Crook, A. E., & Beier, M. E. (2015). *Using SJTs to understand relations between prosocial knowledge, coping with stress, and OCBs*. Unpublished manuscript.

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.

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

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

McCarthy, J. M., Van Iddekinge, C. H., Lievens, F., Kung, M., Sinar, E. F., & Campion, M. A. (2013). Do candidate reactions relate to job performance or affect criterion-related validity? A multistudy investigation of relations among reactions, selection test scores, and job performance. *Journal of Applied Psychology, 98*, 701–719.

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanen, A. E., & Lerma, M. N. (2015). *A context-independent situational judgment test to measure prosocial implicit trait policy*. Unpublished manuscript.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. R. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, & organizational psychology* (pp. 165–199). London, UK: Sage.

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

Scoring SJTs for Traits and Situational Effectiveness

Robert J. Harvey
Virginia Tech

Lievens and Motowidlo (2016) addressed three of the most important unanswered questions regarding situational judgment tests (SJTs): (a) Should we view them as tests that can assess relatively generic constructs that predict performance across settings, (b) what constructs can they assess, and (c) how should they be scored? They suggested fundamentally changing the SJT development process by targeting the specific constructs we measure, using scoring systems that address both the targeted traits and their situational effectiveness, examining construct validity, and evaluating the criterion-related validity of SJT traits (Lievens & Motowidlo, pp. 11–12).

Robert J. Harvey, Department of Psychology, Virginia Tech.

Correspondence concerning this article should be addressed to Robert J. Harvey, Department of Psychology, Virginia Tech, 121 Williams Hall, Blacksburg, VA 24061. E-mail: harveyrj@vt.edu

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

These recommendations are highly significant on both practical and conceptual grounds.

Practically, they lay the groundwork for developing cost-effective, off-the-shelf SJTs that can be used in a wide range of occupations. SJT development has typically been costly and nonstandardized, being based on the questionable assumptions that (a) a high degree of situational customization of scenarios and responses is mandatory and that (b) scoring must be done only in an item-level, situationally customized fashion based on the effectiveness of each behavioral response in each setting. If targeted-trait SJTs that are composed of more generic content and scored using standardized methods can still predict job effectiveness, such SJTs could be used far more widely and cost effectively than current ones.

Conceptually, targeting traits and using trait-focused scoring represent fundamental psychometric changes. SJT construction methods differ sharply from those used in traditional psychological testing, as SJTs typically do not (a) target the constructs to be assessed, (b) specify the types of behaviors that should be sampled in each domain, or (c) use standardized rules for scoring to estimate scores on each underlying trait. In effect, Lievens and Motowidlo have recommended that SJT developers adopt many traditional test-development methods.

Overall, I concur with their recommendations, although in some respects I advocate moving even further toward traditional test-development methods. Clearly, SJTs are *used* by industrial–organizational psychologists in the same way we use other standardized, self-report, psychological tests: to provide individual-differences data predictive of success in a work environment. The fact that when we *develop* SJTs we routinely neglect to perform what would otherwise be seen as essential test-development steps has long been a curious aspect of SJTs. Simply put, SJTs are psychological tests, and in my view the methods used to develop and evaluate other psychological tests should be applied to SJTs as well (e.g., Murphy & Davidshofer, 2005). I differ with Lievens and Motowidlo on three main issues involving how we (a) target and identify the SJT traits, (b) score those traits, and (c) tailor SJTs to specific situations.

1. Which Traits To Measure?

Little agreement currently exists regarding the identity of the traits measured by SJTs. As Schmitt and Chan (2006) noted, because of the distinctive methodology they use, "the dominant constructs [SJTs assess] are probably conceptually distinct from established constructs such as cognitive ability and personality traits" (p. 153). However, some view SJTs as simply an alternative method for measuring traditional psychological constructs (see McDaniel & Nguyen, 2001), whereas others claim that SJTs are predictive

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

because they measure novel *tacit knowledge* and *practical intelligence* constructs that reflect "the ability to adapt to, shape, and select real-world environments" (Stemler & Sternberg, 2006, p. 109).

