ABSTRACT

MORGAN, KYLE EDWARD. Rethinking Intelligence Tests: Using Multidimensional Item Response Theory to Assess Ability. (Under the direction of Dr. Adam Meade).

Cognitive ability tests are frequently employed for the purposes of employee selection. The result of many of these assessments is a single score that describes a candidate's aptitude. Under classical test theory (CTT), this score is the number of questions to which a candidate responds correctly, while traditional item response theory (IRT) calculates a candidate's ability level. However, by representing an individual with only a single score, these models may be unable to capture the complexity of cognitive abilities. In addition, tests that are heavily weighted toward general mental ability (GMA), or *g*, frequently are faced with the problem of adverse impact in selection settings. The presence of adverse impact can lead to costly legal problems for an organization, and thus should be minimized. The current study suggests multidimensional item response theory (MIRT) has the potential to improve prediction of job-related outcomes while also minimizing the occurrence of adverse impact. To test this assertion, cognitive ability data from job applicants and incumbents were modelled using CTT, unidimensional IRT, and MIRT. Results showed that analysis using CTT and unidimensional IRT both provided significant criterion-related validity, but demonstrated adverse impact. MIRT, however, exhibited criterion-related validity and little adverse impact.

Rethinking Intelligence Tests: Using Multidimensional Item Response Theory to Assess Ability

by
Kyle Edward Morgan

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2014

APPROVED BY:

_____	_____
Dr. Adam Meade Committee Chair	Dr. Lori Foster Thompson
_____	_____
Dr. Bart Craig	Dr. Samuel B. Pond III

BIOGRAPHY

Kyle Morgan completed his undergraduate studies at Louisiana State University Honors College in Baton Rouge, Louisiana. He graduated with a B.S. in Psychology in May of 2009. He then began pursuit of a Ph.D. in Industrial/Organizational (I/O) psychology at North Carolina State University in Raleigh, NC. He received his M.S. in I/O psychology in 2012

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Rethinking Intelligence Tests: Using Multidimensional Item Response Theory to Assess Ability

Intelligence is one of the most widely studied subjects in psychology, and is also one of the most controversial. Its importance to psychologists is due to the influence that it has on everyday life. Intelligence has been shown to predict many important social outcomes, ranging from performance in school to juvenile delinquency (Moffitt, Gabrielli, Mednick, & Schulsinger, 1981; Neisser et al., 1996). Intelligence is a critical construct to industrial-organizational psychologists specifically because of its use as a predictor of job performance. Despite its importance, some have argued that the field of IO psychology has done little to further explore the construct in recent years (Scherbaum, Goldstein, Yusko, Ryan, & Hanges, 2012). Thus, this study is an answer to that call to find new avenues of research for such a significant topic. Specifically, this study uses multidimensional item response (MIRT) theory to explore whether assessing specific cognitive abilities has the potential to maintain or improve the predictive nature of intelligence tests, while also minimizing adverse impact.

There is some confusion regarding the variety of terms that relate to intelligence. Intelligence, besides being the lay term for general competence or problem-solving ability (Sternberg, 2000), is often a generic term used interchangeably with "cognitive ability," "general mental ability (GMA)," or "*g* factor" when referencing an individual's mental ability (Viswesvaran & Ones, 2002). When used in the context of testing, however, it is important to distinguish between the two. Cognitive abilities are specific mental capacities that an individual possesses that allows the individual to perform a given task. Individuals

have many specific cognitive abilities (e.g., quantitative reasoning, spatial reasoning, etc.), though all these abilities are highly interrelated. This shared variance, thought to be indicative of an overarching ability, is what is referred to as GMA, or the *g* (general) factor (Ones, Dilchert, Viswesvaran, & Salgado, 2010). Most tests are designed to measure specific cognitive abilities or aptitudes, but if items are put together that assess multiple cognitive abilities, then the resultant test is thought to assess the shared variance that is GMA, or "intelligence" (Ones, Dilchert, Viswesvaran, & Salgado, 2010).

Despite the critical importance of intelligence to IO psychology, the field is lagging behind others (e.g., neuropsychology, cognitive psychology, and education) in an attempt to understand the nature of the construct (Scherbaum et al., 2012). The lack of new research is likely due to the success that psychologists have had in using existing tests of intelligence as predictors of job performance. Many psychologists may believe that this success means that the topic of intelligence with regard to selection has been exhausted, or that we know all we need to know (Goldstein, Scherbaum, & Yusko, 2009; Goldstein, Zedeck, & Goldstein, 2002; Murphy, 1996). There is, however, still much to be learned about this construct (Goldstein, Zedeck, & Goldstein, 2002). Research is still needed to further explicate the relationship between job performance and intelligence; for example, determining specifically how intelligence is manifested in the workplace. Researchers should also explore how the relationship between intelligence and job performance changes with the increasing complexity of jobs in the current environment (Scherbaum et al., 2012). A further examination of alternative models may also be warranted. Some researchers, such as

Gardner (1993) and Sternberg (1985) reject the idea of a common *g* factor in favor of models that portray intelligence as a multifaceted construct that is a combination of specific abilities. In addition, despite the ubiquitous nature of cognitive ability testing, there is still much controversy over the nature of intelligence and its use in employee selection, specifically as it relates to racial group differences in test scores, as well as the utility of specific abilities over GMA (Murphy, Cronin, & Tam, 2003).

One area in which the topic of intelligence can be further explored is in the development of cognitive ability measures. Specifically, the development and scoring cognitive ability tests used in employee selection should be revised to more closely reflect the current theories of intelligence (Scherbaum et al., 2012). Even advanced psychometric methods such as unidimensional item response theory (IRT) may be inadequate for modeling cognitive ability test data. This study was designed to extend research on intelligence testing in a selection context by assessing whether a multidimensional IRT approach to the assessment of intelligence could increase criterion-related validity, which would improve the test's prediction of job performance, or minimize adverse impact, which would decrease the likelihood of legal challenges to a selection system, when compared to a unidimensional approach.

The Structure of Intelligence

Most lay persons have some conceptualization of what constitutes intelligence, yet experts differ on what they believe it actually is and how it functions. This has led to the proposal of many competing theories (Bowman, Markham, & Roberts, 2002, Sternberg,

2000). At the basic level, experts agree that intelligence can be defined as "a very general mental capacity that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience" (Gottfredson, 1997, p. 13).

Though the structure of intelligence is still constantly debated, there is a growing consensus among researchers that intelligence is multidimensional, with each individual possessing many cognitive abilities, and that these abilities are arranged hierarchically with a general "*g*" factor influencing all of the second order abilities (Deary, 2000; McGrew, 2009). Arguably, one of the first researchers to draw attention to the structure of intelligence was Charles Spearman (1904). Spearman noticed that there was a positive relationship between several measures of sensory discrimination, and that they were also related to measures of academic achievement. This led him to the conclusion that there was a common intellectual ability, *g*, which accounted for these positive relationships (Brody, 2000). Spearman attributed any variance not accounted for by *g* to *s*, or a second specific source of variance that lies in more specific abilities.

