Industrial and Organizational Psychology, *9(1), pp 3–22 March 2016.*
 doi:10.1017/iop.2015.71

Focal Article

Situational Judgment Tests: From Measures of Situational Judgment to Measures of General Domain Knowledge

Filip Lievens
Ghent University

Stephan J. Motowidlo
Rice University

Situational judgment tests (SJTs) are typically conceptualized as contextualized selection procedures that capture candidate responses to a set of relevant job situations as a basis for prediction. SJTs share their sample-based and contextualized approach with work samples and assessment center exercises, although they differ from these other simulations by presenting the situations in a low-fidelity (e.g., written) format. In addition, SJTs do not require candidates to respond through actual behavior because they capture candidates' situational judgment via a multiple-choice response format. Accordingly, SJTs have also been labeled low-fidelity simulations. This SJT paradigm has been very successful: In the last 2 decades, scientific interest in SJTs has grown, and they have made rapid inroads in practice as attractive, versatile, and valid selection procedures. Contrary to their popularity and the voluminous research on their criterion-related validity, however, there has been little attention to developing a theory of why SJTs work. Similarly, in SJT development, often little emphasis is placed on measuring clear and explicit constructs. Therefore, Landy (2007) referred to SJTs as "psychometric alchemy" (p. 418).

To shed light on these pressing issues, this focal article builds a case for reconceptualizing SJTs as measures of a form of general domain knowledge.

Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Belgium; Stephan J. Motowidlo, Department of Psychology, Rice University.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: filip.lievens@ugent.be

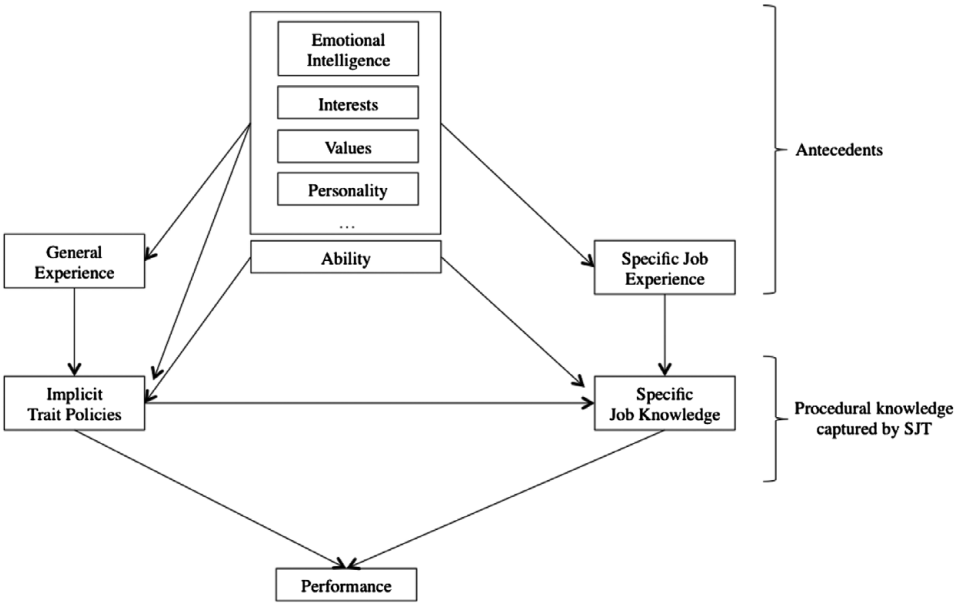

Figure 1. Expanded model of the knowledge determinants and antecedents of situational judgment test (SJT) performance.

The particular form of domain knowledge that we consider here is knowledge about the utility of expressing certain traits. For instance, when agreeable action leads to better job performance than disagreeable action does, people who know this have more general domain knowledge about the utility of agreeableness at work. So, we define general domain knowledge as knowledge about the utility or importance of traits such as these for effectiveness in a job that actually requires expressions of these traits for effective performance.[1] Note that this form of knowledge is not the same as general cognitive ability, although people with more cognitive ability may be better able to learn this knowledge. So, in our reconceptualization, cognitive ability (but also other variables such as personality, see Figure 1) is one of the antecedents of general domain knowledge.

As SJTs are measurement methods that can tap into a variety of job-relevant content domains, we clarify from the outset that a variety of forms of general domain knowledge about trait expression can be relevant for SJTs. As many SJTs pertain to interpersonal relations (Christian, Edwards, & Bradley, 2010), the type of general domain knowledge that is most relevant for answering SJT items involves how to behave effectively when dealing with

[1] General domain knowledge resembles the concept of social desirability in that it captures whether someone knows that an action that expresses a trait such as agreeableness is "desirable" because it is more likely to be effective in a specific work situation.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

others. But other forms of general domain knowledge can also be relevant for SJTs that involve matters related to, for instance, ethical leadership or law enforcement and security. So, given that most SJTs do not deal with strictly technical issues that would involve technical knowledge (only 3%; Christian et al., 2010), we believe that, for most SJTs, general domain knowledge about traits/trait composites such as prosociality, integrity, and conscientiousness that contribute importantly to effectiveness in different jobs will be relevant.