Lievens and Motowidlo made a similar argument, saying that SJTs predict because they measure *general domain knowledge.* They noted that "the particular form of domain knowledge that we consider here is *knowledge* about the *utility of expressing* certain traits" (p. 4, emphasis added) at work, that "general domain knowledge is not acquired from specific job experience but reflects effects of *fundamental socialization processes* and *personal dispositions*" (p. 5, emphasis added), and that it is something that "people learn before they ever apply for a particular job" (p. 5).They granted that cognitive and personality constructs may be related to domain knowledge, but only indirectly, as antecedents (p. 4 and Figure 1).

I agree with Lievens and Motowidlo and with Stemler and Sternberg (2006) that SJTs may offer a useful means for assessing individual-differences constructs that lie in a category variously labeled "tacit knowledge," "practical intelligence," or now, "general domain knowledge." However, I am not yet convinced that the types of constructs they cite as examples of domain knowledge (including "ethical leadership," "how to behave effectively when dealing with others," and "prosociality, integrity, and conscientiousness"; Lievens & Motowidlo, pp. 4–5) fundamentally differ from the types of traits long measured in the personality domain.

Although it is interesting to distinguish "domain knowledge" defined as one's stable views regarding the utility of expressing certain traits in work situations versus "personality traits" defined as one's tendencies to prefer various interpersonal behavioral styles, at some point the distinction becomes somewhat arbitrary, at least as a practical matter. Both define relatively stable, dispositional, individual-differences constructs on which people vary, and that may predict effectiveness in a work setting. Indeed, what data would be required to prove that general domain knowledge constructs are different from the traits measured using facet-oriented personality instruments (especially ones using work-focused items)? Simply finding low correlations between domain-knowledge SJT scales and personality facets is not conclusive, given that many personality facets exhibit low redundancy with other facets.

Of greater importance is the question of how we should identify the constructs that SJTs measure and whether these traits are relatively generic and predictive across a range of occupations (which would allow for the development of standardized, off-the-shelf SJTs) or whether we must tailor the selection of the constructs to measure to each situation. Lievens and Motowidlo only briefly addressed this question, noting that the first step is "identifying specific traits that underlie effective performance" (p. 12) for each job,

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

suggesting various sources to consult (e.g., on-site task and critical-incident analysis, competency models).

I suggest pursuing the more general goal of identifying a taxonomy of traits that predict performance across occupations, which would form the basis for developing standardized off-the-shelf SJTs that do not require situational content customization. Lievens and Motowidlo identified "how to behave effectively when dealing with others" and "prosociality, integrity, and conscientiousness" (p. 5) as examples of constructs to target. I agree, as these are the types of generic traits that may predict well across occupations (although the *ways* in which they may predict may vary across situation; see point 3 below).

With respect to the question of which constructs should be targeted, the Stemler-Sternberg SJT methodology (e.g., Stemler & Sternberg, 2006) is relevant; it is based on a seven-category taxonomy that views response strategies as having a *comply, consult, confer, avoid, delegate, legislate,* or *retaliate* motive. Although the Stemler-Sternberg technique has received relatively little research attention, the results of Harvey (2013) supported the notion that its taxonomy may be useful when developing trait-targeted SJTs. Analyses of item ratings of 20 SJT scenarios developed using this method (for each scenario, one response item was written for each strategy) indicated that scales formed from the items in each pool demonstrated reasonably good internal consistency and discriminant validity with respect to *g*-loaded and personality traits. However, some of the scales were strongly intercorrelated, and they did not define seven distinct constructs in exploratory factor analysis (although the factors effectively defined *combinations* of the seven a priori Stemler-Sternberg traits).

In addition to a priori targeting of traits as proposed by Lievens and Motowidlo and by Stemler and Sternberg (2006), I suggest we also employ other scale-development techniques commonly used in traditional psychological test development, such as exploratory and confirmatory factor analysis. Especially in the early stages of identifying a taxonomy of constructs to be assessed using off-the-shelf SJTs, it is perfectly appropriate to rely on exploratory analyses to identify the constructs that underlie the less situationally tailored SJT items that would be applicable across occupations. Subsequent scale-development work can focus on developing specifications of the types of behaviors that should be sampled for each construct and using confirmatory factor analysis to further document their dimensionality.

