Abstract

Previous research has examined how the method used to score a situational judgment test (SJT) affects the validity of the SJT (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). We seek to add a new type of SJT scoring approach referred to as item response theory. To this end, we compared the summed score approach of scoring SJTs with item response theory and multivariate items response theory in the prediction of job performance. Results from two samples show that both item response theory and multivariate item response theory have promise for scoring SJTs and may increase the validity of SJTs as a single predictor of job performance and in the presence of general mental ability (GMA) and personality. However, issues concerning the factor structure of SJTs may affect the viability of some item response theory models.

SITUATIONAL JUDGMENT TEST VALIDITY: A MATTER OF APPROACH

The process of staffing refers to attracting, selecting, and retaining individuals to work in organizations (Ployhart, 2006). The second step of staffing, selecting, is also referred to as personnel selection and can include employment testing which provides job relevant information about job applicants to the potential employers in an attempt to help employers make better hiring decisions (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). Employment testing and selection methods can influence multiple levels of the organization. At the individual level, selection methods help organizations hire more effective employees (e.g., job performance) (Schmidt & Hunter, 1998). Personnel selection also affects both group- and organizational- level outcomes via the accumulation of knowledge, skills, ability, and other characteristics (KSAOs) within an organization. The resource based view the firm has been used to explain why valid selection methods can provide sustained competitive advantages relative to organizations with less valid selection methods as a result of more effective employees and the accumulation of KSAO's within the organization (Ployhart & Weekley, 2010; Nyberg, Moliterno, Hale, & Lepak, 2014).

Situational judgment tests (SJTs) measure an applicant's judgment regarding work-related situations (Weekley & Ployhart, 2005). Research concerning the testing of an individual's judgment has been around for almost a century (Moss, 1926), but the more recent incarnations of SJTs relate to work concerning Motowidlo, Dunnette, and Carter (1990) who used the tenet of behavioral consistency to explain why presenting individuals with low-fidelity approximations of work situations can predict subsequent job performance. More specifically, that how a person responds in low-fidelity simulations of the work environment may mirror how a person will react when presented with similar situations in an actual work environment.

Relative to other selection tests, SJTs have several documented advantages including: lower mean demographic differences relative to cognitive ability tests, better applicant reactions, inexpensive administration costs, and useful levels of validity (Chan & Schmitt, 1997; McDaniel et al., 2001, McDaniel et al., 2007; Whetzel & McDaniel, 2009; McDaniel, Psotka, Legree, Yost, & Weekley, 2011). Validity is the most important component that organizations and researchers will consider in the implementation of a selection method. As such, the ability to increase the validity of a selection method can offer substantial benefits to organizations in terms of the their selection practices and the host of benefits associated with valid selection procedures at the individual, group, and organizational levels (Nyberg et al., 2014; Ployhart & Weekley, 2010).

To that end, we seeks to compare the validity of a current scoring method used in SJTs, summed score, with that of IRT and MIRT in an attempt to understand if there are differences in the validity of scoring methods as well as what the benefit maybe of using a multivariate approach to scoring SJTs.

Current SJT Scoring Approaches

Approaches to scoring SJTs have been limited because SJTs do not have objectively correct responses and, in many cases, multiple answers could be perceived as correct. Unlike an addition problem (5 + 5 = ?), choosing between behavioral alternatives will involve aspects of individual differences (e.g. personality) and the environment from which the situation arose. Accordingly, there does not seem to be an unambiguously correct method to score SJTs. As such, there have been a number of methods used to score SJTs including: summed score, distance scores, and correlation-based approaches. Of these scoring approaches, the only one appropriate for scoring SJT stems where respondents are asked to choose between behavioral alternatives is the summed score approach, as the other methods require a Likert-type response

indicating either effectiveness or willingness to engage in a particular behavior. Though some research has focused on scoring Likert-type response formats has occurred (McDaniel et al., 2011; Legree, Kilcullen, Psotka, Putka, & Ginter, 2010), little research has examined methods for scoring SJTs where respondents are asked to choose between behavioral alternatives.

Item Response Theory

A new approach to scoring SJTs could be item response theory (IRT). IRT is a collection of mathematical models and statistical item analyses used in test scoring and has been referred to as *modern psychometrics* because it has replaced classical test theory in several substantive research areas (Thissen & Steinberg, 2009; Morizot, Ainsworth, & Reise, 2006). An area where IRT has been especially influential is the education testing where it is used to both score and as a screening tool for potential questions on the Graduate Record Exam (Roid, 2003). IRT has even been referred to as the most important statistical method about which researchers know nothing (Kenny, 2009). This lack of understanding is generally reflected in the organizational sciences; however, some research has used IRT to explore issues, such as: job attitudes, personality, GMA, performance ratings, vocational interests, and employee opinions (Carter, Dalal, Lake, Lin, & Zickar, 2011).

IRT models use item and person characteristics to estimate the relation between a person's underlying trait level (e.g., GMA) and the probability of endorsing an item (LaHuis, Clark, & O'Brien, 2011). Different IRT models assess different parameters and varying levels of complexity. The one parameter model of IRT assesses an item's difficulty, as represented by b (beta). The b parameter is the location, where location is defined as how much of latent trait θ (theta) would be needed to have a 0.50 probability of endorsing the correct item. Latent traits underlie and cause behavior but are not directly observable. A two parameter model includes

both the *b* parameter and the *a* parameter, which measures an item's ability to differentiate between individuals on the latent trait. IRT models may include other parameters. For example, the three-parameter logistic models can be used to assess a parameter related to guessing referred to as *c* (Zu & Kyllonen, 2012).