Our key reasons for reconceptualizing SJTs as measures of general domain knowledge about trait expression are based on recent conceptual and empirical developments in the SJT field. Conceptually, recent theorizing argues that SJTs that are developed to measure procedural knowledge tap not only a component of job-specific knowledge that people learn while working at a particular job but also a component of general domain knowledge that people learn before they ever apply for a particular job. In addition, there is recent, compelling empirical evidence that the "situational judgment" part of the term, "situational judgment tests," is inaccurate because many SJT items can be solved without situational information and situational judgment is not really measured by SJTs. Finally, recent research has shown that SJTs can be developed to measure general domain knowledge and that SJTs that are deliberately saturated with such general domain knowledge can predict performance in both real and simulated work settings.

Essentially, this article argues that (a) SJTs predict job performance because they measure procedural knowledge about how to behave effectively in various work situations; (b) one component of that procedural knowledge is general domain knowledge about the utility of expressing various traits at work; (c) this general domain knowledge is not acquired from specific job experience but reflects effects of fundamental socialization processes and personal dispositions; (d) this type of knowledge can predict performance in work situations; and so (e) SJTs should be developed to measure this type of knowledge deliberately and systematically.

As outlined below, this reconceptualization of SJTs into general domain knowledge has important implications for designing and interpreting SJTs differently than has been typically the case. Most important, it calls for developing SJTs to measure clear and explicit constructs that can predict job performance and yield insights about relations between psychological constructs in theoretical networks.

We begin this article with a brief background on the origins and major research streams of SJTs. Next, we explain the conceptual and empirical pieces of evidence that challenge the traditional SJT paradigm in more detail. We continue with presenting design strategies for reconceptualizing SJTs as measures of general domain knowledge. In the final parts, we discuss the advantages and implications of this approach for future research.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

The Traditional SJT Paradigm: Origins and Major Streams of Research

Similar to other selection procedures, SJTs have a rich history that dates back to the 1920s. However, their modern history started when Motowidlo, Dunnette, and Carter (1990) brought them again to the forefront of industrial and organizational (I-O) psychology. The SJT developed at that time was developed to predict interpersonal performance (leadership, assertiveness, flexibility, and sensitivity) and problem-solving performance (organization, thoroughness, drive, and resourcefulness) in response to clients' request for a paper-and-pencil supplement to a structured, behavioral interview. Although the SJT was not designed to measure any particular psychological construct, authors of that report hoped "it might be interesting eventually to discover what constructs are associated with behaviors sampled by the simulation" (p. 641).

Whereas that SJT was developed through a consulting project, the article that reported this work framed the study in broader terms as an effort to ascertain "how much fidelity is necessary before a simulation can become usefully predictive" and "explore the predictive usefulness of low-fidelity simulations" (p. 640). Similar to simulations and sample-based instruments, the notions of point-to-point correspondence with the criterion served as main theoretical underpinning of SJTs (Lievens & De Soete, 2012). So, the article introduced SJTs as potential cost efficient alternatives to assessment center exercises because they presented respondents only with a written situation and required them to pick the best/worst option out of a series of response options instead of asking them to show actual behavior.

Since that study was published, at least three streams of research on SJTs have developed. One stream continued the effort to examine the validity and efficiency of SJTs. Several meta-analyses summarized results of this mostly concurrent validity research and reported corrected estimated population correlations of .26 and .34, with job performance as the criterion (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Importantly, the McDaniel et al. (2007) meta-analysis also revealed that SJT scores explained 1% to 2% of additional variance over both cognitive ability and Big Five trait scores.

Another stream of research aimed to improve on the original SJT format reported in 1990. As a result, SJTs emerged as a versatile instrument with a variety of different make-ups. Overall, it was found that the criterion-related validity of SJT scores increased when SJTs were based on a careful job analysis, on less detailed questions, and on a video-based presentation format (Christian et al., 2010; McDaniel et al., 2001).

A third stream of research followed up on the suggestion offered by Motowidlo et al. (1990) to identify trait constructs associated with SJT scores. As a general conclusion, SJTs emerged as methods for producing scores that

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

could predict job performance reasonably well and that were also correlated with a variety of trait constructs. It is important to note, however, that these studies showed only what constructs were correlated with SJT scores. They did not establish either (a) whether SJTs actually measure those trait constructs (and in fact the correlations are too low to support the argument that SJTs actually measure them) or (b) whether those correlations support the construct-related validity of SJTs because there was no theory about what constructs *should* or *should not* be conceptually related to whatever it is that SJTs measure.

This short literature review shows that over the years an impressive and useful body of research evidence has steadily accumulated. However, at the same time, important questions remained unanswered. At this point, the SJT literature is often about recipes for test development. Researchers have either followed the original recipe published by Motowidlo et al. (1990) or tinkered with it to develop different test formats, hoping to improve criterion-related validity. Essentially, most SJT research was driven by mostly practical considerations, thereby neither questioning nor (re-)examining the theoretical underpinnings of SJTs as sample-based instruments and low-fidelity simulations. This also left the constructs actually measured by SJTs poorly understood. Moreover, the assumption that processes underlying SJTs indeed tapped into situational judgment remained unchallenged.

