Are Predictions based on Situational Judgment Tests Precise Enough for Feedback in Leadership Development?

Nigel Guenole [1,2]

Oleksander Chernyshenko [3]

Stephen Stark [4]

Fritz Drasgow [5]

IBM Smarter Workforce Institute[1], Goldsmiths, University of London [2], Nanyang Technical, University [3], University of South Florida [4], University of Illinois at Urbana Champaign [5]

Please send correspondence to Nigel Guenole at nigel.guenole@uk.ibm.com, or IBM, 47 Mark Lane, Tower Hill, London EC3R 7QQ, United Kingdom.

Abstract

Situational judgment tests (SJTs) have much to recommend their use for personnel selection, but because of their low reliability the role of SJTs in behavioural training is largely unexplored. However, research showing that SJTs cannot measure homogenous constructs very well is based exclusively on internal analyses, for example, alpha reliability and factor analysis. In this study, we investigated whether patterns of correlations with external criteria could be used to show that SJT dimension scores are homogenous enough for feedback purposes in leadership development. A multidimensional SJT was designed for 268 high potential leaders on a development programme and used in conjunction with a multisource feedback instrument that measured the same competency framework. The SJT was criterion keyed using against the multisource feedback instrument using an *N*-Fold cross validation strategy. Convergent and divergent correlations between the SJT scores and corresponding multisource dimension scores suggested that SJT scores can be constructed in a way that permits dimension level feedback that would be useful in leadership development.

Are Situational Judgment Tests Precise Enough for Leadership Development?

Situational judgment tests (SJT) are a type of measurement method that can be used to assess a variety of managerial dimensions including social skill, conflict resolution style, or leadership capability (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel & Nguyen, 2001, Weekley & Ployhart, 2006). In the personnel selection and development literature, SJTs are classified as low-fidelity work samples (Motowidlo, Dunnette, & Carter, 1990). Typical SJTs consist of several scenarios representing challenging work-related situations. The content of a specific scenario can be presented to respondents in a written, audio, or video format, although the written format is by far the most common. Once an item stem is presented, respondents are asked to choose the most effective and/or least effective response among a set of seemingly equally desirable alternatives. Each alternative typically describes an action that could be taken in response to the scenario situation and has an associated "effectiveness" value.

Numerous authors have outlined the case for SJTs in selection context (e.g. (Clevenger et al, 2001, Cullen, Sackett, & Lievens, 2006). Although they can be costly to develop, SJTs are often still more affordable to develop and run than assessment centres or work shadowing programmes. They can also be relatively easily deployed via the Internet or a local area network within organizations, and require considerably less testing time than these other methods. SJTs can also be objectively scored in a manner more like maximum performance measures (e.g., assessment centre simulations or cognitive aptitude tests) than typical performance measures. This means they SJT questions are less susceptible to response distortion issues commonly associated with Likert-type self-report measures. In addition, they also lead to favourable candidate reactions (Anderson, Salgado, Hulsheger, 2013).

The validity of the SJT measurement method also explains their use in applied settings. McDaniel et al. (2001) showed with meta-analysis that the average corrected

criterion validity of well-developed SJTs was .34 for predicting job performance. Mechanisms that have been proposed to explain the relationship by Motowidlo, Dunnette, & Carter (1990) and Ployhart & Ehrhart (2003) include a) that SJT scenarios reflect samples of behaviour, and scores correlate with future performance because past behaviour is a good predictor of future behaviour (behavioural consistency); b) that responses to SJT scenarios reflect respondent signalling about their intentions to behave in particular ways in future situations that are like the scenarios, and c) that responses reflect job knowledge required for effective performance, and individuals apply the knowledge they show on the SJT in subsequent situations in the workplace. Researchers have also noted an attractive feature of SJTs is their incremental validity over other assessment methods and low adverse impact against women and ethnic minorities (Chan & Schmitt, 2002; Clevenger et al, 2001; Motowidlo et al., 1990; Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998; Weekley & Jones, 1997, 1999).

SJTs in leadership development

While SJTs have traditionally been used in personnel selection contexts, there is reason to believe they could have useful applications in training programmes. Because our sample is comprised of leaders, we focus specifically on leadership development programmes. A crucial advantage of SJTs for leadership development is that, due to the ability to make items highly contextualized, they can be considered *samples* of work performance rather than *signs* of future work performance (Sackett & Lievens, 2008). The degree of contextualization of SJTs and other assessment methods is referred to as the fidelity of the assessment method (e.g. Lievens & Patterson, 2011). This increased opportunity for item contextualization with SJTs allows test designers to prepare items that are more reflective of the complex situations in which leaders are required to exert influence than traditional Likert style items allow. Before situational judgment tests can be used in the same fashion for development as assessment centres or multisource feedback, it is important to demonstrate that SJTs can be used to deliver precise feedback on specific dimensions where each dimension correlates with

meaningfully different work-related outcomes. We note that research showing such an effect would have implications for personnel selection and development. However, such a finding is not as critical in personnel selection contexts where individual dimension scores are not as emphasized as overall scores. On the other hand, in development settings, narrow dimension scores are as, if not more, important than overall scores. It is these narrow scores that tell candidates where to focus their development efforts. Moreover, our primary focus is on feedback in leadership development contexts because our sample was comprised of participants on a leadership development programme.

