**On Designing Construct Driven Situational Judgment Tests:**

**Some Preliminary Recommendations**

Nigel Guenole 1, 2

Oleksandr S. Chernyshenko 3

Jeff Weekley 2

Please send correspondence concerning this manuscript to Nigel Guenole via email at n.guenole@gold.ac.uk

**Abstract**

Situational judgment tests are widely agreed to be a measurement technique. It is also widely agreed that SJTs are a questionable methodological choice for measurement of psychological constructs, such as behavioral competencies, due to a lack of evidence supporting appropriate factor structures and high internal consistencies. Nevertheless, both the peer review literature and applied instances of SJTs reveal no shortage of attempts to measure psychological constructs with SJTs. While acknowledging the lack of evidence supporting the ability of SJTs to assess psychological constructs, we see that a number of practical benefits would accrue to organizations if industrial psychologists were able to design SJTs that measure psychological constructs. Accordingly, this article presents a discussion of steps that psychologists can take to enhance the chances of observing validity evidence for construct based SJTs.

**On Designing Construct Driven Situational Judgment Tests:**

**Some Preliminary Recommendations**

Situational judgment testing is an approach to psychological measurement that was first used in personnel selection settings in the middle part of the 20th century (Chan & Schmitt, 2005). However, until the 1990s, much greater researcher and practitioner attention was given to more traditional selection methodologies such as interviews, self-report personality questionnaires, and multiple-choice ability tests. In an influential paper, Motowidlo, Dunette and Carter (1990) highlighted the utility of SJTs as an assessment methodology for predicting performance, sparking renewed interest in this assessment technique. Since this time, considerable effort has focused on establishing the validity of SJTs, examining what they measure, and experimenting with different SJT formats.

The essence of most of descriptions of SJTs was reflected in the recent definition from Lievens and Motowidlo (2015). They noted that SJTs capture a test taker’s *contextualized* responses to *samples* of job situations. Responses to these situational samples are then used to make predictions about a candidate’s future performance. In contrast to other contextualized assessment methods such as assessment centers, regardless of the fidelity (i.e., realism) with which the stimuli are presented, candidate responses are captured using low-fidelity methods. Responses do not involve enactment of the behavioral response that would be expected on the job. Instead, responses involve selection from amongst a number of closed-ended responses regarding knowledge of the appropriateness of intentions to behave in different ways. These responses can then be subjected to the same psychometric investigations regarding reliability and validity that psychometricians undertake for other selection methods such as cognitive ability and personality questionnaires.

Researchers have now significantly progressed the scientific understanding of SJTs (e.g. (Krumm, Lievens, Hüffmeier, Lipnevich, Bendels, & Hertel, 2015, Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015). In turn this has increased SJT usage in practice. For instance, it is now common for SJTs to be used to aid selection decision-making in medical college admissions (Lievens, Buyse, & Sackett, 2005) and for identifying development areas for business leaders (Guenole, Chernyshenko, Stark, & Drasgow, 2014). For up-to-date summaries of research and applied practice related to SJTs, readers may consult Campion, Ployhart, and MacKenzie (2014), Lievens, Peeters, and Schollaert (2008) and Weekley, Hawkes, Guenole, and Ployhart (2015).

**Interpretability of SJT Scores**

One important disadvantage of the SJT methodology is that SJT scores often have poor internal psychometric properties. To understand how this process plays out, it is important to consider the task candidates are asked to perform when answering an SJT. First, candidates usually read or watch a challenging managerial scenario. They are then usually asked to rate the effectiveness of a series of possible responses, or to rank or partially rank the effectiveness of the different response options (e.g. rank all, pick the best, pick best and pick worst). These responses can then be assigned weights based on subject matter expert judgments or empirical keying against some performance criterion, and responses can be aggregated across items using classical psychometrics or model based approaches such as item response theory and confirmatory factor analysis. The candidate responses can also be scored according to the degree of association (e.g. Pearson’s correlation) between the ratings (or rankings) and the subject matter expert judgments of effectiveness.

A troublesome psychometric issue with SJT scores arrived at in this way is that they produce scores with low internal consistency reliability. In the most recent summary of the coefficient alpha reliability of SJTs, Catano, Brochu, and Lamerson (2012) presented a meta-analysis of alpha coefficients from 39 studies involving 56 coefficients based on 45,062 participants. This meta-analysis showed that the internal consistency reliabilities were on the whole quite poor, the meta-analytic alpha = .46 with a 95% confidence interval of +/- .05. Chan and Schmitt (2005) stated in their review that SJT unreliability is often at the level of individual responses to items, meaning the content of responses, not just the stems, is typically highly heterogeneous.

While this heterogeneity may be beneficial for predicting work performance behaviors that are also often heterogeneous, it poses challenges for simple interpretation of SJT scores. The way item level heterogeneity manifests itself in internal psychometric analyses of SJT scores is in either or both of the following characteristics: a) weak general factors where the variance explained by the first factor is low, and/or b) complex factor loading patterns where items have substantial loadings on multiple dimensions, rather than high loadings on one dimension and near zero loadings on others.