As a result of the parameters that IRT models can estimate, IRT does not provide a simple summed score; rather, the process of scoring tests using IRT models is called *pattern scoring* because "different patterns of responses to the set of items on a test gives different scores, even when the number of correct responses is the same" (Reckase, 2009, p. 62).

All IRT models are different equations for modeling the shape and location of the item response function (IRF; Morizot et al., 2007). The IRT literature uses the terms item characteristics curve (ICC) and IRF interchangeably (Ayala, 2009). Figure 1 is an IRF that illustrates a logistic ogive, such as IRT estimates. A person with a standardized θ of zero has a 0.50 probability of getting the item in Figure 1 correct. Figure 1 shows that *b* is typically measured on a standardized scale between -3 and +3. As the item difficulty increases, *b* also increases. The parameter *a* is represented by the slope of the logistic ogive. As *a* increases, the model is increasingly able to discern between people with a small relative difference in the underlying latent trait.

Insert Figure 1 about here

The three main assumptions of the IRT models are monotonicity, unidimensionality, and local independence (Ayala, 2009). In IRT, monotinicity means that as the latent trait increases, the probability that a respondent will endorse the correct item also increases. The unidimensionality assumption is met when the responses are a manifestation of only a single

latent trait θ (Ayala, 2009). Tests for dimensionality can include factor analysis or its derivative scree plot (Reckase, 1979; Drasgow & Parsons, 1983). Reckase (1979) suggested that having a single dominant factor that explains significantly more than any other single factor will ensure that the IRT unidimensional model can be applied to heterogeneous datasets. Though guidelines for a single dominant factor are subjective, the first factor in Reckase's (1979) dataset accounted for 20% of the variance, which was robust against violations of unidimensionality. Drasgow and Parsons (1983) underscored that the single dominant factor must be prepotent, referring to a single factor that accounts for more variance than each of the other individual factors. In cases where a prepotent factor does not exist, IRT item-level characteristics can be significantly biased (Reckase, 1979; Drasgow & Parsons, 1983).

A concept related to unidimensionality is that of local independence, also referred to as conditional independence (McDonald, 1999). Local independence implies that how an individual responds to a single item depends only upon that individual's location on θ. After the latent trait is controlled for, local independence is proven when items have no remaining significant correlations. If correlations exist, then the dataset is locally dependent and does not meet the assumptions of the IRT model. As is the case with the unidimensionality assumption, violations of local independence can result in biased parameters, model misfit, and an overestimation of model validity (Reckase, 2009).

Although the unidimensionality assumption and local independence are related, some cases may meet the unidimensionality assumption but have a locally dependent scale. For example, researchers have found cases in which scale construction influenced respondent choices independent of the scale content (e.g., ordering bias or common method variance). In cases where multiple items in a scale suffer from the same bias, response patterns may demonstrate a

correlation independent of the items being measured (Goldstein, 1980; Schwarz, 2000). Thus, violations of unidimensionality can cause violations of the local independence assumption, but violations of the local independence assumption do not necessarily imply multidimensionality.

IRT has five distinct advantages relative to classical test theory. First, IRT scales are item invariant (Ayala, 2009). If two different tests are presumed to measure the same latent construct, an individual's latent trait scores are directly comparable across tests because of the information available about the item and scale level via the SIF and IIF. As a result, the information contained within the items, and not the actual items, matters to researchers. In contrast, Classical Test Theory (CTT) scores are not directly comparable across scales, even if the scales purport to measure the same construct, because no mechanism measures item information and compares the two scales in terms of the SIF (Reckase, 2009).

The second advantage is that IRT examines an individual's response pattern to measure an individual's relative place on a latent trait. As noted, IRT does not assume that all items are equally difficult, and it incorporates other parameters of scale items into the respondents' score (Reckase, 2009). This is in contrast to CTT, in which the sum of the raw item scores is the total test score (Warne, McKyer, & Smith, 2012). Because IRT examines the pattern of responses within the item to assess multiple parameters, it can provide a better estimate of an individual's latent trait.

The third and fourth advantages of IRT relative to CTT relate to the characteristics of the response options. The third advantage is that some IRT methods cater well to polytomous response options without an objectively correct response (Kenny, 2009). Fourth, IRT response formats may be better suited to ambiguous, difficult-to-interpret responses (Zu & Kyllonen, 2012).

The fifth advantage is that the scale information function (SIF) allows researchers and employers to better understand whom they are differentiating between. Specifically, they can determine whether their selection tool differentiates equally well across varying levels of an important latent construct or only between specific levels of a latent construct. Consequently, depending on an employer's goals, items with particular characteristics can be added or removed to find new employees with specific level of some latent trait.

Several of these advantages are pertinent to SJT scoring. IRT can examine unordered polytomous data with ambiguously correct items, which most researchers agree is a characteristic of SJT items (McDaniel & Whetzel, 2005; Bergman et al., 2006). Second, some IRT models may be particularly well suited for data in which identifying *a priori* the best response is difficult (Kenny, 2009). One concern with SJTs is that responses with averages in the mid-range of a Likert-type scale do not add, and may reduce, test validity (McDaniel et al., 2011). In addition, tests that instruct respondents to choose best and worst options can have high degrees of variance for some items, which can indicate item ambiguity and can reduce the validity of SJTs. It has been shown that some IRT models may cater well to this type of data (Zu & Kyllonen, 2012). Finally, pattern scoring methods are particularly important on SJTs because although individuals' mean scores may be similar, the responses they select could be very different.