Evidence from analyses of SJTs scores to date suggests that SJTs do not seem to be assessing homogeneous characteristics. On the contrary, they are known to be highly heterogeneous (Chan & Schmitt, 2006, Lievens, Peeters, & Shollaert, 2008; Weekley & Ployhart, 2006, Whetzel & McDaniel, 2009). To this point, however, attempts to measure constructs with SJTs have been based primarily around internal analyses such as factor analysis or internal consistency analyses. No research has examined whether SJT scores show meaningful patterns of correlations with external variables suggesting that SJT subscales are assessing distinct constructs. The central goal of this study is to examine whether SJT dimension scores are homogenous enough to predict distinct outcomes, as is required for feedback in leadership development, despite the fact that the results of internal analyses alone indicate that SJT dimension scores are highly heterogeneous. If this were the case, feedback on SJT dimension scores could be interpreted in terms of the candidate's strengths and weaknesses.

One research design that would address this issue is to examine the correlations between a multidimensional SJT of a given leadership model and multisource ratings of the same dimension model (i.e. isomorphic content alignment between predictors and criteria). This design would allow us to see whether the layperson assumption about validity holds. In psychometric parlance this can be considered an evaluation of convergent and divergent validity via a multi-trait-multi-method correlation matrix (Campbell & Fiske, 1959). If the

scores for the same dimensions across measurement methods could be shown to be related, the applied relevance of SJT scores for on-the-job behaviors would be more explicitly clear than has been shown to date. It is very important to note that an SJT and a multisource feedback instrument assessing the same competency model represent maximum and typical measurements of the same constructs. Whereas in a typical MTMM design it traits are measured by different methods, in the current design traits are being measured with one method (SJT) and performance related manifestations of these traits are being measured with another method (multisource feedback). Therefore, it would be unreasonable to compare the magnitude of the 'convergent' correlations between the same construct across methods against any other standard than the typical magnitude of SJT – job performance correlations. While corrected correlations with job performance have been reported at high as .35 (McDaniel et al., 2001), uncorrected correlations are often much lower. Lievens et al. (2006) for example made a case for the utility of SJT to performance correlations at low as .11.

Hypothesis development

In hypothesizing about why this expected pattern of relationships might hold between SJT dimension scores and corresponding multisource dimension ratings we considered three theoretical/conceptual perspectives. The first was Motowidlo and colleagues' theory that SJTs represent past samples of behavior that predict subsequent behavior (Motowidlo, Dunnette, & Carter, 1990). By explicitly improving the point-to-point correspondence between SJT dimensions and performance outcomes by isomorphic alignment between the content models underpinning the predictors and criteria, the correlations between corresponding constructs assessed via different measurement methods would be expected to be stronger. While earlier work (e.g. Lievens, Buyse, & Sackett, 2005) has shown the importance of appropriate theoretical alignment between SJT predictors and criteria, until now the issue of content isomorphism between specific SJTs measures and performance criteria has not been considered. Put another way, knowledge of when and how to use aspects of a behavioral repertoire delineated in a dimension framework ought to determine the extent

to which those precise aspects of behavior are appropriately used in practice.

The other theoretical perspectives we considered were the widely accepted distinction in the SJT literature between measurement methods and constructs, and the distinction between maximum and typical performance. Our expectations based on the construct-method distinction were that the correlations between the *same* constructs across methods should, on average, be higher than the correlations between different constructs across methods. We anticipated maximum performance capability on a dimension should have greater implications for typical performance on that *same* dimension than it does for typical performance on any other dimension. Based on these considerations we made two hypotheses.

Hypothesis one. Correlations between dimension scores on the SJT and the corresponding multisource performance rating for that dimension from the multisource feedback instrument will be positive and significant.

Hypothesis two. The average correlation between SJT dimension scores and corresponding 'on-target' multisource dimension scores will be greater than the average correlations between all other SJT dimension scores and multisource dimension scores.

Insert table 1 about here

Method

Behavioural framework

The model of leadership capability that we chose to assess was the High Performance Behaviours (HPB) model (Guenole et al., 2011, 2012, 2013). The HPB dimensions emerged from a qualitative review of the research literature related to effective managerial behaviour, and included key research programs such as the Ohio State studies (Stogdill, 1950), the Michigan studies (Likert, 1961), and studies carried out at Harvard (Bales, 1950). The design goal for the HPBs was to stipulate a fairly comprehensive set of leadership dimensions, each with clearly defined boundaries, and that covered the spectrum of behaviours embodied by effective leaders. In total, twelve dimensions are included in the HPB model, similar to what

Fleishman et al (1991) found in their comprehensive analysis of taxonomies of leadership behaviours. The dimensions of the HPB model are defined in job-related language and grounded in job analysis. The competencies are Information Search (IS), Concept Formation (CF), Conceptual Flexibility (CX), Empathy (EM), Teamwork (TW), Developing People (DP), Influence (IN), Building Confidence (BC), Presentation (PR), Proactivity (PO), Continuous Improvement (CI), and Customer Focus (CU). Definitions of each of the twelve behavioural dimensions are presented in table 1.