Many test developers and users prefer the reverse of these situations, i.e., a situation of strong factors where items load highly on one factor and low on others. This latter loading pattern is one of the criteria for meeting Thurston’s (1947) simple structure. It greatly simplifies interpretation of test scores because items can be clearly interpreted as measuring one factor and not another. The net result of heterogeneity is that SJTs have difficulty measuring *homogenous constructs,* which limits the method’s use for theory testing purposes where researchers need clarity about what is being measured. The heterogeneous nature of SJT scores is also problematic where provision of feedback is the end objective, because practitioners usually want scores to relate to underlying abilities and allow specific coaching.

**A research literature driven by practical considerations**

One explanation for the low reliabilities and relative absence of general factors might be that researchers are focusing on formulaic development of SJTs rather than on research investigating what SJTs measure in a scientific manner. For instance, Lievens and Motowidlo (2015) recently lamented the fact that despite considerable progress in the practice of developing SJTs, much of this literature resembled recipes for developing SJTs rather than scientific advances in our understanding of what SJTs are and how they work. These authors argued that recent developments in the science of SJTs question the fundamental assumptions that have driven much of the research into SJTs to this point, and that these new developments ought to inform future research and practice. In particular, they suggested that SJTs should be designed to measure generalized domain knowledge that could then be examined in the context of nomological networks for construct validation.

Reasons for their proposition are broadly as follows. SJTs measure procedural knowledge which can be decomposed into job specific procedural knowledge learned at work, and generalized domain knowledge learned through socialization. They recommend that the latter be the preferred target for measurement using SJTs because it predicts as well as job specific knowledge but does not require having performed the job. More specifically, these authors argued that the procedural knowledge represented by scores on SJTs reflects a candidate’s beliefs about the effectiveness of the different levels of traits that each response options convey, or in other words, a candidate’s implicit trait policies. Individual differences in these trait policies are explained by the idea of dispositional fit. That’s to say, an individual high on agreeableness may have greater knowledge about the efficacy of an agreeableness response from first hand experience, and may therefore rate a response reflecting agreeableness more highly than a candidate with lower levels of agreeableness. This led Lievens and Motowidlo to recommend that, in addition to measuring targeted levels of effectiveness, response options should as closely as possible also mirror levels of a targeted construct or trait. Behind the often disappointing factor analytic results for SJTs, then, might be because the proposed dimensionality often imposed for testing reflects construct models that do not correspond to the underlying generalized domain knowledge, or knowledge about the traits that underpin response options. To resolve this, Lievens and Motowidlo proposed response scales to SJT stems that represent both a) effectiveness and desirability, and b) the level of the trait, and that the closer that these two ratings correspond to one another the better.

Lievens and Motowidlo (2015) also questioned the need to use the traditional scenario and multiple response options, since research has shown that withholding the scenario and only presenting the responses still permits solution of the scenarios, and asking for a rating of a single scenario can can lead to useful associations with performance. In broad terms, we agree with Lievens and Motowidlo. In this article, we build on their suggestions by offering a set of recommendations for designing construct based SJTs based on recent past attempts at construct driven SJT development, and personal experiences designing construct driven SJTs in industry. Importantly, we note that some of this advice is general psychometric advice, but it is advice that we frequently find is not discussed in the context of SJT development.

**Paper structure**

The principal objective of this article is to provide some preliminary guidance on how SJT designers might design *construct driven* SJTs and thus increase interpretability of SJT scores. We first review development steps reported in the methodology sections of recent peer reviewed papers where the authors sought to measure constructs using the SJT method. The goal was to see whether the developers report any special practices that should be incorporated into recommendations for the design of construct driven SJTs. We then integrate findings from this literature review, along with our own experiences of designing construct based SJTs, into a discussion of the steps preliminarily recommended to develop construct-oriented SJTs. At each step, we also discuss an empirical example of a construct driven SJT designed to measure a multi-dimensional model of leadership capability.

**Recent attempts to design construct driven SJTs**

In this section, we examine several recent articles that purport to measure unidimensional constructs with the SJT method (Becker, 2005, Bledlow & Frese, 2009, Motowidlo, Hooper, & Jackson, 2006, Mumford et al., 2008, Peus, Braun, & Frey (2013), Sharma, Gangopadhyay, Austin & Mandal (2013), Guenole, Chernyshenko, Stark and Drasgow (2014), and Westring, Oswald, Schmitt, Drzakowski, Imus, Kim & Shivpuri, S. (2009). The aim here was identify any design practices that researchers used that might enhance the success with which constructs could be measured by the SJT method.

-------------------------------------------

Insert table 1 about here

--------------------------------------------

The results of this qualitative survey of construct driven SJT designs are presented below in Table 1 and revealed the following themes. First, rather than being inferred in an exploratory fashion, the constructs being measured were unanimously pre-determined. Indeed, part of the reason for poor results until this point for factor analysis of SJT responses could be because SJT designers were using ad hoc realistic work scenarios rather than scenarios designed to measure particular constructs; the latter by their nature are more homogenous. Trying to discover what is measured after the scenarios are captured or by analyzing response data post hoc may limit the chances of measuring homogeneous constructs. (We turn shortly to ensuring the SJT options measure their intended construct, an important issue given that it is the response option that is scored, but one that has received inadequate attention).