The first attempt to elaborate on the variance due to sources other than *g* came from Louis Thurstone (1931). Thurstone developed multiple-factor analysis by decomposing the variance on a test into several independent factors, on each of which an individual could be scored. This led Thurstone to the conclusion that there were several primary abilities, including verbal comprehension, spatial reasoning, and inductive abilities. Initially, Thurstone's theory was at odds with Spearman's because the idea of *g* would confound the

simple structure required for Thurstone's analysis, and Spearman's method for assessing *g* would be unsuccessful if large amounts of covariance were attributable to factors independent of *g*. Eventually, Raymond Cattell (1941) resolved the discrepancy between these two theories by defining intelligence hierarchically; that is, with *g* as a superordinate factor and correlated specific abilities being subsumed as second-order factors.

The most comprehensive analysis of intelligence as a construct came with Carroll's (1993) work in which he presented the results of his factor analysis of over 450 datasets. Caroll (1993) extended the hierarchical theory of intelligence, proposing that intelligence was structured into three levels, or strati. Carroll's (1993) examination of the structure of intelligence has since been integrated with Cattell's (1963) theory of fluid and crystallized intelligence into what is known as the Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities (McGrew, 2005). The CHC model maintains the three strati proposed by Carroll (1993). At the top stratum is the *g* factor, or what is commonly referred to as GMA. The second stratum is composed of broad abilities, such as fluid intelligence, crystallized intelligence, and processing speed. Finally, the third stratum outlines over 70 narrow abilities (see Figure 1).

Much of the empirical work done to date supports the hierarchical view of cognitive abilities. Tests of cognitive ability have been shown to load not only on *g*, but also to the second-order variables as well (Carroll, 2003; Reeve, 2004). This is to be expected given that *g* is typically measured through the use of test batteries designed such that individuals are required to solve a variety of complex problems requiring the use of second-order

abilities, such as quantitative ability, verbal ability, or spatial reasoning (Ones, Viswesvaran, & Dilchert, 2005; Schmidt & Hunter, 2004).

Intelligence and Job performance

The usefulness of measures of intelligence in predicting performance on the job is, by now, well-established (Schmidt, 2002). The relationship between tests of cognitive ability and job-related outcomes has been documented in reviews that encompass over 1,300 meta-analyses summarizing data from over 22,000 studies incorporating over 5 million individuals (Ones et al., 2010). This wealth of research is likely due to the fact that intelligence is seen as the best predictor of job outcomes (Schmidt & Hunter, 1998). Cognitive-ability tests predict training outcomes in the range of r = .5-.7, and overall job performance in the range of r = .35-.55 (Ones, Viswesvaran, & Dilchert, 2005).

Given the success of GMA in predicting desirable job outcomes, there is belief among some researchers that there is little value in considering specific abilities beyond GMA (Hunter, 1986), or that they, at best, provide moderate incremental validity (Ree & Earles, 1991; Ree, Earles, & Teachout, 1994). This assertion, however, has been challenged by recent meta-analytic research suggesting that there may be merit to reexamining the role of specific abilities. Goertz, Hülsheger, and Maier (2014) demonstrated that specific abilities (e.g., reasoning, numerical facility, verbal comprehension, and memory) predicted training success at a rate comparable to GMA. They also show that while GMA is one of the best predictors of job performance across jobs, specific abilities can provide similar validities in certain occupations. In addition, some researchers argue that the predominance of GMA

over specific abilities in regards to job performance is due to an artifact of analysis. Lang, Kersting, Hülsheger, and Lang (2010) point out that under traditional incremental validity analyses, GMA is listed as the first step in the regression analyses, with specific abilities listed in subsequent steps. This means that all variance shared by GMA and specific abilities is subsumed under GMA. These authors demonstrate that other methodologies, such as relative importance analyses (e.g., LeBreton, Hargis, Griepentrog, Oswald, & Ployhart, 2007) using a nested models design, can demonstrate greater predictive validity for specific abilities than for GMA.

GMA determines a large amount of an individual's specific capabilities; it is, however, difficult to measure directly. As such, intelligence tests seek to measure GMA through the demonstration of second-order abilities (Ones et al., 2010), such as solving a mathematical problem (quantitative reasoning) or rotating an object in space (visual processing). If the second-order abilities on a cognitive ability measure were directly aligned with necessary job skills (such as deductive reasoning), one could infer that these abilities could predict specific aspects of job performance above and beyond GMA, especially given that certain jobs require differing abilities. For example, it would be expected that someone with a high level of quantitative knowledge would perform better in the job of statistician, while someone with a higher level of auditory processing would perform better as a musician. Past research has demonstrated that while GMA is an important predictor across jobs, it does not fully capture the complexities associated with the specific abilities required of different jobs (Gottfredson, 1984; Prediger, 1989).

One of the largest issues plaguing the use of intelligence tests for selection purposes is that of adverse impact. Adverse impact is defined by the *Uniform Guidelines on Employee Selection Procedures* as "a substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group" (Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, & Department of Labor, 1978, Section 16B). In these situations, employers must demonstrate the job-relevance of the selection tool being used. According to *Uniform Guidelines*, adverse impact is said to be present when one group is selected at a rate below 80% of that for another group. In practice, this is typically demonstrated when ethnic or gender minorities are hired or promoted at a rate below that of the majority group (white males). In general, individuals of African descent score approximately 1 standard deviation below whites, while those of Hispanic descent score approximately .7 standard deviation below whites (Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). When used in a selection context, these tests are likely to cause a differential selection rate of minorities, resulting in adverse impact. For this reason, many employers simply forgo the use of ability tests, opting for less-valid tests that may improve the diversity of hiring (Pyburn, Ployhart, & Kravitz, 2008).

One method with potential to mitigate the effects of adverse impact on cognitive testing is to use tests of specific rather than general cognitive abilities. These tests often demonstrate reduced subgroup-differences in scores (Hough, Oswald, & Ployhart, 2001; Jensen, 1987; Kehoe, 2002; Naglieri & Jensen, 1987), and so can result in reduced levels of

adverse impact. In a recent study by Wee, Newman, and Joseph (2014), the authors were able to demonstrate reduced adverse impact without a reduction in the quality of their selection system. They did so by optimizing the weights assigned to various specific cognitive abilities rather than unit-weighting such abilities, or simply using a measure of general ability. Wee et al.'s (2014) study demonstrates that the use of specific abilities rather than GMA has the potential to limit instances of adverse impact, while still maintaining the validity of a selection system.

Modern Intelligence Testing

Though there are many methods that can be employed to determine an individual's level of ability, the most common method for measuring cognitive abilities is through standardized testing. Standardized tests have the advantages of being reliable and easy to both administer and score (Ones, Dilchert, & Viswesvaran, 2012). These tests are often designed to measure specific abilities or constructs related to intelligence such as scholastic aptitude and are often generically referred to as cognitive ability tests. Other tests, however, are designed to measure intelligence itself (g) using one or more item types. These tests are focused more broadly than cognitive ability tests and are often referred to as intelligence tests. They are usually scored along a continuum and normed to a mean of 100 and a standard deviation of 15 (Neisser et al., 1996). Traditionally, cognitive ability tests have been scored and analyzed through factor-analytic models. These approaches were key to early understandings of the structure of intelligence (Embretson & McCollam, 2000). In recent

decades, however, factor-analytic models have been overtaken by IRT models for the purposes of test development and scoring (Embretson & McCollam, 2000).