Participants

The sample for this analysis is comprised of 268 managers in a multinational pharmaceuticals business. The sample was 60% male. These managers were participating in a leadership program designed for high potential staff. These managers were middle managers or first line managers thought capable of moving into more senior management roles. This is consistent with the intended application of the instrument, which is designed for large-scale selection or development into low to mid-level management roles where assessment centers are too costly to implement. In step one of the leadership development program, managers completed the new SJT measuring the HPBs. Participants then received feedback on their performance against the model, identifying those dimensions on which they demonstrated strong knowledge and the dimensions that showed room for development. To provide a richer perspective on their developmental needs, in a second step of the program, all participants took part in a multisource feedback program. Each was asked to nominate feedback providers who, along with the program participants themselves, would rate the participant on their performance on the dimensions underpinning the SJT. Having the SJT and the multisource feedback instrument data on the same participants served two functions 1) participants were provided an indication of how others see them in relation to the dimensions measured, relative to how they see themselves and 2) subsequent completion of the multisource feedback instrument that measured the same dimensions permitted development of an empirically based scoring key for the current study.

Measures

Leadership SJT Development

Generating initial item stems. Fifty-nine job incumbents with extensive leadership experience were asked to identify leadership critical incidents for the 12 dimensions of the HPB model in a combination of interviews (47) and workshops (12). Participants came from the following industries: Banking (3), Energy (3), Finance (1), Government (4), International Development (1), Law (6), Manufacturing (1), Media (4), Pharmaceuticals (19), Technology (1), Telecommunications (2), and Transportation (13). This group included Senior Vice Presidents of Operations and Development, Heads of R&D, Finance, Security, and Engineering, and Senior Managers from Sales, Maintenance, Planning, or HR departments. Forty-six per cent of the job incumbents were female; all had at least a bachelor level university degree; the majority of incumbents supervised more than 10 subordinates (the number of direct and indirect reports ranged from 2 to 450); and more than half of them had more than 10 years of managerial experience. In total we developed 94 scenarios.

Subject matter expert (SME) rating exercises. Scenarios were edited to a common format and a first SME exercise was undertaken to confirm which scenarios measured each dimension. Five consultants with deep knowledge of the HPB model from conducting leadership development workshops but who did not participate in the initial item writing were asked to serve as SMEs and rate each of the 94 scenarios in terms of the dimension it assessed. If the majority of SMEs (3 or more) agreed on the dimensional designation for a particular stem, then this stem was classified into that leadership dimension; if SMEs disagreed, the stem was designated as "unclassified". SMEs agreed on dimensional designations for 61 of the 94 scenarios, disagreed on 33 scenarios, and some of the scenarios were changed from their initial dimensional designations. The 33 unclassified scenarios were revised and/or split into smaller scenarios to focus on only one aspect of leadership performance behaviour. A second SME study using seven new judges was conducted. To evaluate the extent to which SMEs agreed in their primary and secondary dimensional

ratings, we used the intraclass correlation (ICC1). ICC is commonly used to measure inter-rater reliability for two or more raters and is the ratio of between-groups variance to total variance. The resulting ICC (average measure of reliability for the one-way random effect model) was .88.

In this study we developed a 36-item SJT where the best three scenarios per behaviour were included for each HPB dimension. An example of a situational item stem resulting from this process from the Information Search HPB dimension is presented in appendix A. This scenario illustrates that our scenarios are at the high end of the detail continuum, suggesting that there is likely to be a cognitive load on participants. However, the complexity in the scenarios was necessary to represent the richness of the information provided to us by the participants in the scenario generation, which in turn mirrors the complexity of the situations respondents face in the work environment.

Response alternatives. The first set of responses was obtained from the initial behavioural interviews and critical incident workshops in which job incumbents were asked to recall what action was actually chosen in the real situation. In addition, as part of the critical incident workshop, job incumbents were asked to write short descriptions of how they would respond to a specific situational stem and what were other plausible effective and ineffective responses. Because we wanted high homogeneity, we emphasized the need for a subtle change to the response options that were generated. For this SJT, as far as was possible, responses writers were encouraged to generate responses on a continuum reflecting more or less of the dimension being assessed. This was not always possible, because in numerous scenarios did not reflect gradations of the underlying continuum. Wherever it was possible, however, we followed this principle. Appendix A shows an example of the four response alternatives for the Information Search scenario presented earlier, along with the corresponding intended effectiveness ratings.

Response instructions. 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. Behavioural 'would do' instructions tend to overemphasize a leader's "typical behavior" and, as McDaniel et al. (2003) have shown, this makes SJT scores correlate highly with personality. For example, the meta-analytic correlation between SJTs scores with behavioral instructions and the Emotional Stability personality dimension was found to be .51, suggesting considerable overlap in the constructs assessed. Thus, we implemented the following instructions: "Below is a list of four possible actions you could take in response to the situation. If you were a leader, which action would be most effective and which action would be least effective? Please select the "Most" option for the "most effective" action and the "Least" option for the least effective action. In response to each SJT item, participants were asked to indicate which of the four response options they believed was most effective (and subsequently coded 1) and which of the options they considered to be least effective (subsequently coded -1). An example of an SJT item for the information search competency is presented in appendix A.