A clear drawback of IRT models is that they are assumed to measure a single and continuous latent construct. SJTs likely violate this assumption given that previous research has indicated that SJTs measure a number of constructs, including GMA and personality (McDaniel et al., 2011; McDaniel & Whetzel, 2005). However, researchers have argued that the unidimensionality assumption is not realistic when attempting to measure multidimensional

items and scales (Drasgow & Parsons, 1983; Morizot et al., 2007). In situations where constructs are multidimensional, researchers need to ensure that the traits being measured are sufficiently unidimensional to produce unbiased and accurate item parameters (Morizot et al., 2007). Some research suggests that IRT is able to withstand some departures in unidimensionality (Reckase, 1979; Drasgow & Parsons, 1983). However, even in situations where the unidimensionality assumption is not met, multivariate item response theory (MIRT) can be used. MIRT is specifically used in situations with multivariate scales and items (Reckase, 2009; Wright, 2013).

Researchers have created a number of IRT models, many with different theoretical underpinnings, to handle unique response formats and scoring methods (Ostini & Nering, 2006). Most IRT models can use the same types of keys used by more traditional SJT scoring methods (e.g., consensus). Six models that are relevant to SJT scoring and MIRT are the nominal response model (NRM); two-parameter logistic model (2PL); three-parameter logistic model (3PL); and two MIRT models, the M2PL and M3PL.

Bock (1972) designed the NRM to score polytomous response formats (Ostini & Nering, 2006). The model assumes that a continuous latent variable accounts for all the covariance among unordered items. Research has used the NRM when item responses are unordered (Zu & Kyllonen, 2012), such as when responses do not have a pre-established correct order. Importantly, the NRM does not require a scoring key (Zu & Kyllonen, 2012). Although responses are clearly more, or less, correct relative to the other multiple choice options, the model estimates the correct response based on the items' relation with θ. As such, a crucial advantage, and perhaps purpose, of the NRM model is that it finds implicit ordering in unordered categorical data, such as data from SJTs (Samejima, 1972).

The one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL), which have been used to score SJTs, are based on logistic regression. These IRT models can accommodate dichotomous response formats or data that can be coded into dichotomous response formats. The 1PL, 2PL, and 3PL describe the probability of a correct response to a stem as a function of a stem characteristic (e.g., ambiguity) and the respondent's latent ability (Zu & Kyllonen, 2012). The 1PL model is the least complex model. This binary model, only includes *b,* the parameter that measures item difficulty. The 2PL model captures variance from item difficulty and *a,* an item's ability to discriminate among test takers. Finally, the 3PL model incorporates the item's difficulty (*b*), the item's ability to discriminate among respondents (*a*), and a parameter that measures the respondent guessing the best item (*c*).

Multidimensional item response theory (MIRT) may overcome the unidimensionality assumption associated with IRT while providing some of the same benefits of IRT. MIRT has been viewed as a special case of factor analysis and structural modeling. It comprises a family of models designed to determine the stable features of individuals and items that influence responses across dimensions (Reckase, 1997; Reckase, 2009). Like IRT, MIRT assumes that the measured construct relations are monotonically increasing. MIRT models include those that require simple structure (between-item dimensionality, or within-item unidimensionality) and those with multidimensional items (within-item multidimensionality). Models requiring simple structure assume that different items measure different dimensions but that each item measures only one dimension. Within-item multidimensionality assumes that a single item may measure multiple constructs.

In general, MIRT captures much of the same item information as IRT, though it does so in multidimensional space. As such, the MIRT θ is a vector that measures multiple elements

(Wright, 2013). In addition, MIRT captures a *d*, which measures item difficulty in multidimensional space. Given the multidimensional aspect of *d*, it is not directly equivalent to item location parameter *b* in IRT. Unlike *b*, *d* could include multiple locations for the same level of difficulty. Compensatory models estimate the *a* parameter for each latent trait that the item is assumed to measure. The *a* in MIRT models has a similar interpretation in unidimensional IRT as the ability of an item to discern between respondents; however, in MIRT models, *a* is measured for each latent trait. For example, an item measuring two dimensions will have an associated a_1 and a_2.

MIRT models are either compensatory or non-compensatory. MIRT models that assume between-item multidimensionality are non-compensatory because high scores on one dimension do not compensate for lower scores on another dimension (Reckase, 2009). Non-compensatory MIRT models are assumed to have items that measure only a single underlying θ, which is unlikely with SJTs (McDaniel & Whetzel, 2005).

In contrast, within-item multidimensionality models are compensatory because a high score on one dimension can compensate for a low score on another dimension (Reckase, 2009). Multidimensional models are most appropriate for SJTs where individual items can measure multiple latent constructs because research has shown that individuals employ all of their faculties when responding to a question, such as personality and intelligence (Motowidlo & Buyse, 2010).

Examples of compensatory models include the M2PL and M3PL.The M2PL and M3PL vary in the parameters that are estimated but are both used to score dichotomous responses. The M2PL model estimates *d (difficulty)* and a *(discrimination)*, with items having an estimated *a* for each underlying dimension. The M3PL also estimates a guessing parameter, *c*, to account for the

observation that respondents may correctly answer a question that should require higher levels of θ (Reckase, 2009; Lord, 1980).