One of the main drawbacks of using factor analysis when studying intelligence is that the traditional factor-analytic approach is ill-suited for studying the dimensionality of dichotomous test items (Embretson & Reise, 2000). This is due to the fact that the linear factor analysis of correlations between dichotomous items produces a factor structure that is confounded by item difficulty (Stone & Yeh, 2006). Items of a similar difficulty level tend to correlate more highly with one another than items of non-similar difficulty. These correlated items have the potential to produce spurious factors based on their similar distributions rather than because the latent construct is truly multidimensional (Green, 1983). Because most measures of intelligence involve tests with dichotomously scored (right/wrong) answers, the issue of spurious correlations presents a problem for examining the structure of intelligence.

Where test design is concerned, IRT offers many advantages over classical test theory (Embretson & McCollam, 2000). First, individuals' ability can be scaled along an interval- or ratio-level scale. Under CTT, individual scores are typically the sum or percentage of items answered correctly. With IRT, individuals are scored along a continuum that represents an estimate of an underlying ability or a latent trait, commonly referred to as Θ. Second, the estimates of an item's difficulty are not sample dependent, and thus a test can be administered to different populations while maintaining a comparable scoring system. Finally, IRT models are much better suited to handling missing data. In CTT, participants

with missing values are unable to be scored without estimating these missing responses. In IRT, however, information is gleaned at the item level, and thus a score can still be obtained from the completed items despite missing values. Because missing data are not as detrimental to IRT, tests can be designed in a way that different individuals can be presented with different questions. When test-takers do not all receive the same test, it is much more difficult for a test to be illicitly distributed, aiding in test security.

As indicated above, standard IRT models calculate an estimated amount of a latent ability or trait. This means that with such models, there is an assumption that the items on the test are all measuring the same underlying trait or ability. This assumption is referred to as the unidimensionality assumption (Embretson & Reise, 2000). When this assumption is not met, a unidimensional IRT model may be inappropriate. When it comes to tests of cognitive ability, the assumption of unidimensionality is likely inaccurate. Psychological processes are increasingly being seen as more complex than initially believed (Reckase, 1997; Snow, 1993). This means that, in actuality, a candidate's response to a test item is dependent upon a combination of a number of cognitive processes. For instance, successful completion of a word problem requires the respondent to have a sufficient level of reading ability to assess the problem, and a sufficient level of mathematical ability to solve the problem. This is consistent with the hierarchical view of intelligence in that performance is governed by the *g* factor and one or more broad abilities. Thus, given the complexity of intelligence, unidimensional IRT may not be the best-suited approach for measuring intelligence.

When the assumption of unidimensionality is violated, researchers wishing to utilize IRT are able to turn to multidimensional item response theory (MIRT). MIRT refers to a class of models in which the characteristics of an individual are described using a vector of latent constructs (Reckase, 1997) rather than a single construct (such as *g*). That is, two or more Θ parameters are used to describe an individual (Embretson & Reise, 2000). Thus, multidimensional models are appropriate when a test assesses more than one underlying trait.

Similar to factor analysis, MIRT models are often classified as exploratory or confirmatory (Embretson & Reise, 2000). Exploratory models estimate item and person parameters on two or more dimensions in order to maximize model-data fit. Confirmatory models, conversely, estimate parameters for dimensions that are specified *a priori* based on a theory of the processes that govern item performance. Confirmatory MIRT models can be further classified into one of two groups: compensatory and noncompensatory. In a compensatory model, within a single item, high ability on one dimension can compensate for a low ability on another dimension. In a noncompensatory model, there is a threshold of ability that must be met on all dimensions in order to give a correct response to an item. Despite substantial differences in calculation, these two models typically fit test data equally well (Spray, Davey, Reckase, Ackerman, & Carlson, 1990). Given the similarity of results between the two and the complexity of measurement of the noncompensatory model, the compensatory model has become the dominant model used in research literature (Reckase, n.d.).

MIRT models have been used to successfully model cognitive abilities in many educational applications (Hartig & Hohler, 2008; Lakin & Gambrell, 2012; Li, Jiao, & Lissitz, 2012). These applications have recently been extended to research in an employment setting. Results by Makranksy and Glas (2013) demonstrated that the use of MIRT models for cognitive ability computer-adaptive tests (CATs) can improve test precision, allow for shorter tests, and improve utility of selection systems.

Current Study

The purpose of the current study is to determine if the use of a multidimensional analysis is better suited than a unidimensional analysis for tests of cognitive ability in an employment setting. To do so, we examine three criteria: how well a multidimensional model fit the data, whether scores produced by MIRT analysis are predictive of job performance, and whether analysis through MIRT has the potential to reduce adverse impact.

The first criterion through which we determine suitability is model fit. In IRT, fit can be assessed at the item, person, or model level; fit indices can also be compared between two different models (Embretson & Reise, 2000). By modeling the cognitive ability data with both the unidimensional and multidimensional IRT models, the fit indices can be compared to determine which model better reflects the observed data. Based on the literature reviewed above regarding the nature of intelligence, the following hypothesis is proposed:

H1: A multidimensional model of intelligence will yield better model fit than a unidimensional model of intelligence.

Given that job performance is such an important phenomenon, investigation of ways to conceptualize and measure intelligence to help predict job performance is warranted. Thus, the second criterion to determine whether MIRT is suited to tests of cognitive ability is in the prediction of job performance. The use of MIRT may improve the prediction of job performance by separately scoring the specific abilities measured by cognitive ability tests and determining which abilities are most job-relevant. Because of the disagreement in the literature regarding the utility of specific abilities beyond that of GMA, the following research question is presented:

RQ1: Will a multidimensional model of intelligence identify specific abilities that can better predict job performance than a unidimensional model of intelligence?

Regardless of whether cognitive ability tests highly predict job performance, their use in employee selection is still contentious if these tests demonstrate adverse impact. Thus, the third criterion to determine the suitability of using MIRT for tests of cognitive ability is in the reduction of adverse impact. Previous research has established that looking at specific abilities rather than GMA has the potential to reduce the likelihood of adverse impact (Hough, Oswald, & Ployhart, 2001; Wee, Newman, & Joseph, 2014). By utilizing MIRT to assess candidate abilities, we hope to distinguish these specific abilities from GMA in the prediction of job performance. As such, it may be possible to use MIRT as a scoring system for cognitive ability tests to reduce adverse impact. Thus, the following research question is presented:

RQ2: Will a multidimensional model of intelligence reduce adverse impact as compared to a unidimensional model of intelligence?