Interestingly, the constructs covered in our survey were almost always specified to have multiple dimensions, except for the SJT by Becker (2005) that treated integrity as a unidimensional construct. Because existing literature suggests that SJTs are generally heterogeneous, one approach to mitigate unintended multidimensionality appears to be breaking heterogeneous items into multiple unidimensional constructs and then focusing efforts on measuring one construct at a time. Items can be combined to measure a specific construct, irrespective of the stem with which it associated (i.e., items from a specific stem could contribute to the scores of multiple constructs).

The number of scenarios reported varied widely, from as many as 18 per dimension (Sharma et al., 2013) to as few as three per dimension (Guenole et al., 2014). On the basis of this review it is difficult to make recommendations regarding test length other than to say that, all things being equal, longer is better. As all things are never equal, test length should be a function of desired reliability, construct breadth (e.g., broader constructs will likely need to be assessed by more items than narrow constructs), and intended test time (operationally, seat time constraints may make reliable construct measurement impossible even if the predictive validity of the overall SJT is acceptable).

Study authors generated the content for the scenarios and responses in most instances in the studies we surveyed. The authors did this based on theory rather than relying on surveys or focus groups. We expect this is because study authors felt they could more accurately target the content required based on knowledge of the literature than by relying on job incumbents. When focus groups or surveys are used to generate scenarios, it is important that the focus groups explicitly focused on construct definitions to ensure the relevance of the content that is generated to the intended construct. There was no apparent advantage to either approach, provided that attention is given to ensuring that the items are as unidimensional as possible and clearly related to the intended construct.

The number of response options authors reported using ranged from 3 - 4 (Sharma et al., 2013) to 8 (Peus et al., 2013). The number of response options used in construct driven SJTs is important because it relates to a second feature we surveyed, whether levels of the construct are used as options or if the options represent distinct constructs. From a psychometric standpoint, it is desirable that the response options represent degree of a particular construct rather than qualitatively distinct classes, because we wish to describe candidates as having lower or higher standing on the construct being measured.

However, having response options represent different levels of a single construct does introduce new issues. For example, we expect that it would be very difficult to develop eight options representing distinct levels of a psychological construct. It will also be difficult to make these levels appear approximately equally desirable to candidates, thus introducing faking as a concern in high stakes settings. Fewer rather than more response options seem preferable if the options represent different levels of a single construct (3 to 4 options seem about the most that is feasible). Finally, while there was quite a lot of discussion of SME sorting of scenarios, there was less discussion of SME sorting of response options. We believe more focus on ensuring the response options measure the intended constructs is necessary.

The length of scenarios and options also varied widely in the papers we reviewed. The length of situation stems ranged from 17 words up to 141 words, while the length of response options across studies ranged from 17 to 35 words. The only consistent trend was that the stems are always longer than the response options. From a rational point of view, the longer the item stem or response, the greater the chance for unintended multidimensionality.

In terms of scoring approaches, both subject matter expert (SME) and empirical criterion keying were reported, but lead to no clear consensus as to which approach ought to be adopted. An interesting observation is that none of these studies reported the use of hybrid keys, i.e., SME or theoretical keys supplemented by empirical keying. We expect that hybrid keys could be particularly useful given that our psychological theories and knowledge of SJT design principles to date have not proven sufficiently advanced to allow development of SJTs with strong psychometric properties (i.e., high internal consistencies, clear factor structures). We suspect that, in many instances, the correct key will need to be ‘discovered’ empirically by tweaking the initial scoring key.

Most studies reported conducting item analysis and subsequent item improvements or elimination. As is the case for the development of other psychometric instruments, writing more scenarios and items than are ultimately required will permit judicious scenario/item selection for the best possible psychometric properties. The first of these critical psychometric properties that the item selection is intended to improve is reliability. In our survey of articles, there was little consistency in the reliability estimates reported or the item analysis methods used to maximize reliability. If measuring a construct were a goal of the SJT development process then within-construct estimates of internal consistency reliability would appear necessary.

When non-cognitive constructs are assessed with many options and every option’s effectiveness is rated, our experience suggests that alpha estimates can compare favorably to alpha estimates in other selection approaches, e.g. personality testing. Within this context, making sure that each response option represents a single construct, and accumulating construct-specific response options across stems potentially presents the greatest chance of observing construct relevant variance in test scores. On the other hand, when only a small number of scenarios are used and ranking or partial ranking (pick the best or worst) instructions are used, our experience is that the responses tend to be more heterogeneous and test retest reliability is a better option.

Finally, the second psychometric property that the item selection in the preceding paragraph is intended to improve is validity. There was also considerable variation in approaches to external validation. For example, numerous studies reported correlations with self-report Likert style scales. However, overall, there was inadequate rationale offered for the choice of external scales. A notable exception was the study by Motowidlo, Hooper, & Jackson (2006), which examined the relationship between SJT scores and corresponding personality scale scores. In addition, the significance of the correlation was the primary criterion for validity. However, it was not clear from the hypotheses what the expectations should have been for these correlations, aside from being positive and significant. Other studies used a multi-trait-multi-method that permitted a stronger test of SJT validity because it makes hypotheses about the relative sizes of expected correlations.