Although factor analysis has not been widely used when developing SJTs, exploratory factor analysis of SJT items written using the Stemler-Sternberg method and other methods was used in Harvey (2015b) to identify and score SJT scales. Using a sample ($N = 1{,}367$) from diverse occupations, results indicated that SJT factors were easily interpretable as

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

relatively generic individual-differences constructs, with many paralleling the Stemler-Sternberg traits. This was seen in an abstract six-factor solution (with factors of *Avoid, Consult, Arrogant/Selfish, Legislate, Noncritical, Hard Worker/Team Player*) and a more granular 16-factor solution (*Avoid, Legislate, Inaction, Impress the Boss, Selfless, Consult, Hard Worker, Blame Finder, Entitled/Arrogant, Consensus Seeker, Help Coworkers Solve Problems, Consult With Boss, Help Coworkers Develop, Go the Extra Mile, Find Optimal Solutions, Seek the Facts*).

Obviously, all of these SJT traits may not predict performance in all jobs in all settings. However, their generic nature suggests that they could potentially be useful predictors for any job in which interpersonal interactions with clients or coworkers are important. Although the Harvey (2015b) results represent only a first step toward the goal, there is every reason to believe that after a reasonable amount of additional factor analytic research the SJT field can achieve a similar degree of general agreement regarding latent dimensionality as is seen in the traditional personality domain (e.g., Goldberg, 1999; Goldberg et al., 2006).

2. How To Score?

Lievens and Motowidlo recommended scoring SJT items "according to both their effectiveness and *the underlying trait they represent*" (p. 10, emphasis added). I agree that we must score for the traits underlying SJTs; however, I have concerns regarding the assumptions underlying their methods, as well as whether situationally customized item-level scoring is even necessary. Regarding the former, they assume (pp. 13–16) that test developers and subject matter experts (SMEs) can accurately identify (a) the pole of each trait that is associated with good performance, (b) which trait is measured by each response option (and the *level* of the trait it defines), and (c) the level of job effectiveness of the behavior (presumably the latter two will be similar). I find all of these assumptions to be troublesome.

For example, I performed a criterion-related validation study (Harvey, 2015a, 2015b) using factor-analytically scored SJT scales in a group of software-development jobs. Based on critical-incidents analyses, situational moderation of validity was expected (the moderator was the degree of end-user contact, an index of stress). Results indicated that SJT scales predicted differently in lower- versus higher-stress contexts and that the poles associated with high performance were often ones that SMEs would likely *not* rationally identify as desirable (e.g., avoiding problems, finding blame with coworkers, being arrogant, and claiming credit). The difficulty seen in determining which SJT scale would predict, or even which pole would define higher performers, raises fundamental questions regarding the degree to which we should expect SMEs in general to additionally be able to accu-

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

rately match SJT responses to underlying traits and determine the *level* of each trait it reflects.

Accordingly, I recommend a very different approach for SJT scale scoring, based on the same techniques long used for standardized cognitive and personality tests. In the initial stages of instrument development, if explicitly targeted traits are not used (or if there are questions regarding the validity of the taxonomy), exploratory factor analysis can be used to identify the dimensions underlying SJT item pools, and factor score estimates can be used as scale scores. Once the underlying dimensionality is clarified and larger item pools are written, the standard scale scoring methods used for personality inventories can be used (i.e., for each scale, by combining ratings of each response option to form an overall composite score, after recoding reverse-scored items). More complex methods from item response theory (IRT) can be also used (e.g., Roberts, Donoghue, & Laughlin, 2000).

In sum, I reject one of the most basic assumptions underlying traditional SJT development, to the effect that SJT scoring must involve a consideration of the level of job effectiveness of each behavioral response option in the context of performing each job of interest and that we must use a highly labor-intensive (and subjective) process in which SMEs or other experts determine item effectiveness in each setting. I agree fully with Lievens and Motowidlo that we should target the traits measured by SJTs in advance, that highly situationally tailored content is not necessary for scenarios and behavioral responses (pp. 10–12), and that we should compute a scale score on each underlying SJT trait. Why not take the final step (i.e., truly standardized scoring) needed to make the SJT development process fully consistent with the techniques used to develop all other standardized psychological tests?

3. How To Determine Situational Effectiveness?

Obviously, the above discussion begs a critical question: Isn't it still necessary to have *some* degree of situational customization when using SJTs to predict job success? Clearly, it is. However, I recommend that, instead of following a labor-intensive and subjective process that performs the situational tailoring at the *item* level, we customize SJTs at the *scale score* level when determining how to *use* each scale to screen applicants in a given situation (i.e., *after* using standardized scoring methods to compute the scale scores). As Lievens and Motowidlo (pp. 15–16) effectively recommended, and as was illustrated in Harvey (2015a), SJTs can be situationally customized easily and nonsubjectively at the scale level by conducting a criterion-related validation study to determine which scales predict for the job in question and which pole of each scale is associated with success in that setting.