Method

Participants

Data for this study were obtained from a large research organization. Two different types of datasets were used. The first dataset contained information from the development of a cognitive ability measure. Participant data from the sample (N = 5,528) came from a diverse group of job seekers, generally college graduates, who completed the cognitive ability measure as a practice test before they completed one that would be used for selection purposes. These participants were primarily from the United States and the United Kingdom. Participant data from the second type of dataset came from job incumbents for whom cognitive ability test scores and performance data were available. These data came from two independent studies: (1) retail clerks and cashiers at a discount tool and equipment retailer (N = 285) and (2) claims representatives and claims specialists at an insurance agency (N = 132).

Measures

Cognitive Ability. A proprietary measure of cognitive ability from a large international consulting organization was used. This measure consists of a battery of 312 test questions covering 4 domains: verbal and numerical ability, inductive reasoning, and deductive reasoning. Participants in the first dataset answered between 6 and 27 items, while participants in the second dataset answered a maximum of 12 items evenly divided among

the four domains. Participants were allowed 10 minutes to complete the assessment. Time, however, was not factored into the participant's score. For example items, see Appendix A.

Job Performance. For each job incumbent who completed the ability measure, a number of job performance ratings were also obtained (see Appendix B). Three composite rating scores were generated for each individual. The first rating, performance area ratings, are ratings of competencies specific to the job in question, such as report writing and using numbers. The second, key performance indicators, is a composite rating of factors that focus on actions incumbents perform on the job in various areas, such as numerical and verbal ability, knowledge, and learning. The final ratings, overall performance rating, are ratings of overall or global job performance.

Analyses

Several different models were constructed to represent scores on the cognitive ability measure. First, a scale score was computed by determining the number of items that a participant answered correctly divided by the number of items that were attempted. This is the traditional method of scoring assessments under the CTT framework. All of the other models used were IRT models. For both the unidimensional and multidimensional IRT analyses, the 2 parameter logistic (2pl) model was used. The unidimensional form of the 2pl model estimates the correct response on item i for individual j through the following equation:

$$P(U_{ij} = 1 | a_i, b_i, \Theta_j) = \frac{e^{a_i(\Theta_j - \beta_i)}}{1 + e^{a_i(\Theta_j - \beta_i)}} \quad (1)$$

where Θ_j refers to an individual's ability, a_i refers to the discrimination of an item, and b_i refers to the item difficulty (Reckase, 2009). The 2plm was used because the number of response options varied among the item types, making the calculation of an overall guessing parameter more difficult.

For the multidimensional form of the 2plm, the form of the equation remains relatively unchanged:

$$P(U_{ij} = 1 | a_i, d_i, \Theta_j) = \frac{e^{(a_i \Theta_j' + d_i)}}{1 + e^{(a_i \Theta_j' + d_i)}} \quad (2)$$

The *a* term has been distributed through the exponent, resulting in the form $a\Theta$-ab, with the ab replaced by d, which represents an intercept term. Thus, the $a\Theta'$ term represents a multidimensional space in which *a* is 1 x *m* vector of item discriminations and Θ is a 1 x *m* vector of ability estimates with *m* representing the number of dimensions (Reckase, 2009). The exponent can be expanded to demonstrate the interactions between the *a* and Θ terms in the multidimensional space:

$$a_i \Theta_j' + d_i = a_{i1}\Theta_{j1} + a_{i2}\Theta_{j2} + \cdots + a_{im}\Theta_{jm} + d_i + \sum_{l=1}^{m} a_{il}\Theta_{jl} + d_i \quad (3)$$

In addition to the standard multidimensional models, a bifactor MIRT model was also fit to the data. The bifactor model is a specific case of the MIRT model in which every item loads onto a single dimension, or factor, and a maximum of one other factor. This structure imitates the hierarchical structure associated with cognitive ability (McGrew, 2009). For this study, all items were loaded onto a single factor and one of four additional factors, depending on the type of item. In this model, the first factor corresponds to the variance shared by all items (i.e., a g factor). The second factor represents the variance unique to deductive

reasoning, the third representing inductive reasoning, the fourth representing numerical ability, and the fifth representing verbal ability.

IRTPRO 2.1 (Cai, du Toit, & Thissen, 2011) software was used to estimate the IRT model parameters and individual ability scores. Because of the complexity of the desired models, the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010) was used to estimate model parameters. This algorithm is the most suitable for estimating models with more than two or three dimensions (SSI, 2011). Expected *a posteriori* (EAP) estimation was used to calculate individual θ estimates. This estimation method has been shown to exhibit smaller average error in the population than other estimation methods (SSI, 2011).

Results

Dimensionality Assessment

Because each participant only responded to a small fraction of the dataset, traditional dimensionality assessments were not possible. Factor analysis could not be utilized because there were no sets of complete data for any individual. With the amount of missing data, neither imputation nor pairwise deletion were feasible. Given these limitations, a correlation matrix could not be constructed. In addition, MIRT dimensionality procedures typically employed, such as the use of DIMTEST and DETECT, are unable to handle missing data. Thus, to statistically determine how many dimensions best represented the data, model fit estimates were obtained for models that contained between one and five dimensions. These models, along with the bifactor model, were compared to test the first hypothesis.

In order to compare the IRT models, there are a few criteria that can be used to determine which model best fit the data (Li, Jiao, and Lissitz, 2012). First, the log-likelihood (-2LL) statistics were compared using a likelihood ratio (χ^2 difference) test. In addition, AIC and BIC statistics were compared, with lower values indicating better fit (Akaike, 1974; Schwarz, 1978). For the likelihood ratio test, the models were compared incrementally, with each model being compared against the one with one fewer dimension.

As shown in Table 1, the three-dimensional model exhibited the best fit of all the models based on the -2LL test statistics, AIC and BIC. In addition, the bifactor model was compared against both the unidimesional model and the four-dimensional model. The bifactor model demonstrated significantly better fit than the unidimensional model ($\Delta\chi^2(302) = 706.96$, $p < .001$). This demonstrates that a model in which items are influenced by several different abilities provides a better description of the data than one in which a single ability determines individual responses. The bifactor model also provided significantly better fit than the 4 factor model ($\Delta\chi^2(598) = 13013.99$, $p < .001$). This result shows that a model in which all items are related through a common factor better fits the data than one in which there is not a common factor, lending support to the hierarchical nature of cognitive abilities. Taken together, these results provide support for the first hypothesis by demonstrating that a multidimensional model does, in fact, fit the data better than a unidimensional model. Because the three-dimensional model demonstrated the best fit of the multidimensional models, it was retained for further analysis. In addition, the bifactor model was also retained as model choice is often driven by theoretical considerations (Hartig & Hohler, 2008).

Criterion-related Validity. In order to test the first research question, the criterion-related validity was determined for each of four scoring systems. To do so, the scores derived from each model were related to the job performance measures: performance area ratings, key performance indicators, and overall performance ratings. For the summed score and unidimensional IRT model, each individual's ability is represented by a single score. Thus, these scores were related to performance ratings through the use of Pearson correlations. For the multidimensional IRT models, each individual is represented by several ability estimates, corresponding to the number of dimensions proposed in the model. Intercorrelations between the dimensions for each multidimensional model can be found in Table 2. Because multiple ability estimates are obtained, multiple regression was used to determine the relationship each of these abilities has on the performance ratings. r^2 values were also obtained to determine how well specific ability scores generated by the MIRT analysis explained the variance in job performance. These values were compared to the r^2 values for the summed score and unidimensional models.