Multisource feedback instrument. Participants completed ratings against the HPB framework, and were rated by their manager, two or more peers, and two or more direct reports. Whilst two or more peers and reports took part in the process, in this analysis we had access to the first rater of each type that program participants nominated. The multisource-degree feedback instrument that they completed was a 60-item measure. Each HPB dimension was measured with five six-point Likert-style items. Previous multi-trait multi-method confirmatory factor analyses of data from this instrument have shown it has sound psychometric properties (Guenole et al., 2011). An example of a scale item for the information search competency is "Uses many different sources or methods to gather information about work issues."

Analyses

Empirical keying with N-fold cross validation. All analyses were executed with the statistical computing environment R 3.0.2. In a first step we created dimension totals for the HPB multisource-feedback instrument against which the corresponding SJT scales could be

keyed. To do this we calculated scale composites for each HPB as the simple sums of HPB ratings from across rater groups. We wished to remove the possibility that the validity coefficients observed would be spuriously inflated by scoring participants using a key based on a sample of which they were a part. Therefore, we implemented *N*-fold cross validation (Brieman, Friedman, Olshen, & Stone, 1984). *N*-fold cross validation 'holds out the responses of person *j* and computes an empirical key based on the remaining *N*-1 persons, which is used to score person *j*' (Bergman et al. 2006, p225). Interested readers can obtain the R scripts for this process from the corresponding author. We selected a base rate of 20 candidates as a minimum for implementing our decision rules, which were as follows: a) we gave a point if an option was positively correlated with performance and an examinee chose it as their best, and b) we gave a point if an option was negatively correlated with performance and an examinee chose it as their worst. Only minor variations from these rules occurred, for example, in certain instances an option was scored as zero due to near zero option to performance correlations. In other cases, where the base rate requirement for positively endorsing an item was not met, but where the base rate was met on the negative endorsement, we implemented the reverse of these rules i.e. penalizing for wrong choices. Therefore, overall, the key is a hybrid-scoring key. Each of the three SJT items per dimension was keyed in this manner. Next, the SJT dimension score was computed by summing the item scores within HPB dimensions.

Accuracy of measurement. If the SJT scale scores are to be used for feedback, users will want to know that the dimensions are well measured. This is typically examined with the standard error of measurement (SEM). A reliability coefficient is required for its computation for each scale, but Cronbach's alpha is an inappropriate reliability estimate because the heterogeneity of responses violates the assumption of Cronbach's alpha that items are homogenous. The emerging consensus in the SJT literature is other forms of reliability are better, and in particular, test-retest reliability or related coefficients such as the coefficient of reliability and stability (Schmidt, Le, Ilies, 2003). We do not have data to

permit these estimates currently. Therefore, we estimated the standard error of measurement for each scale using high (.82) and conservative (.46) test-retest SJT reliability values from a recent meta-analysis by Catano, Brochu, & Lamerson (2012).

Examining the validity of the empirical key. To assess the utility of the key, we examined the zero order inter-correlations of HPBs measured using the SJT and HPBs measured using the multisource feedback instrument. Specifically, we were interested first in whether the correlations between the same dimension assessed via SJT and multisource feedback were positive, significant, and also in their magnitudes (hypothesis 1). Second, we were interested in whether correlations between the same dimensions measured in different ways (i.e., the hetero-method-mono-trait correlations) were greater than measures of the different dimensions measured in different ways (i.e. hetero-trait-hetero-method correlations) (hypothesis 2). If evidence of this effect were found, it would suggest support for the convergent and discriminant validity of the new SJT measure. Because this technique extends beyond a single measurement model approach we consider the approach structural (as opposed to a measurement model) multi-trait multi-method analysis. Finally, we also examined the overall validity for the SJT across all dimensions.

Results

Insert table 2 about here

The multisource feedback scales against which we criterion keyed were observed to have good internal consistency. The value were as follows: IS .75, CF .66, CX .75, EM .70, TW .74, DP .80, IN .76, BC .79, PR .83, PO .66, CI .76, CU .80). Some researchers have argued that alpha produces over-estimates of reliability due to idiosyncratic variance being treated as true variance. Therefore, here we also report inter-rater reliabilities for each of the dimensions. These are indeed lower than alpha. The values were as follows: IS=.41, CF=.29, CX=.27, EM=.35, TW=.30, DP=.46, IN=.41, BC=.50, PR=.44, PO=.37, CI=.37, CU =.45. However, these values are similar to meta-analytic estimates of inter-rater reliability of job performance reported by Viswesvaran, Schmitt, and Ones (2002). The

standard errors of measurement presented in table 2 show that under both the conservative and favorable assumptions about the test retest reliability of the current test the SEMs are tightly positioned around the scale means. We do not present the mono-method-hetero-trait correlations due to space constraints, but for the SJT the average was .02 indicating that the SJT scores were uncorrelated. The largest was .19. The average mono-method-hetero-trait correlation was .43 for the multisource feedback instrument, while the maximum was .70, indicating these scores were moderately correlated.