In the last few years, however, this has changed considerably. Recent theoretical and empirical developments have shaken the fundaments underlying the traditional SJT paradigm and are supplementing research on how to design SJTs for maximum criterion-related validity with new research on how to design SJTs specifically to measure psychological constructs that are theoretically related to patterns of behavior that constitute job performance. After reviewing these recent theoretical and empirical developments, we argue that it might be better to reconceptualize SJTs as measures of a form of general domain knowledge instead of as measures of situational judgment.

Challenges to the Traditional SJT Paradigm

Theoretical Advancements and Evidence

Theory of knowledge determinants underlying SJT performance. Starting with the assumption that SJTs measure procedural knowledge about effective action in the work situations they describe (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; McDaniel & Nguyen, 2001; Motowidlo, Borman, & Schmit, 1997; Motowidlo, Hanson, & Crafts, 1997; Weekley & Jones, 1999), Motowidlo and colleagues developed a theory of knowledge determinants underlying SJT performance (Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006a, 2006b). They drew on the extensive literature on knowledge acquisition (e.g., Beier & Ackerman, 2005; Hambrick, 2003; Van

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

Overschelde & Healy, 2001) that makes a key distinction between prior general domain knowledge and domain-relevant knowledge that people acquire through specific experiences relevant to that domain.

The theory of knowledge determinants underlying SJT performance posits that the procedural knowledge about effective actions in work situations that SJTs measure consists of two components. One component is knowledge about effective and ineffective patterns of behavior in a particular job. This job-specific knowledge can be learned only through exposure to that job or jobs like it. The other component is general knowledge about costs and benefits of expressing various traits in situations like those described in SJT items. This general domain knowledge can be learned outside the work situation and before people ever apply for a particular job. Importantly, general domain knowledge is thus not acquired from specific job experiences. Rather, general domain knowledge reflects effects of fundamental socialization processes (parenting, schooling, etc.) and personal dispositions.

There is empirical support for this theory. Motowidlo and Beier (2010) reported a study showing that when SJT scores were decomposed into a component that was especially saturated with general domain knowledge and another component that was especially saturated with job-specific knowledge, both components predicted job performance about equally well. This is an important detail because it means that SJT measures of general domain knowledge should be able to predict the job performance of applicants even if they have had no prior relevant job experience that would have taught them job-specific knowledge. It also suggests that both general domain knowledge and job-specific knowledge are causal antecedents of job performance. To this point, Lievens and Patterson (2011) showed that procedural knowledge, which we presume includes both job-specific and general domain knowledge about effective and ineffective courses of behavior in job-related situations, was a precursor of effective assessment center performance and job performance (see also Lievens & Sackett, 2012).

General domain knowledge in the form of implicit trait policies. The theory of knowledge determinants underlying SJT performance draws a connection between general domain knowledge and the concept of implicit trait policy (ITP). Motowidlo, Hooper, and Jackson (2006a, 2006b) introduced this concept of ITP to explain how people process information in SJT response options when evaluating the effectiveness of the response options. ITPs represent the degree to which people use information about personality traits that response options express when evaluating the effectiveness of the response options. Thus, ITPs are "policies" in the same sense as widely used in the policy-capturing literature to represent how people weigh information when making evaluative judgments about matters such as job search, compensation, employee discipline, job analysis, employment interviews, and so on

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

(Karren & Barringer, 2002). Instead of matters such as these, however, ITPs concern evaluations in the form of effectiveness judgments about response options in SJT items, and the policies that ITPs attempt to capture are about how people use cues about response options' personality expressions when forming these evaluative judgments.

The connection between ITPs and general domain knowledge is this: If it is true that behavior that expresses some personality trait such as agreeableness, for example, contributes to effective job performance, people who believe this have more general domain knowledge. When presented with a list of actions (in an SJT) that vary in the level of agreeableness they express, people will tend to judge actions that are more agreeable as more effective. In this way, people will demonstrate an implicit policy that weighs agreeableness more heavily when judging actions' effectiveness. Thus, ITPs that represent the strength of the connection between agreeableness and effectiveness also represent how well people know that agreeable action is effective, and knowing that agreeable action is effective is general domain knowledge.

Individual differences in implicit trait policies. ITPs for traits such as agreeableness are presumed to vary across individuals. Motowidlo's (2003) notion of "dispositional fit" makes the point that although ability, experience, and conscientiousness are thought to be implicated in the acquisition of all kinds of knowledge and skill, another mechanism may be implicated in the acquisition of knowledge about effective behavior in interpersonal situations like those represented in SJT items. When the most effective responses to interpersonal situations are responses that express a high level of a trait such as agreeableness, the notion of dispositional fit argues that people who are high on that trait are more likely to "know" that responses that express that trait are more likely to be effective. Having more knowledge in this case means that people weigh agreeableness more heavily when judging whether an SJT response option is effective. This knowledge is akin to ITPs for agreeableness and leads to the prediction that ITPs for agreeableness are correlated with individual differences in agreeableness.