To determine the criterion-related validity of the traditional CTT scoring method, the summed score for each individual on the cognitive ability measure was correlated with each of the three performance outcomes. This model was used because it is often the one employed in traditional testing environments. The results of this analysis can be found in Table 3. The summed score was shown to significantly predict key performance indicators in both studies (Study 1: $r = .17$, $p < .01$; Study 2: $r = .19$, $p < .05$) and performance area

ratings in the first study (r = .18, p < .01). The amount of variance in job performance ratings explained by these scores is shown in Table 4.

Next, individual scores from the unidimensional IRT model were correlated with the three performance indicators. The unidimensional model was chosen because the cognitive-ability measure in question was designed to be assessed using such a model, as well as because it is increasingly being used in many applied settings. The unidimensional model was shown to significantly predict performance area ratings in Study 1 (r = .14, p < .05). Next, the three ability estimates were generated by the three-dimensional model were correlated with each of the three performance indicators. As shown in Table 3, no ability estimates on any dimension were significantly related to any performance indicators. In addition, this model explained little of the variance in any performance rating (r^2=.00-.02).

The final model chosen was the bifactor model because this model aligns most closely to the current theories of intelligence. This model produced five ability estimates: one that corresponded to each item type, and one for a general factor. These ability estimates were also correlated with each performance indicator. As shown in Table 2, the second factor, that corresponding to deductive reasoning, was significantly related to overall performance (r = .19, p < .05) and key performance indicators (r = .20, p < .05) in Study 2. Interestingly, the general factor was not significantly related to any performance indicator. For Study 2, ability estimates in the bifactor model explained much more of the variance in job performance ratings than those of the unidimensional or summed score models (r^2=.05-.06). In response to Research Question 1, these findings provide evidence that in certain

circumstances specific abilities do have the potential to provide criterion-related validity separate from GMA. Further, an analysis of test data using MIRT has the potential to substantially improve the variance in performance explained by test of cognitive ability.

Tests of Adverse Impact

In order to test the second research question, adverse impact rates were assessed by establishing cutoff rates to simulate a selection system. Individual scores for each model included in the criterion-related validity analysis were tested separately. The three-dimensional model was excluded because it failed to provide criterion-related validity.

First, individual scores were standardized into z scores. Individuals were then placed into five equal bands representing percentiles at 20% increments according to their z score. Next, pass rates were determined for each racial sub-group at various passing criteria (e.g., top 20%, top 40%, etc.). Finally, the pass rates for each group were compared to the group with the highest passing rate.

These passing rates and the corresponding adverse impact ratios may be found in Tables 5-7. The criteria used to determine the presence of adverse impact was the 4/5 ratio rule as established by the *Uniform Guidelines*. As shown, the summed score and unidimensional models have a pass-rate ratio of less than .8, and thus exhibit adverse impact against Blacks and Hispanics at all levels except when the cutoff is set such that 80% of candidates pass. The general factor of the bifactor model exhibits similar levels of adverse impact. However, when looking at the deductive reasoning factor of the bifactor model, which demonstrated criterion-related validity in one sample, no adverse impact toward blacks

is present. There is, however, evidence of adverse impact toward Hispanics at the most restrictive criterion. In answer to Research Question 2, these findings demonstrate that the isolation of specific abilities and their use in a selection system has the potential to reduce adverse impact.

Discussion

When validating tests in an organizational setting, organizations have many options from which to choose in terms of scoring their selection instruments. Which method they use may rely on a host of factors, including ease of use and accuracy of prediction. As the computer software used to perform these complicated analyses advances, more-elaborate statistical procedures are able to be utilized. These advances allow for researchers and practitioners alike to develop their theories and models for complex constructs such as intelligence.

The finding that a multidimensional model fit the data better than a unidimensional model supports previous research that tests of cognitive ability involve the complex interaction of several different cognitive abilities (Reckase, 1997; Snow, 1993). In addition, the improvement in fit from the four dimensional model to the bifactor model demonstrates that responses to the different item types are not independent of one another, but rather they share a common variance factor that can likely be attributed to general ability. These findings are consistent with much of the previous research on the dimensionality of cognitive ability (Lubinski, 2004; Reckase, 1997; Snow, 2003) and contributes to the ongoing development of theories about the structure of intelligence.

Though other studies have demonstrated the utility of using MIRT dimensions to predict job performance (Makransky & Glas, 2013), this is the first to do so using real-world performance data. This study failed to demonstrate the utility of using unidimensional IRT models over a simple scale score in predicting job performance. This finding is contrary to the assumption that because unidimensional IRT models estimate an underlying ability, they should be more useful in predicting outcome measures (Thissen & Orlando, 2012). This is, however, consistent with a simulation study performed by Xu and Stone (2012) that found summed scores to be comparable to IRT models in predicting outcomes. Those authors concede that IRT models may work better in situations where tests contain items at the extreme ends of the ability distribution, but these situations are not common in practice (Xu & Stone, 2012).

In the first criterion-related validation study, the summed score outperformed the unidimensional IRT model for predicting performance area ratings, and was the only significant predictor of key performance indicators. Neither the three-dimensional model nor the bifactor model were significant predictors of any of the job performance indicators in this study. However, in the second criterion-related validation study, the deductive-reasoning factor of the bifactor model significantly predicted overall performance and key performance indicators, while scores from the unidimensional model did not correlate with any performance outcomes. Furthermore, for Study 1, the bifactor MIRT model explained as much variance in almost all performance ratings as the unidimensional IRT model, and much more than the unidimensional IRT model in Study 2. This gives support to the idea that for

some situations, specific abilities may contribute in addition to, or even beyond, scores of GMA. This result may indicate that for certain jobs, different factors may be more predictive of performance.

Perhaps one of the most noteworthy findings of this study is the analysis of adverse-impact rates. Consistent with previous studies (Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997), the cognitive-ability measure demonstrated subgroup differences such that racial minorities scored lower than the majority group. If such a test were used in a selection context, it would likely exhibit adverse impact. However, if one were to use only scores on the second factor of the bifactor model, not only would these scores be predictive of performance in certain contexts, but it would also potentially eliminate adverse impact from the cognitive ability component of the selection system.

There is currently much controversy surrounding the EEOC's (1978) definition of adverse impact. McDaniel, Kepes, and Banks (2011) argue that these guidelines have not kept up to date with the science of best practices, and run contrary to the other guiding documents for the field of personnel selection, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003). However, until the EEOC guidelines are revised, the 4/5ths rule will still be the guiding premise of selection practitioners because any legal challenge to these selection systems,

regardless of the verdict, is a time-intensive and costly affair (McDaniel, Kepes, & Banks, 2011). Therefore, organizations are constantly seeking ways to minimize the risk of adverse impact while still preserving the validity of their selection system (Ployhart & Holtz, 2008). The results of this study show that utilization of MIRT for cognitive-ability measures may provide one such solution.