Research into SJT Scoring using IRT. Unlike more traditional methods of scoring SJTs, research into scoring SJTs using IRT is limited. Zu and Kyllonen (2012) compared scoring methods based on more traditional methods and item response theory in their evaluation of SJTs that measured emotional management in youths and student teamwork/collaboration. The authors used two classical test scoring methods: first, the number of correct questions, based on an expert key, and second, a partial credit model with keys based on respondent consensus and partial scores based on the proportion of the sample choosing that item. They used the IRT methods of NRM, generalized partial credit model (GPCM), 1PL, 2PL, and 3PL. In their first study of emotional management in youths, Zu and Kyllonen found that the NRM, on average, predicted outcome variables better than other scoring methods, but their second study did not find that any method was clearly more effective. Thus, the authors used two different SJTs in two different samples and found two different results, leading to inconclusive findings regarding the effectiveness of the various scoring methods. To understand the results, they conducted a secondary analysis of the items and found that the test used in the first study included more ambiguous items than the test in the second study. They concluded that in cases where item ambiguity is high, NRM may be the more appropriate scoring method.

Wright (2013) examined the dimensionality of SJTs using MIRT and factor analysis. Wright first used an exploratory factor analysis and confirmatory factor analysis (CFA) to form four factors of the SJT and then used MIRT to derive another three factors, for a total of seven SJT factor scores. In addition, Wright calculated a total SJT score as a summed score of the number correct in the SJT. Through these analyses, Wright attempted to predict job performance

in the presence of personality and a measure approximating GMA. Because the study assessed the SJT as one having multiple factors, Wright did not report a correlation between an overall SJT score and the criterion variable of supervisor-rated job performance; however, the researcher performed a hierarchical regression to predict job performance by entering the overall SJT score, CFA-derived factors, and MIRT-derived factors into a regression equation that included personality and an approximation of GMA. The overall SJT score increased the *R*-squared value from 0.14 to 0.23, the addition of the four CFA derived factors increased the *R*-squared value to 0.27, and the addition of the three MIRT-derived factors increased the *R*-squared value to 0.41. These results showed that each of the SJT scores, though correlated with one another, added incremental validity to the prediction of job performance.

Both Zu and Kyllonen (2012) and Wright (2013) found higher levels of validity using their respective IRT and MIRT models than with other methods for scoring SJTs. As such, IRT and MIRT show promise for scoring SJTs with higher levels of validity and provide practitioners with an efficient method to increase the validity of their selection tools.

Hypotheses

Research has shown that different scoring methods may influence the validity of the SJT in predicting job performance (Legree et al., 2010; Zu & Kyllonen, 2012; Bergman et al., 2006). The traditional summed score method will be compared to the NRM, 2PL, 3PL, M2PL, and M3PL models based on a single key developed by subject matter experts and we posit:

H1a: The method of scoring the SJT will influence the SJT validity in the prediction of job performance.

The aforementioned advantages associated with IRT and MIRT, including the ability to estimate item parameters in the scoring of individuals as well as pattern scoring, should be a better measurement approach as compared to the summed score. As such, we offer the following:

H1b: IRT and MIRT methods will yield higher validity than the summed score approach.

Research has shown that SJTs are multidimensional (McDaniel, et al., 2011) and that they measure aspects of personality, intelligence, and other constructs (as illustrated by incremental validity). As such, taking a multidimensional approach to scoring SJTs may allow the researcher to capture factors that measure criterion relevant traits, which could in turn increase the validity of SJTs. Consequently, we offer the following:

H1c: MIRT will provide higher levels of criterion-related validity than other methods of IRT scoring.

In line with my previous hypothesis concerning IRT and MIRT as potentially better scoring methods and that we have no reason to believe that IRT methods will increase multicollinearity with existing personnel selection methods, we offer the following:

H2a: IRT methods will yield greater incremental validity over and above personality and GMA relative to the summed score approach.

Again, in line with the observation the SJTs are multidimensional, capturing criterion relevant variance associated with the different factors that SJTs produce should increase the incremental validity and thus we offer the following:

H2b: MIRT methods will yield greater incremental validity over and above personality and GMA relative to IRT methods.

Methods

Sample. We have two samples that were both collected for a concurrent validation of an SJT used in the selection of retail managers. Sample 1 consists of 1,859 employees and Sample 2 consists of 1,094 employees.

Procedure. Respondents were asked to complete the survey while at work, using company time. Management support was communicated, but respondents were given the option to participate. The survey took approximately 3 hours to complete and respondent anonymity was communicated prior to responding to the survey.

Measures. The SJT for Sample 1 included 77 items in the pick best and pick worst format while the SJT for Sample 2 included 19 pick best items. The test formats were paper-and-pencil. For each stem the respondents were presented with a situation and several potential responses to the situation described in the stem.

The data set also contains cognitive ability and the Big 5 personality scales. Cognitive ability was measured using a 90-item cognitive ability test. This 90-item test included mathematical reasoning, verbal reasoning, mathematical equations, and vocabulary questions. Each multiple choice question had five possible responses to choose from.

Personality was measured with five separate 25-item scales each designed to capture each of the big five dimensions. In total, there were 125-items to measure conscientiousness, openness, agreeableness, neuroticism, and extroversion. Though the personality scale is proprietary, it has been used in previously published research (Weekley and Ployhart, 2005).

The criterion consists of a measure that researchers created based on job specific tasks taken from existing job descriptions, training manuals, policy manuals, and other sources of job documentation. These tasks were rated on a five-point scale by the respondent's manager. The ratings were averaged to create a measure of task performance.

Analyses. Initial scoring procedures comprised the traditional summed score approach as well as IRT and MIRT models, including: NRM, 2PL, 3PL, M2PL, and M3PL. Both the M2PL and M3PL will be used to score the two dimensional and three dimensional models. These dimensions were derived from previous research (Wright, 2013) and factor loadings were assessed using exploratory factor analyses with a varimax rotated solution.