**Suggestions for the development construct driven SJTs**

In this section we offer a 5-step procedure for developing construct driven SJTs. Our approach is based on known SJT design steps originally outlined by McDaniel and Nguyen (2001) and Weekley and Ployhart (2006), but we make specific comments about considerations that need attention if the SJT is designed to measure underlying constructs rather than just to predict important outcomes. We also provide an empirical example of a construct driven SJT developed by Guenole et al (2014) to measure 12 leadership competencies.

**Step 1. Develop situational item stems**

In considering the reason for the poor record of SJTs at measuring psychological constructs, Chan and Schmitt (2005) wrote ‘We believe that the multidimensionality of SJT situations and situation responses may be traced back to the multidimensionality of the criterion construct that the SJT is designed to predict’. In other words, sampling a multidimensional criterion space for scenarios and responses and then putting the behavioral scenarios into a single SJT leads to a multidimensional SJT at the item level”.

Building on this observation, we suggest that when building construct driven SJTs, scenario writers should be asked to write scenarios that represent a specific dimension being measured. This differs from an approach that might ask item generators to focus on construction of realistic work scenarios. It also differs from approaches that try to discover dimensionality post hoc through content sorting of scenarios by subject matter experts. The latter approaches, which are empirically rather than theoretically guided, are the most common in practice (Weekley et al., 2006; Peus et al., 2013) but so far have not been effective in designing construct driven SJTs.

In applied settings SJT development routinely involves custom construct models that have not been validated using techniques such as multi-trait-multi-method (MTMM) analyses using any measurement methodology, let alone SJTs. This can present a significant challenge to the test designers who are essentially being asked to design homogenous measures of poorly defined constructs. Therefore, wherever possible, steps should be taken to validate the underlying model with the alternate methodology (e.g., Likert ratings, assessment centers) to ensure that sufficient evidence exists for the constructs to be measured with the SJT method. Good SJT stems should provide enough information that a respondent can decide on a possible course of action, but not too much detail that would lead a candidate to favor an obviously correct response.

The information should be general enough so that respondents from different backgrounds can interpret the information in similar fashion. To produce high quality item stems, it is usually recommended to conduct one-on-one behavioral interviews with actual job incumbents and/or conduct “critical incidents” workshops where participants are asked to recall actual work-related situations relevant to the specific constructs to be measured. Note that participants should ideally be from the same population for which the SJT assessments will be conducted, so SJTs developed from interviews with first level managers might not be appropriate for senior executives.

Our experience with designing construct driven SJTs is that one may need to have several rounds of scenario generation. We have found that starting out with focus groups generates a large number of construct driven scenarios, but the content coverage (e.g., number of industries or variety of situations) is often narrow. After the initial analysis and elimination of redundant items occurs, it is often necessary to go back to job incumbents to generate new items in a more focused manner. For instance, this might take the form of one-to-one interviews where information required for specific content areas that can be directly probed. Results from behavioral interviews and/or workshops are then edited to a common format where 1) the problem/situation is described in full detail, 2) the actual action/response made by the interviewee is recorded, and 3) the result of that action is detailed.

**Step 2. Conduct item stem analysis and classify stems into dimensions.**

Once the set of initial item stems has been edited to a common format they should be sorted for any redundancy, and re-classified into their intended dimensions by an independent group of subject matter experts. Our experience is that it is desirable that the initial scenario pool contains 2-3 times more stems than needed for the final SJT form. Having more scenarios than will ultimately be retained is more important in the development of construct driven SJTs, where a scenario’s relevance to a specific construct must be demonstrated before it is retained. A subject matter expert (SME) rating exercise is typically conducted to achieve this goal. Ideally, a group of expert judges who did not participate in the scenario development should be selected. While there is no firm rule on the number of judges that should be recruited for this task, but we have found from 5 to 7 works well and that an odd number is preferable to avoid ties. The goal of the rating exercise is to allow the best scenarios to be retained and the poor scenarios removed or edited for improvement.

In addition to classifying each stem, SMEs may be asked to rate the difficulty, realism, and quality of the scenarios. Item statistics for each scenario can then be generated and used to decide whether to retain or exclude particular scenarios into the final test form (this is particularly useful when parallel forms are being designed). We have found using a combination of means, standard deviations, and inter-rater reliabilities of the SME judgments to be helpful in deciding about the appropriateness of item stems. If an item stem cannot be classified into a specific dimensional category due to rater disagreement, it is usually modified or deleted from further consideration. In many ways, this process is similar to pre-testing of personality or cognitive ability items where a smaller group of respondents is used to compute initial classical test theory or common factor analysis statistics to identify problematic items.

**Step 3. Develop Response Alternatives.**

If SJTs are intended to measure judgment in the face of typical work situations, it is common that responses represent classes of broadly distinct strategies for handling each scenario. Job incumbents usually generate these responses. They are typically asked to write a short description of how they would respond to a specific situational stem. Responses can also come from behavioral interviews or “critical incident” workshops that have recorded the actual actions taken in response to the situation. SMEs may also be used to generate response alternatives with a designated degree of effectiveness. Finally, the test authors may contribute items based on their knowledge of the constructs of interest, the job, and their personal experiences. Generation of response options is a critical stage in developing construct driven SJTs. In standard SJTs, it has been shown that multidimensionality typically exists at the level of response options. This is likely the result of the development strategy reported in many SJTs where response options were generated to have maximally diverse types of behaviors.