The correlations in table 3 show that positive associations are observed between HPB dimensions measured using the SJT and multisource dimension scale scores isomorphically aligned for content. For example, the correlation between Information Search measured by the SJT and Information Search measured by the multisource instrument is .16. Collectively, these findings, represented by the diagonal elements of table 3, provide support for hypothesis one. The correlations between SJT HPB scores and corresponding multisource HPB scales are of useful sizes and all were significant. The lowest correlation was .12. The average mono-trait-hetero-method correlation was .16, while the largest correlation was .20. While these values might seem small, the convergent correlations must be judged in the context of other SJT work correlating scores with performance outcomes. Lievens and Sackett (2007) for example, presented correlations between a high stakes testing SJT and subsequent grade performance that ranged between .11 and .18. The correlations reported in this study therefore compare well with the uncorrected correlations presented in other published SJT validity work. Moreover, the sum of all SJT dimensions was correlated with the sums of all the multisource dimensions to arrive at the total uncorrected validity of the SJT of .33, comparable to the corrected validity reported by McDaniel et al. (2006). While SJT scores do not always correlate highest with their corresponding multisource-degree feedback score, for ten out of twelve competencies the score between the SJT dimension and the corresponding multisource dimension is in the top two correlations.

Insert table 3 about here

To see whether our results supported hypothesis two, we examined the average mono-trait-hetero-method correlation and compared its magnitude to the average hetero-trait-hetero method correlation. If the average mono-trait hetero-method correlation is higher than the average hetero-trait-hetero method correlation, it would indicate evidence for convergent validity of the SJT. In this case, the average mono-trait-hetero-method correlation was .16, and the average hetero-trait-hetero-method correlation was .04. Most of the hetero-trait-hetero-method correlations were non-significant. The results of the criterion keying therefore supported hypothesis two. Not only were positive associations observed between the same dimensions measured in maximum and typical performance measurement methods isomorphically aligned with respect to content, but the strongest relationship between SJT dimensions and multisource dimensions were, on average, between the dimensions that were isomorphically aligned for content. Furthermore, a Wald chi-square test of a model constraining the mono-trait-hetero-method correlations equal to the hetero-trait-mono-method correlations was rejected, further enforcing the conclusion that the mono-trait-hetero-method correlations were greater.

Readers may be interested to see the correlations between the SJT dimension scores and the different rater groups. These are presented in table 4. In general, these correlations were positive and slightly smaller due to multisource dimension scores for each dimension being more reliable when aggregated across raters.

Insert table 4 about here

Observed versus Operational Validities

The observed convergent correlations we have reported are greater than the observed divergent correlations, and other studies have reported smaller criterion correlations for SJTs. Nevertheless, the correlations in table 3 are small. Some might be concerned that the correlations are too small to warrant dimension level feedback. In such a situation, we believe is worth noting here that the reported correlations are observed correlations. That is to say, they have not been corrected for measurement error due to unreliability as correlations

are corrected in a structural equation modelling framework, for example. However, given the availability of appropriate reliability estimates, it is quite possible to correct these observed correlations for unreliability and check the effect.

Ones, Dilchert and Viswesvaran (2007) discuss operational validities as correlations corrected for unreliability in either the criterion, the predictor, or both the criterion and the predictor. The reliability estimates we have available are intra-class reliabilities for the criterion. Correcting the observed correlations for criterion unreliability, but not predictor unreliability, leads to the correlations one could expect to observe in practice. These are presented in table 5. The mean mono-trait-hetero-method correlation under this approach is 26, and all correlations fall between .20 and .35. We did not have appropriate test-retest reliability estimates for predictor corrections, so we used the meta-analytic estimate of .52 reported by Catano et al. (2012) to estimate correlations corrected for predictor and criterion related unreliability. This led to a mean mono-trait-hetero-method correlation of .36, with all such correlations falling between .30 and .49. We do not present these latter correlations, but they are available from the corresponding author.

Insert table 5 about here.

Discussion

SJTs have become an important selection methodology because of characteristics such as sound validity, high user acceptance, reduced impact, and cost effectiveness. Their use for leadership development has not been extensively investigated. However, the SJT method is attractive for leadership settings because the ability to contextualize items allows the scenarios to more accurately reflect the nature of leadership challenges than typical Likert-type items allow. They also are less costly to run from an administration perspective and require less participant time than methods such as assessment centers and work shadowing. However, a reasonable pre-requisite for effective use in leadership development settings is evidence that multiple dimensions can be measured and that these dimensions have different consequences for performance at work. More specifically, it is necessary to say that high

scores on dimensions assessed by leadership SJTs corresponds to more effective workplace displays of the behavioral dimensions the SJT is purported to measure.