This overall hypothesis that ITPs are correlated with individual differences in traits has also been put to the test. Motowidlo et al. (2006a) supported this hypothesis with SJTs developed specifically to measure ITPs for agreeableness, extraversion, and conscientiousness. Their results showed that self-reported agreeableness and extraversion ratings were significantly correlated with ITPs for agreeableness and extraversion, although self-reported conscientiousness was not significantly correlated with ITP for conscientiousness. Results of several other studies about ITPs for prosocial action (i.e., actions that are performed with the intent to aid or benefit another individual; George, 1992) that used a different SJT format reinforced this conclusion (Crook et al., 2011; Ghosh, Motowidlo, & Nath, 2015; Kell,

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

Motowidlo, Martin, Stotts, & Moreno, 2014; Martin, Kell, & Motowidlo, 2015; Motowidlo, Crook, Kell, & Naemi, 2009; Motowidlo, Martin, & Crook, 2013). These studies used a format that Motowidlo, Crook, Kell, and Naemi (2009) termed a "single-response" SJT in which respondents are asked to rate the effectiveness of only one response option instead of ranking or rating multiple response options. In particular, in these studies, ITPs for prosocial action in occupational contexts such as medicine, law, community service volunteering, and engineering were correlated as expected with traits such as agreeableness, benevolent values, social vocational interests, and emotional intelligence. Importantly, these other constructs were not presumed to be identical to prosocial ITP that the SJT was designed to measure—they were presumed to be *antecedents* of the prosocial ITP construct (see also Figure 1). Moreover, these studies also revealed that ITPs for prosocial action were correlated with prosocial performance in roleplays that simulated situations in which others needed help.

Recent Empirical Research Evidence

How "situational" are SJTs?. Recent research has also scrutinized the key notion as to whether SJTs really tap into situational judgment. So, these studies delve into fundamental assumptions underlying SJTs that were for a long time taken for granted. Krumm et al. (2015) conducted three studies to examine how "situational" judgment in SJT actually is. In their first study, they distinguished between two conditions: In one condition, a traditional SJT (a teamwork SJT) was used, whereas in another condition, the situation description was removed from each of the items of this particular teamwork SJT. So, in that latter condition, the SJT items were "decapitated" because respondents received only the item options. Results showed that the provision of context in the form of inclusion of situational stems had less impact than typically assumed. That is, it did not matter for 71% of the items whether situation descriptions were included in terms of the number of correct solutions per item. In terms of the total score, there was a difference of about 3 points (on 30) between the two conditions.

Given that these results might have been contingent on the specific SJT used (teamwork SJT), Krumm et al. conducted a second study. To examine the generalizability of the results, SJT items were chosen from three broad categories: job knowledge and skills (i.e., 10 SJT items assessing pilot judgment), applied social skills (i.e., 10 teamwork SJT items), and basic personality tendencies (i.e., 10 integrity SJT items). Even in the case of the job-specific aviation SJT, 43% of the items could be solved without respondents receiving the situation description. Results further showed that test takers' expertise level, item length, item difficulty, or response instruction did not moderate the results.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

A third study of Krumm et al. conducted verbal protocols and showed that better performance on the SJT items without situation descriptions was related to one particular test-taker strategy, namely, reliance on general domain knowledge. Although it is important to mention that Krumm et al. focused on the traditional written SJTs (instead of video-based/3-D animated or avatar-based SJTs) and that they did not investigate the impact of deleting situation descriptions on criterion-related and construct-related validity, these results suggest that judgment in typical SJTs might be much less situationally determined than is often assumed and that we might have been "somewhat naïve in assuming that inserting situational cues in assessments automatically would allow them to tap into context-dependent knowledge" (p. 412). In other words, this study provides further support for a key idea in the theory of knowledge determinants underlying SJT performance. That is, even SJTs that were designed with specific situations and jobs in mind capture general domain knowledge.

Is situational judgment measured in SJTs? Another recent study (Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015) also tested some of the deep-rooted assumptions underlying SJTs. Contrary to common wisdom, Rockstuhl et al. argued that traditional SJTs actually do not require situational judgment. Instead, they pointed out that applicants are primarily asked to judge response effectiveness. Indeed, when one looks at typical SJT instructions, respondents are required to indicate what the best thing to do is (knowledge-based response instruction) or what they would do (behavioral tendency response instruction). Strikingly, we do not ask them about their judgment of the situation per se. In the context of an intercultural video-based, open-ended SJT, Rockstuhl et al. showed that situational judgment is only measured when there is an explicit instruction to judge the situation ("What are the thoughts, feelings, and ideas of the people in the situation?"). So, they gave respondents both instructions: Make a judgment of the situation, and make a judgment of the effective response. Interestingly, their study further revealed that the judgments made by test takers on the basis of the situation descriptions (i.e., their construal of the situation) had incremental validity over judgments of response effectiveness for predicting job-related criteria in an international context. The added value of situational judgment that can be linked to perspective taking was especially the case for predicting contextual performance. So, this study showed that situational judgment could capture important predictive information. The problem, however, is that it is not measured in traditional multiple-choice SJTs, which echoes the conclusions of Krumm et al. (2015).