Limitations and Future Directions

One of the primary limitations of this study was the nature of the test used. Participants only responded to a small fraction of the items available, meaning the resultant data set was too sparsely populated to conduct traditional dimensionality assessment. Another limitation of these analyses is that they may not be generalizable to different tests, different job contexts, or different performance indicators. This study does, however, provide a proof-of-concept for looking at cognitive ability tests using a multidimensional framework and demonstrates that such analyses may have merit.

Another limitation of this study is the complexity of the model dimensionality. When using MIRT, most analyses focus on describing each item with only one dimension, in what is known as simple structure (Ackerman, Gierl, & Walker, 2003). When each items is allowed to represent multiple dimensions, in what is known as complex structure, the results are often hard to interpret because clusters of items identified by each dimension are frequently quite homogenous (Zhang & Stout, 1999). As such, an interpretation of the three-dimensional model that best fit the data could not be discerned. This limits the amount to

which test-developers would be able to capture multiple specific abilities within individual items.

Finally, there are several practical issues in the use of MIRT to model cognitive ability test data. First, these analyses are complex and require specialized software with which practitioners are likely unfamiliar. Second, even when certain dimensions are found to be predictive of job performance while also minimizing adverse impact, it may be hard to justify weighting these abilities differently than the others because we do not have a comprehensive picture of specific cognitive abilities and their influence on subgroup differences (Outtz & Newman, 2010). Third, explaining to organizational stakeholders how findings from such studies can improve the quality of new hires to organizational stakeholders may prove arduous at best.

This study demonstrates the potential for MIRT-based analysis of cognitive-ability measures to improve adverse-impact rates while still providing benefit as a predictor of job performance. These results have implications for the development of adaptive tests using an MIRT framework which have already been demonstrated to improve test precision, allowing for shorter tests (Makransky & Glas, 2013). Such adaptive multidimensional tests would likely have a positive impact on employee testing, and would allow for the consideration of multiple skill-sets within one examination. This study also demonstrates a method through which organizations can fine-tune their cognitive ability tests. By looking at the dimensionality of the test with respect to criterion-related validity, these organizations can

create a specialized test to be used in a given context by using items that better assess the cognitive abilities that are most related to job performance.

The results of this study, as well as the many studies conducted on the scoring of personnel assessments, emphasize that there is no best-fit solution for all testing situations. Researchers should consider and explore many different solutions to test scoring to determine which best fit the situation, and which provide the most utility given the unique circumstances in which a test is implemented. As demonstrated in this study, a combined approach of empirical and theoretical examination of these tests can provide not only more theoretically sound measurement models, but can also improve the utility of an organization's testing program.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53. doi:10.1111/j.1745-3992.2003.tb00136.x

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. (2nd ed.). Washington, DC: American Educational Research Association.

Bowman, D., Markham, P. M., & Roberts, R. D. (2002). Expanding the frontier of human cognitive abilities: So much more than (plain) *g*! *Learning and Individual Differences, 13*, 127-158.

Brody, N. (2000). History of theories and measurements of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 16-33). New York, NY US: Cambridge University Press.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307-335. doi: 10.3102/1076998609353115

Cai, L., du Toit, S.H.C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling. Chicago, IL: Scientific Software International, Inc.

Carroll, J. B. (1993) *Human Cognitive Abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychology Bulletin, 38*, 592.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1-22. doi:10.1037/h0046743

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780198524175.001.0001

Embretson, S. E., & McCollam, K. (2000). Psychometric approaches to understanding and measuring intelligence. In R. J. Sternberg (Ed.) *Handbook of intelligence* (pp. 423-444). New York, NY US: Cambridge University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, & Department of Labor. (1978). *Uniform Guidelines on employee selection procedures*. 29 CFR, 1607.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York, NY US: Basic Books.

Goertz, W., Hülsheger, U. R., & Maier, G. W. (2014). The validity of specific cognitive abilities for the prediction of training success in Germany: A meta-analysis. *Journal of Personnel Psychology, 13*(3), 123-133. doi:10.1027/1866-5888/a000110

Goldstein, H. W., Scherbaum, C. A., & Yusko, K. (2009). Adverse impact and measuring cognitive ability. In J. Outtz (Ed.) *Adverse impact: Implications for organizational staffing and high stakes testing* (pp. 95–134). New York, NY: Psychology Press.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). g: Is this your final answer? *Human Performance, 15*(1-2), 123-142. doi:10.1207/S15327043HUP1501&02_08

Gottfredson, L. S. (1984). The role of intelligence and education in the division of labor (Report No. 355). Baltimore, MD: John Hopkins University, Center for Social Organization of Schools.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence, 24*(1), 13-23. doi:10.1016/S0160-2896(97)90011-8

Green, S. B. (1983). Identifiability of spurious factors with linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 3-13.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within- and between-item multidimensionality. *Journal of Psychology, 2,* 89-101.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitude, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340–362.

Jensen, A.R. (1987). Individual differences in mental ability. In J.A. Glover & R.R. Ronning (Eds.), *Historical foundations of educational psychology* (pp. 61-88). New York: Plenum.

Kehoe, J. F. (2002). General mental ability and selection in private sector organizations: A commentary. *Human Performance, 15*, 97–106.

Lakin, J. M., & Gambrell, J. L. (2012). Distinguishing verbal, quantitative, and nonverbal facets of fluid intelligence in young students. *Intelligence, 40*(6), 560-570.

Lang, J. B., Kersting, M., Hülsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: the perspective of the nested-factors model of cognitive abilities. *Personnel Psychology, 63*(3), 595-640. doi:10.1111/j.1744-6570.2010.01182.x

LeBreton, J. M., Hargis, M. B., Griepentrog, B., Oswald, F. L., Ployhart, R. E. (2007). Multidimensional approach for evaluating variables in organizational research and practice. *Personnel Psychology, 60*, 475–498.

Li, Y., Jiao, H., & Lissitz, R.W. (2012). Applying multidimensional IRT models in validating test dimensionality: An example of K-12 large-scale science assessment. *Journal of Applied Testing Technology, 13*(2), 1-27.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "General intelligence,' objectively determined and measured'. *Journal of Personality And Social Psychology, 86*(1), 96-111. doi:10.1037/0022-3514.86.1.96

Makransky, G., & Glas, C. A. W. (2013). The applicability of multidimensional computerized adaptive testing for cognitive ability measurement in organizational assessment. *International Journal of Testing, 13*, 123-139.

McDaniel, M. A., Kepes, S., & Banks, G. C. (2011). The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology: Perspectives on Science And Practice, 4*(4), 494-514. doi:10.1111/j.1754-9434.2011.01382.x

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136-181). New York, NY US: Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. doi:10.1016/j.intell.2008.08.004

Moffitt, T. E., Gabrielli, W. E, Mednick, S. A., & Schulsinger, E (1981). Socioeconomic status, IQ, and delinquency. Journal of Abnormal Psychology, 90, 152-156.