Statistical Software. Regression analyses were performed in SPSS and all IRT and MIRT models were estimated using flexMIRT software (Cai, 2013).

Item Fit. flexMIRT also provides information concerning item fit. IRT item level estimates are made using marginal chi-square values and the standardized local dependence (LD) chi-square matrix (Cai, Thissen, & du Toit, 2011). Each item has a single marginal chi-square value associated it. Higher chi-square values establish that items are deviating from the chosen IRT or MIRT model. The standardized LD chi-square value is estimated as the relation between each item included in the model. Similarly, higher standardized LD chi-square values indicate a potential violation of the local dependence assumption.

Results

Before scoring the SJTs using IRT, we ran models to assess the adequacy of the base IRT models for use in scoring the SJT. To accomplish this, we measured the fit of the 2PL model using the RMSEA. The RMSEA was 0.03 in both Sample 1 and 2 indicating good fit for the model. The best practice in IRT is to use the RMSEA to establish that the data fit the IRT assumptions and then use comparative fit metrics (e.g. the Akaike Information Criterion and Bayesian Information Criterion) to evaluate the relative fit of the models with respect to one another. Despite the initial satisfactory RMSEA, the 3PL model in Sample 2 showed significant departures from the unidimensionality and local independence assumptions of the IRT model.

The item level standardized LD chi-square values and the marginal chi-square values are reported in Appendix A. Items 1, 2, 9, and 11 were the cause of the departures from the model assumptions as illustrated by the high chi-square values and the high standardized LD chi-square values. Consequently, two sets of analyses were run for Sample 2. One set of analyses included all 19-items while the second set of analyses removed the items causing violations of the model assumptions resulting in a 15-item SJT measure.

Hypothesis 1a through 1c are related to the validity of the SJTs in the absence of GMA and personality. The results of the first hypotheses are reported in Table 1. Hypothesis 1a stated that the method used to score the SJT would influence the criterion validity. In support of hypothesis 1a, the different methods of scoring produced varying levels of validity. In Sample 1, these validities range from 0.111 to 0.141. The 19-item Sample 2 validity ranged from 0.031 to 0.134, while the 15-item Sample 2 validity ranged from 0.023 to 0.134.

Insert Table 1 about here

Hypothesis 1b related to the comparison between the summed score, IRT scoring methods, and the MIRT scoring methods stating that the IRT and MIRT methods would have a higher level of validity compared to the summed score. Hypothesis 1b received partial support. In Sample 1, the 2PL, 3PL, and NRM models displayed higher levels of validity compared to the summed score, but the balance of the models did not. In 19-item Sample 2, the multivariate models showed greater validity than the summed score, but the univariate models did not. In 15-item Sample 2, there were three models that showed higher levels of validity than the summed score. First, both 3-dimensional multivariate models showed greater validity (e.g. the M2PL and

M3PL) than the other scoring methods. The 3PL model also produced higher levels of validity than the summed score. However, the balance of the models yielded lower levels of criterion validity than the summed score. Thus, across the two groups there was a discrepancy in which scoring method showed the highest level of validity; however, both groups provided scoring methods that had higher levels of validity than the summed SJT score.

Hypothesis 1c stated that MIRT methods of scoring would provide higher levels of criterion related validity than IRT methods of scoring. There is mixed support for hypothesis 1c. In Sample 1, the MIRT methods did not provide higher levels of validity despite the multidimensional models having multiple factors to explain variance in the criterion variable and the multidimensional models showing better fit. Further, the results from Sample 1 showed that the NRM model had a higher level of criterion validity than the other methods. This is particularly interesting because the NRM does not have a presupposed ordering of incorrect and correct items, thus the model is calculating the rightness and wrongness of each response based on response patterns and the unidimensional trait they are assumed to measure. In the 19-item Sample 2, all of the multidimensional models provide higher levels of criterion related validity than any of the unidimensional models. Further, the unidimensional models in Sample 2 show fairly low levels of validity when compared with the summed score and the Sample 1 unidimensional models. However, this was due to the departure from unidimensionality that made the use of unidimensional IRT models less appropriate. The 15-item Sample 2 SJT unidimensional items contained higher levels of validity relative to the 19-Item Sample 2 with the 3PL model containing the highest level of validity amongst the unidimensional models. Further, the unidimensional 3PL model displayed a higher level of validity than the 2-dimensional MIRT models. However, the model providing the most validity in 15-item Sample 2

was again the M3PL 3-dimensional model. Thus, both Sample 2 analyses had the highest amount of validity using the 3-dimensional M3PL model. Note that the NRM was the most valid scoring method in Sample 1, but was the least predictive method across both sets of Sample 2 models.

Hypotheses 2a and 2b dealt with the incremental validity of the SJT scoring method over and above personality and GMA. Table 2 displays the results concerning this incremental validity for Sample 1, 19-item Sample 2, and 15-item Sample 2, respectively. Column 1 contains the scoring method used. Column 2, column 3, and column 4 contain the model R at each step of the regression. Step 1 included only GMA, step 2 included GMA and personality, and step 3 included GMA, personality, and the SJT score(s). Column 5 contains the absolute change in the model R from step 2 to step 3, column 6 displays the significance of the change going from step 2 to step 3, and column 7 contains the percentage change in model R from step 2 to step 3.