I reached the conclusion that SJTs do not require item-level situational customization, which some SJT traditionalists will no doubt find heretical,

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

for several reasons. First, SJTs are psychological tests that can be used to measure psychological traits that predict job success; unless there is a compelling reason to do so, why should we follow a unique and laborious procedure for developing and using them for applicant screening that deviates from the standard methods used for all other psychological tests? Second, the only way in which it will be possible to develop cost-effective, off-the-shelf SJTs that can be used for a wide range of occupations is by fully standardizing the development process (including scale scoring). Although organizations are free to continue to invest substantial resources to develop SJTs using item-level, situationally customized scoring methods, I find the prospect of being able to dramatically lower the cost of developing and using SJTs to be a very appealing one that could substantially increase the extent to which SJTs are used in applicant screening.

Third, and arguably most important, it has been empirically demonstrated that SJT scales developed and scored using standardized methods predicted performance in a sample of $N = 309$ software-development jobs and offered significant unique predictive power when used in a battery including traditional personality facets and a *g*-loaded critical thinking test (Harvey, 2015a). They also showed good discriminant validity with respect to personality facets and a *g*-loaded reasoning test in an $N = 1{,}367$ occupationally diverse sample (Harvey, 2015b). In that study, R^2s predicting each SJT scale from 17 personality scales and the *g*-loaded scale were quite modest, ranging from 7.3% to 10.4% (mean = 8.8%) variance explained for the six-factor SJT scales and 2.4% to 15.2% (mean = 9.4%) for the 16-factor scales.

Of course, other options for achieving situational customization exist, including using a generalization of the methods used to rationally generate the "true" effectiveness levels in item-level SJT scoring keys (e.g., consensus judgment of SMEs) but rating the effectiveness of the more general behavioral tendencies in the scale scores (e.g., *avoid, consult, legislate, help coworkers solve problems*) instead of rating each individual response option. Although that approach is simple, it suffers from the same limitations seen when using SME judgments to rationally determine the target profiles for personality tests. For example, Whitenack and Harvey (2015) found that a large degree of variability exists in SME ratings of target ranges even for similar jobs, raising questions regarding the subjectivity of the process.

In sum, it is possible for us to follow the same fully standardized development processes used for traditional cognitive and personality tests when constructing SJTs and still retain the ability to situationally customize the way in which SJTs are used in different jobs and contexts. Indeed, given the results noted by Harvey (2015a, 2015b) to the effect that in some situations the pole of an SJT scale defining higher-performing workers is not

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

the one that represents the socially desirable behavioral strategy or that has been claimed to be universally valid in meta-analyses (e.g., several of the poles seen among higher performers were decidedly not the ones that define the high-Agreeableness, high-Conscientiousness end of the scale), it is arguably essential that practitioners rely on situational customization and search for situational moderators of validity. Whether we conceptually view SJT scales as measures of "practical intelligence," "tacit knowledge," "implicit trait policies," "general domain knowledge," or simply additional personality facets that diverge from more traditional personality traits, it is important that future research focus on more fully defining the range of such traits that underlie responses to SJTs and continue to advance the goal of producing standardized, cost-effective, off-the-shelf SJTs that predict job performance across diverse occupations and settings.

References

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

Harvey, R. J. (2013, April). *Cognitive- and personality-based discriminant validity concerns regarding Stemler-Sternberg SJTs*. Paper presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, TX.

Harvey, R. J. (2015a, April). Deriving target personality profiles empirically: Not socially desirable, with situational moderation. 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.

Harvey, R. J. (2015b). *Targeted constructs and standardized scoring of situational judgment tests: Examining Stemler-Sternberg and projective SJT formats*. Unpublished manuscript.

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.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103–113.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications*. New York, NY: Pearson.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A generalized item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3–32.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135–155). Mahwah, NJ: Erlbaum.

https://doi.org/10.1017/iop.2015.119 Published online by Cambridge University Press

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.2015.119 Published online by Cambridge University Press