Murphy, K. R. (1996). Individual differences and behavior in organizations: Much more than g. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 3–30). San Francisco, CA: Jossey-Bass.

Murphy, K., Cronin, B., & Tam, A. (2003). Controversy and consensus regarding the use of cognitive ability testing in organizations. *Journal of Applied Psychology*, 88, 660–671.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11*, 21-43.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77-101.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2010). Cognitive abilities. In J. L. Farr, N. T. Tippins (Eds.) *Handbook of employee selection* (pp. 255-275). New York, NY US: Routledge/Taylor & Francis Group.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In O. Wilhelm, R. W. Engle (Eds.) *Handbook of understanding and measuring intelligence* (pp. 431-468). Thousand Oaks, CA US: Sage Publications, Inc. doi:10.4135/9781452233529.n24

Outtz, J. L., & Newman, D. A. (2010). A theory of adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 53–94). New York, NY: Routledge/Taylor & Francis Group.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153-172. doi:10.1111/j.1744-6570.2008.00109.x

Prediger, D. S. (1989). Ability differences across occupations: More than *g*. *Journal of Vocational Behavior, 34*, 1–27. doi:10.1016/0001-8791(89)90061-4

Pyburn, K. R., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*(1), 143-151. doi:10.1111/j.1744-6570.2008.00108.x

Reckase, M. D. (n.d.). Multidimensional item response theory. Retrieved from irt.com.ne.kr/data/MIRT.doc

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36. doi:10.1177/0146621697211002

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

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44*, 321–332. doi:10.1111/j.1744-6570.1991.tb00961.x

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*, 518–524. doi:10.1037/0021-9010.79.4.518

Reeve, C. L. (2004). Differential ability antecedents of general and specific dimensions of declarative knowledge: More than g. *Intelligence, 32*, 621–652. doi:10.1016/j.intell.2004.07.006

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297–330. doi:10.1111/j.1744-6570.2001.tb00094.x

Scherbaum, C. A., Goldstein, H. W., Yusko, K. P., Ryan, R., & Hanges, P. J. (2012). Intelligence 2.0: Reestablishing a research program on g in I–O psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 5*(2), 128–148.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1-2), 187-211. doi:10.1207/S15327043HUP1501&02_12

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.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173. doi:10.1037/0022-3514.86.1.162

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 719-730.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464. doi:10.1214/aos/1176344136

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Spearman**,** C. (1904). "General intelligence**,**" objectively determined and measured. *American Journal of Psychology, 15*, 201– 292.

Spray, J.A., Davey, T., Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1990). Comparison of two logistic multidimensional item response theory models. (ACT Research Report - ONR 90-8)

Scientific Software International, SSI. (2011). *IRTPRO: User's Guide*. Lincolnwood, IL: Scientific Software International, Inc.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Sternberg, R. J. (1985): *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. (2000). The concept of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 3-15). New York, NY US: Cambridge University Press.

Stone, C. A., & Yeh, C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the multistate bar examination. *Educational and Psychological Measurement, 66*(2), 193-214. doi:10.1177/0013164405282483

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-137). London, England: Lawrence Erlbaum.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*(5), 406-427. doi:10.1037/h0069792

Viswesvaran, C., & Ones, D. S. (2002). Agreements and disagreements on the role of general mental ability (GMA) in industrial, work, and organizational psychology. *Human Performance, 15*(1-2), 212-231. doi:10.1207/S15327043HUP1501&02_13

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than *g*: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology, 99*(4), 547-563. doi:10.1037/a0035183

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. Educational and Psychological Measurement, 72(3), 453-468. doi:10.1177/0013164411419846

Zhang, J. M, Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Table 1

Model dimensionality fit statistics

Model	AIC	BIC	-2LL	Free Parameters	-2LL test
1 dimension	103926.96	103926.96	102718.96	604	
2 dimensions	103585.64	103585.64	101773.64	906	945.32**
3 dimensions	103532.76	103532.76	101116.76	1208	656.88**
4 dimensions	118033.99	118033.99	115025.99	1504	13909.23
5 dimensions	105624.75	105624.75	102000.75	1812	13025.24
Bifactor	103824.00	103824.00	102012.00	906	

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion, -2LL = -2Loglikelihood. Degrees of freedom for -2LL test represent difference in free parameters between models.
**$p < .001$

Table 2

Intercorrelations Between Multidimensional Model Dimensions

		Dimensions			
		1	2	3	4
3 Dimensional	1				
	2	-.05*			
	3	-.06*	-.04*		
Bifactor	1				
	2	-.01			
	3	.02	.03		
	4	.01	.00	.02	
	5	.00	-.01	-.01	-.02

*N = 5,945; *p* < .001

Table 3

Validity Coefficients for Summed-Score, Unidimensional Model, 3 Dimensional Model and Bifactor Model

Performance Indicators	% Correct	Uni-dimensional	3 Dimensional			Bifactor				
			1	2	3	1	2	3	4	5
Overall Performance										
Study 1	.10	.05	.04	-.01	.03	.01	.01	-.01	.04	.00
Study 2	.03	.03	.10	.06	-.09	-.04	.19*	.09	.06	.01
Key Performance Indicators										
Study 1	.17**	.11	.07	-.05	.06	.05	.06	.00	.06	-.03
Study 2	.19*	.12	-.01	.05	.05	.10	.20*	.10	-.01	.02
Performance Area										
Study 1	.18**	.14*	.09	.00	.05	.03	.04	-.01	-.02	-.04
Study 2	.17	.14	-.03	.06	.04	.15	.16	.03	.00	-.09

Note. Validity coefficients represent *r* values.
*$p < .05$ **$p < .01$

Table 4

Variance Explained for Summed-Score, Unidimensional Model, 3 Dimensional Model and Bifactor Model

	% Correct	Uni-dimensional	3 Dimensional				Bifactor					
Performance Indicators	r^2	r^2	1	2	3	r^2	1	2	3	4	5	r^2
Overall Performance												
Study 1	.01	.00	.05	-.01	.04	.00	.01	.02	-.01	.04	-.01	.00
Study 2	.00	.00	.09	.07	-.08	.02	-.03	.19*	.09	.06	.01	.05
Key Performance Indicators												
Study 1	.03	.01	.08	-.05	.07	.01	.05	.07	.01	.06	-.02	.01
Study 2	.04	.01	-.01	.05	.05	.00	.11	.20*	.08	.00	-.02	.06
Performance Area												
Study 1	.03	.02	.09	.00	.03	.01	.03	.04	-.01	-.03	-.07	.01
Study 2	.03	.02	-.02	.06	.04	.01	.17	.17	.01	.00	-.09	.06

Note. For the multidimensional models, values represent standardized beta weights for each dimension.
*$p < .05$