Analyses of the internal structure of SJTs to date have suggested that this assumption does not hold. To further test assess this assumption using externally focused analyses we studied the relationships between a multidimensional SJT of leadership capability (maximum performance setting) and a multisource feedback instrument (typical performance setting) assessing the same behavioral model. The criteria for inferring what is measured under this approach are shifted from internally focused reliability and factor structures to patterns of correlations with external variables. We hypothesized that the mono-trait-hetero-method correlations would be positive and significant based Motowidlo's knowledge determinants theory of SJT validity (Motowidlo, Dunnette, & Carter, 1990). Further reasons for expecting this result included that a) the same content domain is assessed across the measurement methods, b) maximum performance on a dimension was expected to have the greatest typical performance implications for the corresponding dimensions.

Our results provided preliminary support for convergent and divergent validity of the SJT dimensions, showing that SJTs can be constructed to yield evidence that meaningful multidimensionality in SJT scores is observable based on patterns of relationships with appropriately selected external performance criteria. More precisely, a key was generated that resulted in positive correlations between the same dimensions measured by both SJT and multisource approaches. Further, the correlations between the same dimensions measured using an SJT and multisource approach were, on average, greater than all other correlations between the SJT and multisource measurement approaches to measurement.

The SJT scores were also uncorrelated. This suggests that any concerns that subscale scores are not differentiated from one another is misplaced. The size of the "convergent validity" correlations was also small. However, it should be kept in mind that these correlations are not convergent validity in the traditional sense of two measures assessing the same construct at the same point in time. Rather, these correlations represent the relationship

between maximum and typical indicators of the same construct assessed using different measurement methods. These correlations need to be judged in the context of other uncorrected validity coefficients between SJT scores and subsequent performance. From this vantage point, the hetero-method-mono-trait 'convergent' correlations are in the same range as past research on SJT literature discussing uncorrected correlations (e.g Lievens & Sackett, 2007).

Comparison of results to construct validity of assessment centers

Researchers may consider it instructive to compare the current research on SJTs to assessment center research, given that both the current study and typical assessment centers adopt MTMM-based designs. Until a recent study by Guenole, Chernyshenko, Stark, & Drasgow (2013), the only one other study in the history of assessment center research by Magdalen et al. (2000) had reported an assessment center with greater mono-trait-hetero-method correlations than hetero-trait-mono-method correlations. In addition, each of four key meta-analyses reported in the academic literature found evidence than assessment centers produce smaller mono-trait-hetero-method correlations than hetero-trait-mono-method correlations (Bowler & Woehr, 2006; Lance et al., 2008; Lievens & Conway, 2001; and Woehr & Arthur, 2003). The results of the current SJT study showing greater mono-trait-hetero-method correlations than hetero-method-mono-trait correlations compare favorably to internal validity evidence for assessment centers. Nevertheless, even with corrections for unreliability it must be acknowledged that the height of the correlations is lower than that for Assessment Center MTMM studies.

Conclusion and Implications

While these findings have important implications for selection and for development, they are particularly useful for leadership development because it makes the feedback messages that test interpreters provide to leadership development program participants clear. If you score highly on a particular dimension, candidates want to know that their co-workers will say they are more effective at using that dimension. Before the current research this

expectation, which corresponds to a very reasonable layperson view of validity, had not been supported by empirical findings. This was in part because SJTs have until now been primarily focused in selection settings where the total score is most important, and in part also due to historical views about the heterogeneity of SJTs. However, based on this study, feedback for leaders against a priori competency frameworks looks to be possible because SJTs can be constructed such that different SJT dimension scores different consequences for important work outcomes. The use of SJTs in leadership development could occur either as a test for the purposes of training needs analysis, in which case the SJT could occur prior to training; or to evaluate learning due to training, in which the SJT might be used in a pre- and post-test design. Alternatively, one-to-one discussion of the scenarios in the test might be used to engender understanding amongst developing leaders about why particular options are likely to be more effective than others. If multidimensional SJT scores are desired, we suggest at this point that they be derived based on criterion keying of relationships with external variables. We also suggest that measurement researchers might wish to investigate fitting their measurement models that have failed to find evidence of meaningful multidimensionality for SJTs previously to data sets where different scales are keyed against different external criteria.

There are several limitations to this study that we now consider. Principal among these is a potential limitation with all criterion keys, i.e., the extent to which the key generalizes to a) other samples using the same measures, and b) other samples using different criterion measures. The use of the *N*-fold cross validation strategy should mitigate this point, but it is an issue we are continuing to investigate. A second limitation is that the SJT dimension scores were based on just three items per dimension. Because reliability is in part a function of test length, a three-item test is not expected to be reliable, and so lower correlations with criterion measures are to be expected. Including more items to assess each HPB dimension in the SJT would likely produce higher correlations with criterion measures, but at the cost of being more burdensome on research participants. Trying theoretical rather than empirical

keys could also be a strategy that would lead to larger criterion related correlations. Overall, this study provides initial evidence for the use of SJTs for leadership development. Much additional research is needed – longer assessments, contrasting the validities for supervisor vs. peer ratings of performance, etc. – but the pattern of correlations in Table 3 indicate that SJTs may prove useful and can be designed to be in tune with lay person expectations of validity.