Insert Table 2 about here

Hypothesis 2a stated that IRT methods would produce higher levels of validity than the summed score approach that has more traditionally been used in SJT scoring. There is mixed support for this hypothesis. In Sample 1, the unidimensional IRT models produced higher levels of incremental validity than the summed score, but the MIRT models showed lower levels of incremental validity than the summed score. In the 19-item Sample 2, all of the multidimensional models showed greater incremental validity relative to the summed score, but none of the unidimensional models did. In the 15-item Sample 2 the 3PL, M2PL with 3-dimensions, and M3PL with 3-dimensions had higher levels of incremental validity than the summed score, but the balance of the models did not. Thus, all groups contained models that provided more

incremental validity than the summed score, but those models differed across the groups and not all models showed greater incremental validity than the summed score.

Hypothesis 2b stated that MIRT models would show greater incremental validity than the unidimensional IRT models. Again, there was mixed support for this hypothesis. In Sample 1, the NRM, which produced the largest amount validity, also explained the largest amount of incremental variance (+12.4%) while GMA and personality were included in the model. In the 19-item Sample 2, the M3PL model with three dimensions explained more variance in the model above personality and GMA (+50.0%) relative to the M2PL with three dimensions (+49.1%). All of the multidimensional models in 19-item Sample 2 showed higher levels of validity than the summed score or unidimensional IRT models. In 15-item Sample 2, the 3-dimensional models showed greater incremental variance explained than the unidimensional models, but the 2-dimensional models did not. Summarizing the results from hypothesis 2c, 19-item Sample 2 supported hypothesis 2c such that all MIRT models added more incremental variance than the unidimensional IRT models. 15-item Sample 2 found partial support for hypothesis 2b such that the 3-dimensional models showed greater incremental validity than the unidimensional models; however, the 2-dimensional models did not. Finally, Sample 1 failed to support hypothesis 2b such that the MIRT models did not provide more incremental variance than the unidimensional IRT models.

Discussion

Both Sample 1 and Sample 2 mirrored earlier research performed using IRT and MIRT for use in SJT scoring methods. Previous research examined the validity of unidimensional IRT by using the NRM, generalized partial credit model, 1PL, 2PL, and 3PL models to score two different SJTs. This research suggested that the NRM may offer the greatest

amount of validity among the unidimensional models (Zu and Kyllonen, 2012). Our Sample 1 results corroborate previous research, with the NRM providing higher levels of criterion validity than any of the other models tested. It is important to note that the previous research was comparing only unidimensional IRT models. That we found support for the use of the NRM in comparison with other unidimensional IRT models and MIRT models provides more evidence that the NRM may be a fruitful scoring method meriting further research.

Beyond the empirical results there are several advantages of using the NRM. The first of which is that the model does not require a scoring key indicating the correctness of a particular item. Further, it can handle the polytomous response formats often used in SJTs. Finally, the NRM does not assume that there is an objectively correct response and instead the model will order responses with respect to θ, which is not necessarily a linear scale from correct to incorrect.

Other research has found that SJTs have multivariate structure and using multivariate methods to score SJTs could have benefits in terms of validity (Wright, 2013). Both 19-item Sample 2 and 15-item Sample 2 corroborate this research. Previous research has shown that SJTs measure multiple constructs (McDaniel et al., 2007; Motowidlo & Beier, 2010; Christian, Edwards, & Bradley, 2010), but it is important to note that the factor structure will be dependent on the datasets and previous research has suggested that the stability of these factor structures is questionable across samples and can capitalize on nuances specific to individual datasets (McDaniel & Whetzel, 2005). This research chose to test 2-dimensional and 3-dimensional models and consequently the EFAs were constrained to measure only 2- and 3-factors. These numbers were based on both previous research (Wright, 2013) and on the observation that SJTs

measure GMA, personality, and another factor, illustrated by incremental variance, which could be referred to by a number of names (e.g. judgment).

Several limitations of this research need to be mentioned. First, a concern in SJT research is the use of incumbents as opposed to applicant samples. As mentioned, incumbent samples tend to suffer from range restriction. In addition, they may also differ in terms of test-taker motivation on the SJT (Tay & Drasgow, 2012). A distinction has been made between high-stakes and low-stakes SJTs such that incumbents are taking the SJT as a low-stakes test (i.e. they don't have anything to gain or lose based on their test performance) and applicants are taking the SJT as a high-stakes test (i.e. they may stand to gain or lose the job based on their test performance) (Lievens, Sackett, & Buyse, 2009). The motivation to do well on the SJT would clearly differ between these two groups. Though we don't expect that the objective comparison of scoring methods would change across sample types, we cannot unequivocally say that the additional variance explained by the IRT and MIRT methods would replicate to applicant samples.

Another clear limitation of this research is the varying factor structure across samples which caused disparate results in the analyses. More specifically, that Sample 1 items fit the unidimensionality assumption whereas Sample 2 did not. Because the factor structure of SJTs has been found to vary by SJT and dataset, using MIRT may pose significant challenges to practice, particularly because what the different factors are measuring are, as yet, unidentified. Future use of factor based SJT scoring approaches will be contingent upon understanding factor structure or, at the very least, having a consistent factor structure across samples. Consequently, until research has shown what constructs the factors are measuring and can provide evidence of convergent and divergent validity, factor based scoring approaches in SJTs will be limited.