Although more research with different SJTs and different SJT formats is needed, the findings of these two recent empirical studies lead to at least two conclusions. As a first conclusion, the findings question the extent to which

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

existing traditional SJTs are measures of situational judgment. It seems that in many cases, general domain knowledge suffices to a large extent to solve SJT items, which is consistent with one of the main axioms of the theory of knowledge determinants of SJT performance. As such, these recent studies also suggest that an SJT is somewhat of a misnomer. As a second conclusion, they question the extent to which high levels of contextualization are actually needed in SJTs.

SJTs as General Domain Knowledge Measures: Design Considerations

These results of recent empirical studies and the aforementioned theoretical developments suggest that it makes sense to reconceptualize SJTs into measures of general domain knowledge. Practically, the next question is, How can this general recommendation be put into practice? Overall, we propose that SJT developers should proceed in the following sequence: (a) delineate the general domain of knowledge the SJT is intended to measure, (b) develop situational stems and response alternatives that describe actions relevant to that domain, (c) score them according to both their effectiveness and the underlying trait they represent, (d) assess the construct-related validity of the SJT scores by examining their correlations with other constructs that are theoretically related to the domain of knowledge, and (e) assess the criterion-related validity of the SJT scores according to their correlations with patterns of job performance in the form of behavior linked to the domain of knowledge the SJT is designed to measure. Below we provide more details about each of these steps, but we acknowledge at the outset that steps we propose here are broad enough to include various alternative procedures, including some that we have not anticipated.

Delineate the General Domain of Knowledge

As noted, we propose reconceptualizing SJTs as measures of general domain knowledge about the utility (i.e., costs and benefits) of engaging in actions that represent either high or low levels of a targeted trait. Then the first step in developing an SJT should be to identify the domain of knowledge that the SJT is intended to measure. This means identifying specific traits that underlie effective performance. This can be done via an analysis of tasks/roles that make up a job or, alternatively, of critical incidents that describe examples of effective and ineffective performance. In some cases, relevant (compound) traits can be identified by reviewing competency models that are available a priori. In fact, (compound) traits identified that way may actually closely mirror behavioral definitions of the competencies themselves. For instance, a work analysis or a competency model might reveal the importance of knowledge about prosocial action or knowledge about conscientious action and so forth.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

Develop Situational Stems and Response Alternatives

If the SJT is designed to follow a traditional multiple-response format, situational stems can be developed in any of the ways that are currently used, either from analysis of critical incidents, from interviews with subject matter experts (SMEs) about challenging situations, from the imagination of item writers, and so on. The crucial point is that however they are developed, the situational stems should describe opportunities for someone to perform actions that are both effective and high on the targeted trait or both ineffective and low on the targeted trait (see Tett & Burnett, 2003).

A bigger challenge, perhaps, is to develop appropriate response alternatives. In the traditional multiple-response format, our approach requires that response alternatives vary in more than what level of effectiveness they represent. They should be effective *and* represent a high level of the targeted trait or they should be ineffective *and* represent a low level of the trait. In addition, if the SJT format forces choices of most/least likely or best/worst response options, all the response options for a situational stem should represent either a high or a low level of the same trait so respondents are not put in the position of having to choose between a high level of one trait and a high level of another trait.

In the single-response format, SJT items can be developed from analysis of critical incidents. The incidents can be collected through a variety of ways. For instance, they can be collected by asking SMEs for examples of occasions when they saw someone do something that struck them (a) as especially effective or especially ineffective, (b) as especially high on the targeted trait or especially low on the targeted trait, or (c) as especially effective and high on the targeted trait or especially ineffective and low on the targeted trait.

Whether in the multiple-response or single-response format, response options should be checked to assure that they capture both the intended level of effectiveness and the intended level of the trait. This can be done by collecting ratings of the effectiveness of response options from SMEs (who are very familiar with the job) and by collecting ratings of the traits they express from other SMEs (who are sufficiently familiar with the behavioral implications of high and low trait levels to judge response options accordingly). If the traits can be defined without arcane psychological jargon, SMEs selected from the same population that provided the effectiveness ratings (though not the exact same persons who provided these effectiveness ratings) could be used, but psychologists (or even doctoral students in psychology) should also be able to provide credible ratings of traits expressed by the response options.

If each response option is scaled for both effectiveness and trait level in this way, it is possible to correlate the two ratings across the full range of response options in the SJT. We have no way to offer firm guidance about how

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

strong that correlation should be, but as a rough start, perhaps it should be at least as strong as what we would expect for an acceptable reliability estimate. If the observed correlation is judged to be too low, it can be improved by dropping or adjusting response options for which the effectiveness rating is not consistent with the trait rating.

Scoring of Response Options

Several alternative strategies are possible for scoring response options. However, assuming that respondents complete the SJT by making judgments about the effectiveness of response options, all scoring strategies involve calculating the correspondence between the response options' trait effectiveness levels and the respondents' judgments about their effectiveness.