References

Anderson, Salgado, Hulsheger (2013). Applicant Reactions in Selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. International Journal of Selection and Assessment, 18, 291-304.

Arthur, W., Woehr, D., & Maledgen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. Journal of Management, 26, 813–835. doi:10.1016/S0149- 2063(00)00057-X

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.

Bales, R.F. (1950), A set of categories for the analysis of small group interaction, American Sociological Review, 15, 257-263.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. Journal of Applied Psychology, 91, 1114–1124. doi:10.1037/0021-9010.91.5.1114

Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and *regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Campbell, D. T. & Fiske D, W. Convergent and discriminant validation by the multitrait multimethod matrix. Psychological Bulletin, 56, 81-105.

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, 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.

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

Goodman, J. S., Wood, R. E., & Chen, Z. (2011). Feedback specificity, information processing, and transfer of training. Organizational Behavior and Human Decision Processes, 115, 253-267.

Guenole, N., Chernyshenko, O., Stark, S., Cockerill, T., & Drasgow, F. (2013). 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, Cockerill, T and Drasgow, F (2012). 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, xx-xx. ISBN 978-1-4094-0386-9

Guenole, Nigel, Cockerill, T, Chamorro-Premuzic, Tomas and Smillie, Luke D. (2011). Evidence for the validity of dimensions in the presence of rater source factors. Consulting Psychology Journal: Practice and Research, 63, 203-218.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center post-exercise dimension ratings. Journal of Applied Psychology, 89, 377–385. doi:10.1037/0021-9010.89.2.377

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology,* 90, 442-452.

Lievens, F., & 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(5), 927.

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

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. Journal of Applied Psychology, 92, 1043.

Lievens, F., & Conway, J. M. (2001). Dimensions and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. Journal of Applied Psychology, 86, 1202–1222. doi:10.1037/0021-9010.86.6.1202

Likert, R. (1961). New patterns of management. New York, NY: McGraw-Hill.

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, 640–647.

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.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. Personnel Psychology, 60, 995-1027.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. International Journal of Selection and Assessment,11, 1-16.

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. Journal of Applied Psychology, 98, 114.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. Annual Review of Psychology, 59, 419-450.

Schmidt, Le, Ilies (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. Psychological Methods 8, 206-224

Stogdill, R. M. (1950). Leadership, membership and organization. Psychological Bulletin, 47, 1–14.

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.

Weekley, J. A., & Ployhart, R. E. (2006). Situational judgment tests: Theory, measurement, and application. Lawrence Erlbaum Associates Publishers.

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

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: unconfounding construct-level convergence and rating difficulty. Journal of Applied Psychology , 87 , 345-354.

Table 1. HPB dimension definitions

Dimension	Code	Definition
Information Search	IS	Gathering a rich variety of information from many different sources about events
Concept Formation	CF	Linking information to form new ideas that explain the underlying causes of events
Conceptual Flexibility	CX	Seeing issues from many different perspectives to compare options prior to taking action
Empathy	EM	Encouraging others to express openly their real thoughts and feelings
Teamwork	TW	Creating effective teams within the unit and across related departments or functions
Developing People	DP	Providing staff with the resources, coaching, feedback and training to develop their capability
Influence	IN	Using persuasive arguments and the goals and interests of others to build support for ideas
Building Confidence	BC	Making your stance on issues clear
Communication	CO	Making clear and concise presentations and establishing effective communication processes
Proactivity	PO	Designing implementation plans and outlining actions and responsibilities
Continuous Improvement	CI	Setting goals and targets and monitoring progress, in order to improve performance
Customer Focus	CU	Setting targets focused on adding value for the customer

Table 2. Standard error of measurement under different assumptions about test-retest reliability

HPB	SJT Mean	Std dev.	SEM 1	SEM 2
IS	2.16	.62	.26	.46
CF	1.70	.88	.37	.65
CX	1.08	.71	.30	.52
EM	.98	.57	.24	.42
TW	1.33	.74	.32	.55
DP	1.20	.72	.30	.53
IN	-.55	.61	.26	.45
BC	.25	1.05	.45	.77
PR	.77	1.24	.52	.91
PO	2.02	1.03	.44	.76
CI	-.61	.53	.23	.39
CU	.52	.86	.36	.63

SEM 1 is calculated under a favourable assumption that test-retest reliability is high (.82), and SEM 2 is calculated under a conservative assumption that test-retest reliability is low (.46). Both values are based on Catano et al. (2012).