In summary, increasing the validity of personnel selection may offer organizations a path to sustained competitive advantage and may influence variables of interest at multiple levels of the organization (Ployhart & Weekley, 2010; Nyberg et al., 2014). At the individual level, better personnel selection methods offer organizations a path to hiring more effective employees. At the group and organizational levels, strategic human resource management has noted that KSAOs can accumulate within groups making the group, and ultimately the organization, more effective. Consequently, the pursuit of more valid selection methods is of strong importance to the personnel selection field. To that end, this research examined IRT and MIRT methods for use in scoring SJTs in comparison with the summed score approach. Overall, this paper provided support for using IRT and MIRT methods to score SJTs in personnel selection whether it is as a single personnel selection method or in combination with GMA and personality. However, due to unstable factor structure, we suggest that unidimensional IRT methods be further examined for use in scoring SJTs and that future research needs to be conducted exploring SJT factor structure before MIRT methods can be instituted as a potential replacement for current methods of SJT scoring.

References

Ayala, R.J. (2009). The theory and practice of item response theory. In Little, T.D (Series Ed.) *Methodology in the Social Sciences,* New York, NY: The Guilford Press.

Bergman, M.E., Drasgow, F., Donovan, M.A., Henning, J.B., & Juraska, S.E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223-235.doi: 10.1111/j.1468-2389.2006.00345.x

Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer Software]. Chapel Hill, NC: Vector Psychometric Group.

Carter, N.T., Dalal, D.K., Lake, C.J., Lin, B.C., Zickar, M.J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the job descriptive index. *Organizational Research Methods, 14,* 116-146. doi: 10.117/1094428110363309

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159. doi: 10.1037/0021-9010.82.1.143

Christian, M.S., Edwards, B.D., & Bradley, J.C. (2010). Situation judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63,* 83-117. doi: 10.1111/j.1744-6570.2009.01163.x

Drasgow, F. & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189-199. doi: 10.1177/014662168300700207

Goldstein, H. (1980). Dimensionality, bias independence, and measurement. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246. doi: 10.1111/j.2044-8317.1980.tb00610.x

Kenny, D.A. (2009). Founding Series Editor Note in: Ayala, R.J. (2009). The theory and practice of item response theory. In Little, T.D (Series Ed.) *Methodology in the Social Sciences,* New York, NY: The Guilford Press.

LaHuis, D.M., Clark, P., & O'Brien E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods, 14,* 10-23. doi: 10.1177/1094428109350930

Legree, P.J., Kilcullen, R., Psotka, J., Putka, D., & Ginter, R.N. (2010). Scoring situational judgment tests using profile similarity metrics. United States Army Research Institute for Behavior and Social Sciences. Technical Report 1272. Retrieved from: www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA530091

Lievens, F., Sackett, P.R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94,* 1095-1101. doi: 10.1037/a0014628

Lord, F.M. (1980). *Applications f item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates. McDaniel. M.A., Hartman, N.S., Whetzel, D.L., & Grubb, W.L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91. doi: 10.1111/j.1744-6570.2007.0065.x

McDaniel. M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740. doi: 10.1037//0021-9010.86.4.730

McDaniel, M.A., Hartman, N.S., Whetzel, D.L., Grubb, W.L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91. doi:10.1111/j.1744-6570.2007.00065.x

McDaniel, M.A., Psotka, J., Legree, P.J., Yost, A.P., & Weekley, J.A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96,* 327-336. doi: 10.1037/a0021983

McDaniel, M.A., & Whetzel, D.L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515-525. doi: 10.1016/j.intell.2005.02.001

McDonald, R.P. (1999). *Test theory: A unified approach.* Mahwah, N.J.: Erlbaum.

Moss, F.A. (1926). Do you know how to get along with people? Why some people get ahead in the world while others do not. *Scientific American, 135,* 26-27.

Morizot, J., Ainsworth, A.T., Reise, S.P. (2009). Toward modern psychometrics: Application of item response theory models in personality research. In R.W. Robins, R.C., Fraley, & R.F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York: Guilford, 2007.

Motowidlo, S.J., & Beier, M.E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment tests. *Journal of Applied Psychology, 95,* 321-333. doi: 10.1037/a00117975.

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. doi: 10.1037//0021-9010.75.6.640

Nyberg, A.J., Moliterno, T.P., Hale, D., Lepak, D.P. (2014). Resource-based perspectives on unit-level human capital: A review and integration. *Journal of Management, 40,*316-346. doi: 10.1177/0149206312458703

Ostini, R., & Nering, M.L. (2006). *Polytomous Item Response Theory Models.* Thousand Oaks, CA: Sage Publications.

Ployhart, R.E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management,32,* 868-897. doi: 10.1177/0149206306293625

Ployhart, R.E., & Weekley, J. (2010). Strategy, selection, and sustained competitive advantage. In J. Farr & N. Tippins (Eds.),

Reckase, M.D. (2009). *Multidimensional Item Response Theory*. New York, NY: Spring.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-26. doi: 10.1177/0146621697211002

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230. doi: 10.2307/1164671

Roid. G.H. (2003). Stanford-Binet Intelligence Scales, Vol. 5, Technical Report. Itasca, IL: Riverside Publishing.

Samejima, F. (1972). A general model for free response data. (Psychometric Monograph No. 18) Richmond, VA: Psychometric Society.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274. doi: 10.1037//0033-2909.124.2.262

Schwarz, N. (1999). Self-reports: How questions shape the answers. *American Psychologist,* 54, 93-105. doi:10.1037/0003-066X.54.2.93

Tay, L. & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods, 15,* 363-384. doi: 10.1177/1094428112439709

Thissen, D. & Steinberg, L. (2009). Item Response Theory. In R.E. Millsap, & A. Maydeu-Olivares (Eds.), *The Sage Handbook of Quantitative Methods in Psychology,* (pp. 148-177). Los Angeles, CA: Sage.