In contrast, with construct driven SJTs, response options should be generated to represent different levels of the same behavior. In fact, this is the key requirement for SJTs to be unidimensional, which we suggest is preferable for construct based SJT design. To facilitate generation of response options that progress from indicators of low levels of a construct to high levels of a construct, we recommend that SJT designers consider the response options to be behaviorally anchored rating scales that progress linearly from low to high levels of a construct. A set of response options assessing the Information Search dimension from the High Performance Behavior model at a low level, for instance, might involve not bothering to search for information. A moderate level response might describe searching internally around the organisation for information relevant to resolving the scenario. Finally, a high level response may involve searching inside and outside the organization.

**Step 4. Decide response instructions and develop a Scoring Key**

McDaniel and Whetzel (2007) noted that while many types of response instructions can be used with SJTs, nearly all of them fall under “Behavioral Tendency” or “Knowledge” categories. In SJTs with behavioral tendency instructions, respondents are asked to indicate what they “would most likely and/or least likely do”. The assessment in this context attempts to tap into respondents’ typical behaviors or styles. In contrast, knowledge instructions ask respondents to indicate “the most/least effective response” and, thus, assesses whether respondents have enough expertise to recognize the correct path of action. The assumption in both response instructions cases, however, is that higher test scores will generalize to the actual future behavior. For the assessment of leadership capability, we believed that knowledge instructions are more appropriate.

By definition, capable leaders are those who are able to vary their response styles according to the situation at hand. To do that, a leader must know which of the available actions has the best probability of success. Whether the leader actually is able to adequately perform the correct action would depend on his/her experience, training, and personality, many of which are assessed in conjunction with SJTs. On the other hand, we believe that behavioral instructions tend to overemphasize a leader’s “typical behavior” and, as McDaniel et al. (2003) have shown, this makes SJT scores correlate more highly with personality. For example, the meta-analytic correlation between SJTs scores with behavioral instructions and the Emotional Stability personality dimension was .51, suggesting considerable overlap in the constructs assessed.

Bergman, Drasgow, Donovan, Henning, and Juraska (2006) have reviewed a number of scoring strategies that can be used to key SJTs. These include empirical keying, modal keying, rational keying, theoretical keying, SME keying, “novice vs. experts” keying, and various hybrid keys. The most common scoring keys for SJTs are developed using SME judgments about the effectiveness of the response options. The variations of this approach generally involve allocation of different amount of points for choosing response options that a SME consensus has deemed more effective. As with other SME approaches, more experts produce better keys than fewer experts, so a minimum of 5-7 experienced SMEs are typically recommended for SJT scoring key development efforts.

Once the mean effectiveness rating for each response across judges is obtained, several scoring strategies can be developed depending on the response instructions given to test takers. If a response instruction asks to select one choice (e.g., pick the best response or what you would most likely to do), the mean effectiveness rating can serve as the item score. Alternatively, effectiveness ratings can be re-coded into discrete “effectiveness scores” with the most effective option being coded as + 1,for example, the least effective -1, and the middle two as 0s. The choice of the effectiveness scores is somewhat arbitrary and depends on a rational decision about how to weight effective/ineffective actions. If actual response data is available and unidimensionality assumption is essentially satisfied, polytomous item response theory models could also be fitted to estimate threshold parameters for each SJT item.

If a response instruction asks respondents to select two choices (most and least likely actions), response alternatives may be assigned effectiveness scores in accordance with their effectiveness ratings (e.g., + 2 for most effective, -2 for least effective), and the two selected alternatives are then combined into a single total score (e.g., the score for the chosen “least likely” response is subtracted from the score for the “most likely” response) or both are recorded separately. Finally, response instructions can ask respondents to provide effectiveness ratings for each alternative; then their ratings can be used as is or compared against SME effectiveness ratings and assigned some sort of an agreement code (e.g., the absolute difference between respondent’s effectiveness ratings and SME effectiveness ratings can be summed, such that larger scores reflect greater disagreement and poorer judgment; alternatively, +1 if item is correctly judged to be the most effective/least effective, -1 if the judgment is in the opposite direction). Note that the latter approach would be particularly appropriate when response options represent different behavioral dimensions.

Although SME keys are most commonly used in SJT scoring, Bergman at el. (2006) have shown that multiple scoring keys should be developed for SJT assessments whenever feasible. This is because different keys essentially relate SJT responses to different notions of successful performance. Theoretical or model-based keys tie SJT scores to known performance models and, as such, are useful for feedback and personnel development purposes. SME or novice/expert keys tell whether respondents think in similar/dissimilar fashion to the judging group. Finally, empirical keys help to separate respondents into successful or unsuccessful performance groups.

The key point with regard to developing scoring keys for construct driven SJTs is that until this point, theoretical/SME scoring keys alone have not produced satisfactory results. Our experience suggests that, in order to develop scoring keys that produce data reflecting the intended psychological construct, it is necessary to incorporate an empirical approach to the design of the assessment. In other words, what is ‘correct’ can only be partially determined based on theory. Sometimes the correct answer to an item will also need to be discovered, i.e. decided correct based on whether or not the response correlates with an outcome of interest. Such keys are considered hybrid keys in the language of Bergman et al. Hybrid keys require creating a theoretical and an empirical key and selecting the scoring approach that leads to the highest correlation with the criterion of choice. We describe an example of an empirical key that can be used in the hybrid now.