The first issue to settle here is whether to base the scoring key on ratings of the effectiveness of response options or on ratings of the traits they express. If the key involves comparing respondents' effectiveness ratings with response options' trait ratings, an index of the correspondence between those two sets of ratings is a direct measure of the effect of response options' trait levels on someone's judgments about response option effectiveness. This would be a direct measure of ITP for that trait and, therefore, of general domain knowledge about the costs and benefits of expressing that trait in the work situations described in SJT items. We suspect, however, it might be a hard sell to convince practitioners to use trait judgments made by psychologists and/or doctoral students for an SJT scoring key. Alternatively, the key could involve comparing respondents' effectiveness ratings with the mean effectiveness ratings by SMEs who are very familiar with the job. Then an index of the correspondence of these two ratings would be less purely a measure of general domain knowledge, *unless the trait ratings and effectiveness ratings are correlated close to 1.0 across response options*. But even if the correlation is imperfect, such an index of correspondence could be interpreted as a measure of procedural knowledge that is deliberately saturated with general domain knowledge about the utility of a particular trait. The extent to which such a scoring key reflects the underlying knowledge construct depends on how closely SMEs' judgments of response options' effectiveness correspond with the other SMEs' judgments of response option trait levels.

In short, we recommend that SJT response options be developed so that their effectiveness, as judged by SMEs, corresponds as closely as possible with the trait levels they express, as judged by the other group of SMEs. Then when the scoring key compares an applicant's responses in the form of judgments about response option effectiveness with mean SMEs' judgments of response option effectiveness, it will capture reasonably well the construct of general domain knowledge that the SJT was intended to measure.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

Assuming that the scoring key takes the form of some index of correspondence between respondents' effectiveness judgments and mean effectiveness ratings by SMEs, there are also alternative strategies for computing the index. If the SJT asks respondents to rate the effectiveness of all response options, the index of correspondence could be computed as the correlation between respondents' effectiveness ratings and mean SME effectiveness ratings across all response options. Assuming that response options are deliberately developed so that they are either very effective and very high on the trait or very ineffective and very low on the trait, the correlational index is essentially a point-biserial correlation and can be estimated with reasonable fidelity as the difference between the mean of effectiveness ratings for all effective and high-trait response options and the mean of all effectiveness ratings for ineffective and low-trait response options. Of course, this can be done simply by reverse scoring effectiveness ratings for all ineffective, low-trait response options and then summing across all response options.

If the SJT asks respondents to choose most/least likely options or best/worst options, the index of correspondence can be calculated by summing the mean SME effectiveness ratings about response options chosen as most likely or best and subtracting mean SME effectiveness ratings for all response options chosen as least likely or worst.

Although we recommend relying on SMEs' mean effectiveness ratings for a scoring key to be used operationally, we also recommend using trait judgments for a second scoring key to be used to help establish construct-related validity. Trait-based scoring keys can be developed in the same way as we described for effectiveness-based scoring keys. Then when both keys are applied to the same group of applicants, although the one based on SMEs' effectiveness judgments is used for selection decisions, the trait-based index of correspondence can be correlated with the effectiveness-based index to confirm that the final SJT score used for selection decisions measures a form of procedural knowledge that is thoroughly saturated with the intended construct of general domain knowledge.

Construct-related validation. In line with the notion of validation being hypothesis testing, in the next step a theory is formed about what other constructs are related as antecedents to the particular type of general domain knowledge the SJT is presumed to measure. That is, what personality, ability, interest, values, and other variables ought to be antecedents of the knowledge domain targeted by the SJT (see also Figure 1). When these constructs are measured alongside the SJT, this theory can be tested by examining correlations between the SJT scores and the ratings on the constructs presumed to be antecedent to the knowledge domain assessed via the SJT.

Criterion-related validation. Finally, we propose to validate the SJT scores using criteria of relevant dimensions of job performance, making sure

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

that the behavioral elements in the criteria are "linked" to the underlying general knowledge domain. So, we recommend following a similar predictor–criterion matching logic as that which has been done in the personality domain where personality constructs are linked appropriately to corresponding behavioral dimensions of job performance (e.g., Bartram, 2005).

Taken together, the sequence just described exemplifies our conviction that we should put the most emphasis on the type of knowledge the SJT is intended to measure (assuming it is knowledge about the utility of performing actions that express well-defined traits or trait complexes). Starting from general domain knowledge about trait-related behavior, we do not recommend or prefer any particular format. We described two possible approaches (multiple-choice and single-response formats). Hence, we welcome further efforts and experiments to develop and test the format best suited for tapping one type of general domain knowledge or another. In other words, the search for the "best" format characteristics should be informed not just by whatever produces higher criterion-related validity but also by whatever produces a better measure of the construct the SJT is designed to measure. If our arguments about measuring general domain knowledge in the form of ITPs have merit, this means we need to try to develop SJTs with format characteristics that will measure those knowledge constructs as validly as possible.