Warne, R.T., McKyer, E.L.J., Smith, M.L. (2012). An introduction to item response theory for health behavior researchers. *American Journal of Health Behavior, 36*, 31-43. doi: http://dx.doi.org/10.5993/AJHB.36.1.4

Weekley, J.A., & Ployhart, R.E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18,* 81-104. doi: 10.1207/s15327043hup1801_4

Whetzel, D.L, & McDaniel, M.A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188-202. doi: 10.1016/j.hrmr.2009.03.007

Wright, N. (2013). New strategy, old question: Using multidimensional item response theory to examine the construct validity of situational judgment tests. (Doctoral dissertation, North Caroline State University, 2013).

Zu, J., & Kyllonen, P.C. (2012). *Scoring situational judgment tests with item response models.* (Report ETS-2012-0160.R1). Princeton, NJ: Educational Testing Service.

Figure 1. Item Response Function

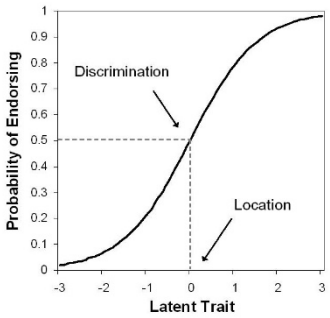

Figure 2. Item Response Function. Reprinted with permission from Morizot, J., Ainsworth, A.T., Reise, S.P. (2009). Toward modern psychometrics: Application of item response theory models in personality research. In R.W. Robins, R.C., Fraley, & R.F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York: Guilford.

Table 1. Multiple R and R-Squared

Sample 1 Multiple R and R-Squared

Method	Multiple R	R-Squared
Summed Score	0.121	0.015
2PL	0.125	0.015
3PL	0.135	0.018
NRM	0.141	0.019
M2PL - 2 Dimension	0.115	0.012
M2PL - 3 Dimension	0.117	0.012
M3PL - 2 Dimension	0.120	0.013
M3PL - 3 Dimension	0.111	0.012

Sample 2 (19-Items) Multiple R and R-Squared

Method	Multiple R	R-Squared
Summed Score	0.103	0.010
2PL	0.058	0.003
3PL	0.037	0.000
NRM	0.031	0.000
M2PL - 2 Dimension	0.113	0.011
M2PL - 3 Dimension	0.132	0.015
M3PL - 2 Dimension	0.126	0.014
M3PL - 3 Dimension	0.134	0.018

Sample 2 (15-Item) Multiple R and R-Squared

Method	Multiple R	R-Squared
Summed Score	0.115	0.013
2PL	0.092	0.008
3PL	0.125	0.016
NRM	0.023	0.001
M2PL - 2 Dimension	0.098	0.010
M2PL - 3 Dimension	0.133	0.018
M3PL - 2 Dimension	0.109	0.012
M3PL - 3 Dimension	0.134	0.018

Table 2. Model R

Sample 1 - Model R

Scoring Method	Step 1	Step 2	Step 3	Abs. Chg.	Sig.	% Chg.
Summed Score	0.110	0.169	0.183	0.014	<.01	8.3%
2PL	0.110	0.169	0.183	0.014	<.01	8.3%
3PL	0.110	0.169	0.187	0.018	<.01	10.7%
NRM	0.110	0.169	0.190	0.021	<.01	12.4%
M2PL - 2 Dimension	0.110	0.169	0.179	0.010	<.05	5.9%
M2PL - 3 Dimension	0.110	0.169	0.181	0.012	<.05	7.1%
M3PL - 2 Dimension	0.110	0.169	0.181	0.012	<.05	7.1%
M3PL - 3 Dimension	0.110	0.169	0.178	0.009	>.10	5.3%

Sample 2 (19-Items) - Model R

Scoring Method	Step 1	Step 2	Step 3	Abs. Chg.	Sig.	% Chg.
Summed Score	0.060	0.106	0.135	0.029	<.01	27.4%
2PL	0.060	0.106	0.119	0.013	<.10	12.3%
3PL	0.060	0.106	0.109	0.003	>.10	2.8%
NRM	0.060	0.106	0.111	0.005	>.10	4.7%
M2PL - 2 Dimension	0.060	0.106	0.142	0.036	<.01	34.0%
M2PL - 3 Dimension	0.060	0.106	0.158	0.052	<.01	49.1%
M3PL - 2 Dimension	0.060	0.106	0.152	0.046	<.01	43.4%
M3PL - 3 Dimension	0.060	0.106	0.159	0.053	<.01	50.0%

Sample 2 (15-Items) - Model R

Scoring Method	Step 1	Step 2	Step 3	Abs. Chg.	Sig.	% Chg.
Summed Score	0.060	0.106	0.143	0.037	<.01	34.9%
2PL	0.060	0.106	0.134	0.028	<.01	26.4%
3PL	0.060	0.106	0.151	0.045	<.01	42.5%
NRM	0.060	0.106	0.109	0.003	>.10	2.8%
M2PL - 2 Dimension	0.060	0.106	0.133	0.027	<.05	25.5%
M2PL - 3 Dimension	0.060	0.106	0.158	0.052	<.01	49.1%
M3PL - 2 Dimension	0.060	0.106	0.140	0.034	<.01	32.1%
M3PL - 3 Dimension	0.060	0.106	0.161	0.055	<.01	51.9%

Note: Several relevant tables are omitted from this submission, but are available by request from the author.