**Step 5. Construct SJT Forms and Interpretation Guidelines.**

The final step in SJT development is form construction. Where construct driven SJTs are designed, we recommend wherever feasible designing multiple SJT forms of the same psychological constructs. The different SJTs measuring the same construct can contribute to replicating scores for individuals on the same construct but using different SJT measures. Our confidence that we have measured the construct we intended is improved if the scores on two different SJTs measuring the same construct are highly correlated.

At least three different approaches to the construction of parallel form situational judgment tests have appeared (see Lievens & Sackett, 2007, Oswald et al., 2007, and Brummel et al., 2009). Lievens and Sackett (2007) discussed an approach that can be used when tests are expected to be multidimensional and when pre-testing is prohibited. The approach Lievens and Sackett recommended posits that similarities and differences between items can be radical or incidental in nature. Radical features of an item refer to the structural features of the problem that impact its difficulty, whereas incidentals refer to nuisance or surface factors that do not impact the item's difficulty. Two items are considered equivalent if the content domains they sample are the same. In incident isomorphism, two items are considered equivalent if the structural, underlying nature of the item is similar, and only surface features of the problem are different. Under item isomorphism, all features of the item are the same, -- they are essentially clones of one another save for the fact that the wording and grammar has been altered.

The next approach is a procedure described by Oswald et al. (2007), itself an extension of an approach described by Gibson and Weiner (1998). The method is appropriate when pre-testing of items is possible. First, the item set is administered to a sample of respondents and the item level characteristics are recorded, including item mean, item standard deviation, item reliability, and, in an extension of Gibson and Weiner, item correlations with an external criterion. Oswald et al. then used a SAS/IML routine to create thousands of parallel forms that representatively sampled items from across all content sub domains. The SAS/IML routine calculated test level mean, standard deviation, and alpha, and criterion related validity for each of the tests based on the item parameters derived in the previous step. Criteria for optimal assembly of a set of parallel forms are then test level statistics including i) the standardized effect size for the mean difference between the proposed form and overall mean across all parallel forms (this eliminated the majority of the parallel forms); ii) alpha reliability over .70, iii) they specified that the correlation with the external criterion for each test be over .15, the mean correlation of all parallel forms with the criterion. In this manner a very large number of parallel forms can be easily developed.

A third relevant approach to parallel form development was described by Brummel et al. (2009). This article focused on assessment center simulations, although the approach is also relevant for parallel form SJTs. Parallel forms are tests measuring the same construct that yield data with the same true scores, standard deviations, inter-item correlations, and factor structures (Cronbach, 1947). Because this definition of parallelism is rarely achievable in practice, psychometricians have degrees of parallelism, including parallel, tau equivalent, essentially tau equivalent and congeneric. The gist of the approach described by Brummel et al is to develop scoring equations to alter the scores of examinees post hoc, ensuring equivalence.

**Conclusion and Recommendations**

The scientific literature is relatively clear regarding construct driven SJTs – there is inadequate evidence that SJTs are suited for measuring psychological constructs. Despite the lack of evidence, many applied practitioners see considerable advantages in being able to assess psychological constructs with SJTs. This would have numerous benefits. For example, it would permit more accurate feedback in development settings and could be the basis for an unobvious, self-reported training needs analysis. When used in MTMM contexts, SJT measurement of psychological constructs could also improve our understanding of construct validity for commonly studied constructs. Finally, knowing what an SJT measures would enable test developers to create test batteries with less redundancy (e.g., increasing the odds that the SJT adds incremental validity over the other components of the battery). Because there is clearly value in measuring constructs with SJTs, in this article we set out to achieve two goals. First, we reviewed recent research where described designing construct driven SJTs for examples of steps SJT designers could take to improve construct validity evidence for SJTs. Second, we used this review as the basis for a modified SJT development strategy where SJTs are intended to measure constructs.

Numerous recommendations were identified, however, all fall under the banner of paying very careful attention to the validity objectives during the design process. For example, we recommended only attempting construct driven SJT design if the target construct has been assessed previously with multiple other measurement methodologies. A second recommendation elaborated on this point and related to dimensionality. At the level of the scenario stem, we recommended avoiding multidimensionality by portioning multidimensional stems into multiple unidimensional stems based on SME judgments. At the level of the response options, construct driven SJTs would see response options that are unidimensional at the item response level by representing different levels of the targeted construct, rather than instances of different behavioral categories.

To further eliminate unintended multidimensionality, we recommend countering haphazard approaches to item design that results in such a wide variety of response formats. Specifically, we recommended avoiding overly long situational descriptions as they present unnecessary opportunities for unexpected multidimensionality in scenarios. Finally, even after having taken all these steps it is important to realize scores may need to be made to ‘behave’ in a nomological network in accordance with expectations. Therefore, some form of criterion keying might be required. Internal structure (i.e. factor analysis) and reliability analyses should then be repeated on the criterion-keyed scores. These analyses should account for the structure of the data as closely as possible and may involve confirmatory factor analysis or item response theory.