SJTs as General Domain Knowledge Measures: Conceptual and Practical Benefits

Reconceptualizing SJTs into general domain knowledge measures has several theoretical advantages. First, this reconceptualization highlights that we should focus *first* on what we want to measure and let SJT format details follow accordingly. Further, we put forward theoretical and empirical arguments that we should focus especially on measuring general domain knowledge. Hereby, we need to be explicit about what kind of general domain knowledge we are trying to measure. For SJTs intended to predict effective interpersonal behavior, this knowledge is likely to involve a compound trait such as prosociality. For other kinds of SJTs, for security guards or police officers, for instance, this knowledge is likely to involve some other compound trait like conscientiousness/integrity/reliability/rule compliance. Then the SJT should be designed to determine whether people know that actions that express high levels of the relevant compound trait in a particular job are effective and that actions that express low and polar opposite levels of the compound trait are ineffective. Importantly, such an SJT is amenable to construct validation in a way that SJTs built to tap some vague notion of "situational judgment" are not.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

Second, our reconceptualization contributes to answering the lingering question of what is being measured by SJTs because the theory of knowledge determinants underlying SJT performance firmly grounds constructs measured by SJTs into widely accepted knowledge frameworks that make a distinction between general domain knowledge and job-specific knowledge. Hereby the theory clarifies that personality and cognitive ability are not directly measured by SJTs. Instead, personality and cognitive ability are conceptualized as antecedents of these two types of procedural knowledge as acquired over the years. Given that our theory stipulates what is being measured by SJTs and what are the antecedents, it logically enables better delineating of what is intended variance versus unintended variance in SJT scores.

Third, as our reconceptualization posits that SJT procedural knowledge is malleable and can be taught, it provides the foundation for using SJTs in training and development applications. More broadly, by conceptualizing SJTs in knowledge frameworks, there is an explicit emphasis and link to the notion of knowledge, which is of key strategic importance to organizations, as reflected in the knowledge-based view of the firm (Grant, 1996).

Apart from these conceptual advantages, our reconceptualization also has practical benefits. It should result in more generic SJTs, thereby increasing their applicability across jobs and settings. For instance, in Motowidlo, Ghosh, Mendoza, Buchanen, and Lerma (2015), an SJT was developed to measure general domain knowledge about the utility of prosocial action across work settings that combined items about prosocial and antisocial action in four different professional occupations (law, medicine, community service volunteering, and human factors engineering). Results suggested that this "generic " SJT can predict prosocial behavior in occupational and social settings very different from those reflected in the situational item content. So, although these SJT items came from different occupations, they still reflected the same prosocial construct, which job analyses had identified as being of key importance for these occupations.

When the single-response strategy is adopted, the SJT item development and scoring is further simplified and made more efficient. As another advantage of the single-response format, it enables focusing more tightly on the same underlying trait. This should make it easier to develop items that reliably express specific traits that are the targets of the particular type of general domain knowledge that an SJT is designed to measure.

SJTs as General Domain Knowledge Measures: Future Research Directions

Our reconceptualization of SJTs as general domain knowledge measures opens a window of opportunity for future SJT research. Generally, future studies are needed to test and refine some of our theoretical ideas and

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

address potential caveats. In particular, we propose the following four avenues for future research.

General Domain Knowledge as Measured by SJTs in a Broader Net

First, the model presented by Motowidlo and Beier (2010) can be expanded to include other traits besides personality traits as originally described (see Figure 1). In particular, research with single-response SJTs suggests that a far richer array of traits might be interesting and important to study as potential antecedents of knowledge constructs represented by ITPs. With respect to prosocial ITPs, for instance, these may include emotional intelligence, social vocational interests, and benevolent values. We show ability as another potential antecedent because intelligence is presumably involved in the acquisition of all kinds of knowledge. It may, however, turn out that general domain knowledge in the form of ITPs for traits such as prosociality and conscientiousness is shaped less by intelligence than by underlying personality traits related to the ITP. For instance, Kell et al. (2014) found that an SJT that measured prosocial knowledge in the medical field was significantly correlated with students' self-rated agreeableness ($r = .31$), with their clinical skill in a standardized roleplay ($r = .20$), and with their clinical performance in primary care ($r = .22$) but not with their verbal reasoning ($r = .04$) or GPA ($r = .03$).

Trainability of General Domain Knowledge as Measured by SJTs

Second, future studies should test some of the assumptions underlying the theory of knowledge determinants of SJT performance. As the theory assumes that SJTs measure two types of knowledge, it is intriguing to examine to what extent SJTs tapping general domain knowledge capture learning effects. Indeed, a very basic question is how do people acquire general domain knowledge? What kinds of experiences teach people that actions that express high or low levels of a trait such as prosociality are effective or ineffective? Motowidlo and Beier (2010) speculated that experiences in the form of fundamental socialization processes may be largely responsible for developing this kind of general domain knowledge, perhaps through parental advice or modeling that teaches or promotes the utility of prosocial behavior (e.g., helping others in need, turning the other cheek, looking after one's neighbors) or discourages antisocial actions, such as showing selfish preoccupation with one's own interests, holding a grudge and getting even, and advancing one's own interests at others' expense. These experiences could conceivably occur in many different forms and in many different contexts throughout the course of development into adulthood. But we need biographical work that identifies patterns of socialization experiences that lead people to develop beliefs that expressing such traits is or is not effective.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

General Domain Knowledge SJTs and High-Stakes Settings

We need more studies on the criterion-related validity of general domain knowledge SJTs, especially in actual (high-stakes) selection settings. Some might argue that criterion-related validity might suffer because general domain knowledge SJTs are more generic (less point-to-point correspondence with the criterion) in nature and might be more fakable and coachable than traditional SJTs. The latter argument is based on the fact that the response options of general domain knowledge SJTs deal with the same trait (or complex bundle of traits and skills, a.k.a. competency) because they are explicitly designed to represent effective and ineffective actions related to the same trait (or compound traits).

