SJTs

In both video-based and virtual SJTs, multimedia technology is used to present scenarios and response options in a filmed or virtually animated format (Lievens & Sackett, 2006; McHenry & Schmitt, 1994; Olson-Buchanan & Drasgow, 2006). Recent meta-analytic results demonstrate that multimedia SJTs show stronger criterion-related validity results than written

Bobby Naemi, Educational Testing Service, Washington, DC; Michelle Martin-Raugh, Educational Testing Service, San Francisco, California; Harrison Kell, Educational Testing Service, Princeton, New Jersey.

Correspondence concerning this article should be addressed to Bobby Naemi, Educational Testing Service, 1800 K Street, NW, Suite 900, Washington, DC 20006. E-mail: bnaemi@ets.org

https://doi.org/10.1017/iop.2015.120 Published online by Cambridge University Press