The approach described above is predicated on the assumption that response options can be written which reflect varying levels of the intended construct, without being too transparent in terms of desirability. Clearly, more empirical research is required to confirm this. There is another, as yet untried, option which is to basically ignore the stems and accumulate response options across them into independent measures of various constructs. This would be analogous to treating each SJT stem like an assessment center exercise and the response options like ratings of various dimensions. By relaxing the constraint that the options represent different levels of the same construct, and allowing response options within a stem to measure different constructs, transparency may be reduced.

This approach would require that each item be rated in terms of effectiveness and compared to a SME mean, so that each item’s score is independent from the other item scores within that stem. Taking the absolute difference or within-person standardized absolute difference (McDaniel et al, 2011) would mean that smaller scores reflect better judgment. The underlying assumption in this approach is that participants higher on a construct will exercise judgment as to the effectiveness of behaviors saturated with that construct which is more accurate (similar to the SMEs) than those lower on the construct Note that not every response option would need to contribute to a trait score (construct) and the items comprising a trait score could be drawn from any of the stems in the SJT. This approach also warrants an empirical test.

To conclude, we reinforce our perspective that to date, there is scant evidence that SJTs can effectively measure psychological constructs. However, early studies tended to focus on SJTs that measured general job performance skills, which are inherently multidimensional. Second, many of the psychometric recommendations in this article have been implemented in isolation. We have not identified a single study that attempted to use all of the design considerations we have listed. We are hopeful that their application in combination will yield better results than we have seen so far in the realm of construct driven SJTs. Indeed, it would be interesting to see a meta-analytic study on the effectiveness of SJTs that measure constructs after a sufficient number of studies have been published that implement the steps discussed in this article.

**References**

Bass, B. M. (1985). *Leadership and performance beyond expectation.* New York: Free Press.

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

Bergman, M. E., Drasgow, F., Donovan, M. A., Juraska, S. E., & Nejdlik, J. B. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14, 223-235*.

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

Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The State of Research on Situational Judgment Tests: A Content Analysis and Directions for Future Research. Human Performance, 27, 283-310.

Catano, V. M., Brochu, A. & Lamerson, C. D. (2012). Assessing the Reliability of Situational Judgment Tests Used in High‐Stakes Situations. *International Journal of Selection and Assessment* 20, 3, 333-346.

Chan, D. & Schmitt, N. (2002) Situational judgment and job performance. *Human Performance, 15*, 233–254.

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.

de Meijer, L. A., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-Driven Development of a Video-Based Situational Judgment Test for Integrity. European Psychologist, 15, 229-236.

Fleishman, E., & Harris, E. F. (1962). Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology, 15, 43-56.*

Guenole, N., Chernyshenko, O, Stark, S, Cockerill, T and Drasgow, F (2011). We're doing better than you might think: a large scale demonstration of assessment centre convergent and discriminant validity. In: Nigel Povah and George Thornton, eds. Assessment Centres and Global Talent Management. London: Gower.

Guenole, N., Chernyshenko, O., Stark, S., Cockerill, T., & Drasgow, F. (2012). More than a mirage: A large-scale assessment center with more dimension variance than exercise variance. Journal of Occupational and Organizational Psychology, 86, 5-21.

Guenole, N., Chernyshenko, O., Stark, S., & Drasgow, F. (2014). Are situational judgment tests precise enough for feedback in leadership development? European Journal of Work and Organizational Psychology, 3, 433-443.

Guenole, N, Cockerill, T., Chamorro-Premuzic, T., & Smillie, L. (2011). Evidence for the validity of dimensions in the presence of rater source factors. Consulting Psychology Journal: Practice and Research, 4, 211-237.

Houston, J.S., Borman, W.C., Farmer, W.L., & Bearden, R.M. (2005). *Development of the enlisted computer adaptive personality scales (ENCAPS), renamed Navy Computer Adaptive Personality Scales (NCAPS).* Technical Report No. 502. Navy Personnel Research, Studies, and Technology (NPRST). Navy Personnel Command.

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.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. Industrial and Organizational Psychology, 9, 1-20.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. Journal of Applied Psychology, 90, 442.

Lievens, F., & Motowidlo, S. J. (2015). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. Industrial and Organizational Psychology, 9, 1-20.

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

Luthans, F., & Lockwood, D. L. (1984). Toward an observation system for measuring leader behavior in natural settings. In J. G. Hunt, D. Hosking, C. Schriesheim, & R. Steward (Eds.), *Leaders and managers: International perspectives on managerial behavior and leadership* (pp. 117-141). New York, NY: Pergamon Press.

Luthans, F., Rosenkrantz, S. A., & Hennessey, H. W. (1985). What do successful managers really do? An observational study of managerial activities. *The Journal of Applied Behavioral Science*, *21*, 255-270.

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*, 60–79.

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.

McDaniel, M.A. & Whetzel, D.L. (2007). Situational judgment tests. In D.L. Whetzel & G. R. Wheaton (Eds.). *Applied measurement: Industrial psychology in human resources management*. Mahwah, NJ: Erlbaum. 235-257.

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

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

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.