So far, scores on general domain knowledge SJTs (see, e.g., the single-response SJTs on prosociality) have exhibited similar validities as the meta-analytic validities of scores on full-blown SJTs. Nevertheless, we recommend more research that examines general domain knowledge SJTs in high-stakes settings. In a similar vein, we need studies that extend the findings with SJTs as measures of ITPs about prosocial and conscientiousness actions to other domains. For instance, leadership and integrity represent two domains that are prevalent in the taxonomy of Christian et al. (2010). Future studies should test the feasibility of developing SJTs that assess general domain knowledge in these two areas.

SJTs are often used as screening devices ("select out") in applicant pools that have little or no relevant job experience. SJTs that assess general domain knowledge are especially suited for such purposes. Moreover, as noted above, research showed that such SJTs could be useful and predictive across jobs and settings (Motowidlo et al., 2015). That said, we acknowledge that, for some SJT purposes (e.g., certification, credentialing, selection into advanced level jobs; Lievens & Patterson, 2011), a high degree of context-specificity might be a key requirement. This might also be the case in some industries (public sector) and for technical jobs. So, for these purposes, sectors, and jobs, the development of job-specific SJT items with high levels of contextualization seems to be advisable.

SJT Formats for Assessing General Domain Knowledge

In this focal article, we put the emphasis on SJTs as measures of constructs and especially as measures of general domain knowledge. Our design considerations attest to this construct-driven focus. We see the single-response SJT format only as one possible and easy to develop format. So, the single-response SJT format is not inherently tied to our reconceptualization of SJTs as measures of general domain knowledge. Thus, we welcome studies that explore a variety of formats for operationalizing our broad sequence of designing SJTs as measures of general domain knowledge and ascertaining

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

their effects on key selection outcomes (reliability, validity, subgroup differences, fakability, coachability, costs, etc.).

Applicant perceptions to SJTs depend on the SJT format. In fact, past research revealed that applicants react more favorably to more fancy formats. For instance, multimedia formats were preferred over paper-and-pencil formats (Chan & Schmitt, 1997; Lievens & Sackett, 2006) and interactive formats over noninteractive ones (Kanning, Grewe, Hollenberg, & Hadouch, 2006; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). A key point is that SJTs as measures of general domain knowledge are still job related and can in principle also be put in audiovisual formats (e.g., video-based, 3-D animated, or avatar-based formats; Fetzer & Tuzinski, 2014). For instance, video clips can be made of the different response options that represent effective or ineffective actions related to the same competency (see Podsakoff, Podsakoff, MacKenzie, & Klinger, 2013). So, efforts to increase the realism in SJT stimulus format or ground them in interactionism (e.g., by explicitly asking candidates to make situational judgments, see Campion & Ployhart, 2013; Rockstuhl et al., 2015) can be undertaken as long as they are explicitly designed to measure general domain knowledge.

References

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203.

Beier, M. E., & Ackerman, P. L. (2005). Age, ability, and the role of prior knowledge on the acquisition of new domain knowledge: Promising results in a real-world learning environment. *Psychology and Aging, 20*, 341–355.

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures: Interactionist psychology operationalized. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). New York, NY: Routledge.

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.

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.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417.

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.

Fetzer, M. S., & Tuzinski, K. (2014). *Simulations for personnel selection*. New York, NY: Springer.

George, J. M. (1992). The role of personality in organizational life: Issues and evidence. *Journal of Management, 18*, 185–213.

https://doi.org/10.1017/iop.2015.71 Published online by Cambridge University Press

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

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*, 109–122.

Hambrick, D. Z. (2003). Why are some people more knowledgeable than others? A longitudinal study of knowledge acquisition. *Memory &Cognition, 31*, 902–917.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment, 22*, 168–176.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research and practice. *Organizational Research Methods, 5*, 337–361.

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

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

Landy, F. J. (2007). The validation of personnel decisions in the twenty-first century: Back to the future. In S. M. McPhail (Ed.), *Alternate validation strategies: Developing and leveraging existing validity evidence* (pp. 409–426). San Francisco, CA: Jossey-Bass.

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *Handbook of assessment and selection* (pp. 383–410). New York, NY: Oxford University Press.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96*, 927–940.

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. (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.

Martin, M. P., Kell, H. J., & Motowidlo, S. J. (2015). *Prosocial behavior: Exploring the role of personality, values, emotional intelligence, and prosocial knowledge*. Unpublished manuscript.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (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). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

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.

Motowidlo, S. J. (2003). Job performance. In W. C. Borman, D. R. Ilgen, & 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., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71–83.

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

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., 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.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology*. Palo Alto, CA: Consulting Psychologists Press.

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.

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

Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. (2013). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *Journal of Applied Psychology, 98*, 99–113.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887.

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464–480.

Tett, P. R., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–551.

Van Overschelde, J. P., & Healy, A. F. (2001). Learning of nondomain facts in high- and low-knowledge domains. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1160–1171.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679–700.