Table 3. Mono-trait-hetero-method and hetero-trait-hetero-method correlations between SJT and multisource dimension scores

	ISMSF	CFMSF	CXMSF	EMMSF	TWMSF	DPMSF	INMSF	BCMSF	PRMSF	POMSF	CIMSF	CUMSF
ISSJT	**.16***	.01	-.04	.07	.04	-.03	-.07	.05	.14*	-.03	.11	-.07
CFSJT	.07	**.19****	.13*	.13*	-.01	.07	.05	-.01	.03	.00	.02	.15*
CXSJT	-.11	-.05	**.12***	.05	.02	-.11	.04	.02	-.04	.00	-.04	-.03
EMSJT	-.02	.02	.04	**.13***	-.01	.04	-.12	-.01	.02	-.11	-.08	.00
TWSJT	-.04	.01	.07	.10	**.16***	.07	.10	-.07	.04	-.02	.03	-.04
DPSJT	.03	.01	.16**	.15**	.07	**.19***	.06	.13*	-.01	-.02	.04	.07
INSJT	.00	.06	.11*	.04	.04	-.03	**.13***	.04	.06	.08	.00	.17*
BCSJT	.08	-.03	-.03	.08	.21**	.15*	.10	**.14***	.07	.15*	.21*	.09
PRSJT	.08	-.05	.08	.15**	.09	.02	-.11	.04	**.14***	.03	-.02	-.08
POSJT	.01	.07	.07	.22**	.13*	.08	.07	.15*	.15*	**.20****	.05	.07
CISJT	.08	.01	.04	.10	.14*	.14	.15*	.11	.03	.17*	**.18***	.20*
CUSJT	-.02	-.08	-.05	.04	.02	-.19*	.07	-.07	.02	.02	-.07	**.15***

**. Correlation is significant p<.01 level (1-tailed).
*. Correlation is significant p< .05 level (1-tailed).

Table 4. SJT correlations with multisource ratings by rater group

	ISSJT	CFSJT	CXSJT	EMSJT	TWSJT	DPSJT	INSJT	BCSJT	PRSJT	POSJT	CISJT	CUSJT
Manager	.06	.08	.06	.08	.10	.02	-.02	.09	.02	.08	.10	.04
Peer	.05	.02	.08	.04	.09	.14	.15	.13	.19	.13	-.04	.09
Report	.08	.14	.08	.05	.07	.17	.02	.10	.13	.09	.06	.11
Self	.08	.12	.06	.08	.06	.02	.11	.07	.06	.15	.00	.00
Combined	.16	.19	.12	.13	.16	.19	.13	.14	.14	.20	.18	.15

.

Table 5. Operational validities - criterion corrected for unreliability in criterion using ICC1 intra-class correlations

	ISMSF	CFMSF	CXMSF	EMMSF	TWMSF	DPMSF	INMSF	BCMSF	PRMSF	POMSF	CIMSF	CUMSF
ISSJT	**.25**	.02	-.07	.12	.08	-.04	-.12	.07	.21	-.05	.19	-.10
CFSJT	.11	**.35**	.25	.22	-.01	.10	.08	-.01	.05	-.01	.04	.22
CXSJT	-.17	-.10	**.23**	.08	.04	-.16	.06	.02	-.06	.00	-.06	-.04
EMSJT	-.03	.03	.08	**.22**	-.02	.06	-.19	-.01	.03	-.19	-.13	.00
TWSJT	-.06	.02	.13	.17	**.29**	.10	.15	-.10	.07	-.04	.04	-.06
DPSJT	.05	.01	.30	.26	.12	**.28**	.09	.19	-.02	-.04	.07	.10
INSJT	.00	.11	.21	.06	.08	-.05	**.20**	.06	.08	.12	.00	.25
BCSJT	.12	-.06	-.05	.14	.38	.22	.16	**.20**	.11	.24	.35	.13
PRSJT	.12	-.09	.15	.26	.15	.03	-.18	.06	**.22**	.06	-.03	-.11
POSJT	.01	.12	.14	.37	.23	.12	.10	.21	.23	**.32**	.08	.11
CISJT	.13	.01	.08	.16	.25	.20	.23	.15	.05	.28	**.30**	.30
CUSJT	-.03	-.14	-.10	.07	.03	-.28	.11	-.10	.03	.03	-.12	**.23**

Appendix A.
Sample scenario and response option for IS.

You are in charge of finances for an aid organisation. You and the CEO have recently been forwarded the accounts report to be reviewed at the next board meeting. You are seriously concerned over the contents of this report. In particular, it shows a downturn in fundraised income due to press reports that some of your funds in Sri Lanka may have been used to fund a terrorist activity. No evidence has yet been found of a direct link although investigations are being undertaken. The financial summary shows a predicted deficit of almost $18m, and the balance sheet shows alarmingly low levels of cash. The situation comes as a surprise, as you and your team had expected a less dramatic summary of the financial problems based on recent developments. Due to the perceived crisis, unless you advise otherwise, the CEO will go public with these results. What do you do?

a) Advise the CEO to hold off going public until you can get more facts. Reassure him that you intend to get more information and that to make a public statement now could be very damaging.
b) Issue a press release to the media letting them know that the allegations are untrue and restating that you do not fund terrorist activities. Contact your biggest contributors/supporters, via letter, to reassure them. Initiate a fund raising campaign and get an expert organisation to come in and support you on this.
c) Get an update on the investigation and double check the figures for the financial summary to make sure the CEO has the most up to date information to make an informed decision. Support whatever course of action the CEO decides on.
d) Call a meeting of your staff and get all perspectives on why things have reached this point without you and your team being aware. Contact the people carrying out the investigation to check progress.