Mumford T. V., Van Iddekinge, C. H., Morgeson, F. P., Campion, M. A. (2008). The team role test: development and validation of a team role knowledge situational judgment test. Journal of Applied Psychology, 93, 250–267.

Olson-Buchanan, J.B., Drasgow, F., Moberg, P.J., Mead, A.D., Keenan, P.A. and Donovan, M.A. (1998) An interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*, 1–24.

Peus C, Braun S, Frey D. (2013). Situation-based measurement of the full range of leadership model. development and validation of a situational judgment test. Leadership Quarterly, 24, 777–795.

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.

Sharma S, Gangopadhyay M, Austin E, Mandal, M. K. (2013). Development and validation of a situational judgment test of emotional intelligence. International Journal of Selection and Assessment, 21, 57–73.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. Situational judgment tests: Theory, measurement, and application, 135-155.

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-Fidelity Simulations. Annual Review of Organizational Psychology and Organanizational Behavior, 2, 295-322.

Weekley, J.A. and Jones, C. (1997) Video-based situational testing. *Personnel Psychology, 50*, 25–49.

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

Westring A., Oswald F., Schmitt N, Drzakowski, S., Imus, A, Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. Human Performance, 22, 44–63.

Table 1. Review of recent descriptions of design of construct driven SJTs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Design issue | Becker (2005) | Motowidlo et al. (2006) | Bedlow and Frese (2009) | Mumford et al. (2008) |
| Predetermined construct? | Predefined - integrity | Predefined - Extraversion, Agreeableness, Conscientiousness | Predefined - Personal initiative | Predetermined - team roles |
| Single or multiple constructs (includes multidimensional)? | Single | Multiple constructs | Single | Multiple |
| Number of scenarios | 20 scenarios | 5 to 6 per dimensions | 12 scenarios | 8 scenarios, one per team role type |
| Who wrote the scenarios? | Study author | Study authors | Authors based on free response survey | Study authors based on theory |
| Typical scenario length (based on sample scenario | 64 words | 64 - 76 words | 61 words | 101 words |
| Number of response options | 4 responses per scenario | 5 to 10 per scenario | 4 responses per scenario | 6 to 8 responses per scenario |
| Who wrote the responses? | Study author | Student researchers, job incumbents | Study authors | Study authors based on theory |
| Typical option length (based on sample responses) | 17 words | 9 to 25 words | 20 words | 26 words |
| Response instructions | Behavioural tendency | Knowledge instructions (rate effectiveness) | Behavioural tendency | Knowledge instructions (rate effectiveness) |
| Responses as continuous levels or distinct classes? | Not mentioned | Not mentioned | Continuous levels | Categorically distinct |
| Scoring key | Criterion keyed | SME key | SME key | Theoretical key |
| Item analysis | Not mentioned | SME ratings of scenario's trait expressiveness | Not mentioned | Yes |
| Reliability | Not mentioned | Analysis of SME ratings | Rater average deviations, test re-test | Mosier's (1943) reliability for weighted composites |
| Validity evidence | Self report integrity, managerial ratings leadership, potential performance | Correlations with corresponding NEO-PI-R | Self supervisor ratings of construct, nomological network of construct | *r* with Team role performance |

Table 1 Continued

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Design issue* | Westring et al. (2009) | Peus et al. (2013) | Sharma et al. (2013) | Guenole et al. (2014) |
| Predetermined construct? | Pre-determined - goal orientation | Predefined - full range leadership | Predefined -EI | Predefined - High Performance Leadership |
| Single or multiple constructs (includes multidimensional)? | Multiple | Multiple | Multiple | Multiple |
| Number of scenarios | 8 scenarios | 15 scenarios | 16, 18, and 12 scenarios for 3 dimensions, 46 scenarios total | 3 per dimension, 12 dimensions, 36 total |
| Who wrote the scenarios? | Study authors based on theory | Study authors based on focus groups | Study authors | Authors based on workshops, interviews and surveys |
| Typical scenario length | 31 words | 50 words | 25 words | 147 words |
| Number of response options | 3 to 4 responses per scenario | 8 responses per scenario | 3 responses per scenario | 4 responses per scenario |
| Who wrote the responses? | Undergraduates asked for realistic responses | Study authors based on theory | 25 Psychologists (correct), 50 non-psychologists | Experts and job incumbents |
| Typical option length | 18 words | 35 words | 25 words | 35 words |
| Response instructions | Behavioural instructions | Knowledge instructions (rate effectiveness) | Behavioural tendency | Knowledge instructions |
| Responses as continuous levels or distinct classes? | Categorically distinct | Not mentioned | Not mentioned | Continuous levels |
| Scoring key | CFA of effectiveness ratings | SME and Criterion keyed | SME key | Criterion keyed |
| Item analysis | Yes | Yes | Factor analysis | Based on low SME agreement |
| Reliability | Not mentioned | Inter-rater reliability | Internal consistency | Discussed meta-analytic estimates |
| Validity evidence | GPA / SAT scores | MTMM with MLQ i.e. same constructs by Likert scale by self report | Low r with self reported EI, correlations with life satisfaction, academic achievement, GMA | MTMM evidence using 360-degree feedback measure of SJT competency model |