Table 5

Adverse Impact Ratios for Summed Score Model

	20% Passing		40% Passing		60% Passing		80% Passing	
	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio
Gender								
Male	20.6%	1.00	38.6%	1.00	59.8%	1.00	80.4%	1.00
Female	18.7%	0.91	36.7%	0.95	57.2%	0.96	77.8%	0.97
Race								
White	23.8%	1.00	43.1%	1.00	64.6%	1.00	84.1%	1.00
Black	9.4%	0.40*	21.9%	0.51*	44.1%	0.68*	67.0%	0.80
Asian	22.8%	0.96	42.4%	0.98	63.3%	0.98	81.7%	0.97
Native	25.0%	1.05	50.0%	1.16	75.0%	1.16	87.5%	1.04
Hispanic	16.9%	0.71*	27.2%	0.63*	45.6%	0.71*	66.9%	0.80

**Indicates adverse impact present*

Table 6

Adverse Impact Ratios for Unidimensional Test

	20% Passing		40% Passing		60% Passing		80% Passing	
	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio
Gender								
Male	20.6%	1.00	39.8%	1.00	60.8%	1.00	81.1%	1.00
Female	18.5%	0.90	38.5%	0.97	58.1%	0.95	78.3%	0.97
Race								
White	22.0%	1.00	44.0%	1.00	64.9%	1.00	84.2%	1.00
Black	12.2%	0.56*	27.8%	0.63*	47.4%	0.73*	70.4%	0.84
Asian	22.5%	1.02	43.0%	0.98	62.6%	0.97	80.5%	0.96
Native	25.0%	1.14	50.0%	1.14	50.0%	0.77	87.5%	1.04
Hispanic	16.2%	0.74*	28.7%	0.65*	48.5%	0.75*	72.8%	0.86

**Indicates adverse impact present*

Table 7

Adverse Impact Ratios for Bifactor Model

	Factor 1								Factor 2							
	20% Passing		40% Passing		60% Passing		80% Passing		20% Passing		40% Passing		60% Passing		80% Passing	
	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio	Pass Rate	AI Ratio
Gender																
Male	21.5%	1.00	40.8%	1.00	61.3%	1.00	80.8%	1.00	18.9%	1.00	39.2%	1.00	59.9%	1.00	79.2%	1.00
Female	17.6%	0.82	37.7%	0.92	57.8%	0.94	79.0%	0.98	20.7%	1.09	40.4%	1.03	60.5%	1.01	80.4%	1.02
Race																
White	22.6%	1.00	43.3%	1.00	63.9%	1.00	83.2%	1.00	20.1%	1.00	40.8%	1.00	61.9%	1.00	80.8%	1.00
Black	11.3%	0.50*	25.6%	0.59*	46.9%	0.73*	68.9%	0.83	17.8%	0.88	40.0%	0.98	58.0%	0.94	79.3%	0.98
Asian	21.2%	0.94	43.0%	0.99	65.5%	1.03	83.6%	1.00	20.4%	1.01	38.3%	0.94	58.0%	0.94	77.9%	0.96
Native	50.0%	2.21	75.0%	1.73	75.0%	1.17	75.0%	0.90	12.5%	0.62*	37.5%	0.92	50.0%	0.81	87.5%	1.08
Hispanic	16.2%	0.72*	34.6%	0.80	49.3%	0.77*	70.6%	0.85	25.0%	1.24	44.1%	1.08	60.3%	0.97	78.7%	0.97

**Indicates adverse impact present*

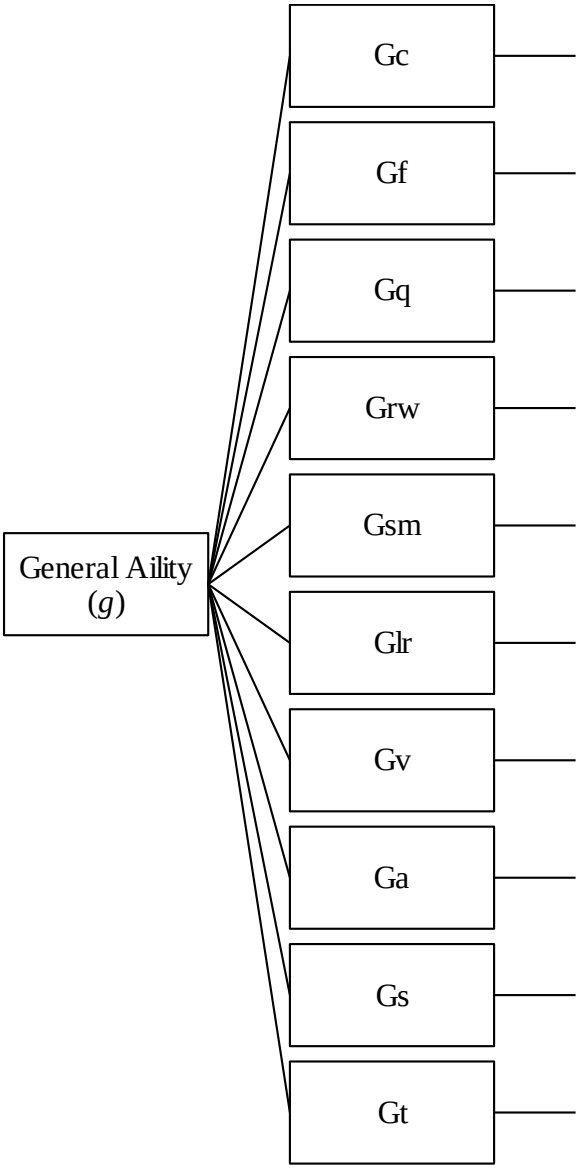


Figure 1. Cattell-Horn-Carroll (CHC) model. Narrow abilities not listed for the sake of space. Gc = Comprehension-knowledge; Gf = Fluid Reasoning; Gq = Quantitative Knowledge; Grw = Reading and Writing Ability; Gsm = Short-term memory; Glr = Long-term Storage and Retrieval; Gv = Visual Processing; Ga = Reaction and Correct Decision Speed; Gs = Processing Speed; Gt = Decision/Reaction Time/Speed

APPENDICES

Appendix A

Example Cognitive Ability Items

Example Verbal Item:

"Many organizations find it beneficial to employ students over the summer. Permanent staff often wish to take their own holidays over this period. Furthermore, it is not uncommon for companies to experience peak workloads in the summer and so require extra staff. Summer employment also attracts students who may return as well-qualified recruits to an organization when they have completed their education. Ensuring that the students learn as much as possible about the organization encourages interest in working on a permanent basis. Organizations pay students on a fixed rate without the usual entitlement to paid holidays or bonus schemes."

Statement 1: It is possible that permanent staff who are on holiday can have their work carried out by students.

Statement 2: Students in summer employment are given the same paid holiday benefit as permanent staff.

Statement 3: Students are subject to the organisation's standard disciplinary and grievance procedures.

Statement 4: Some companies have more work to do in the summer when students are available for vacation work.

Example Numerical Item:

For each question below, click the appropriate button to select your answer. You will be told whether or not your answer is correct.

Newspaper Readership				
Daily Newspapers	Readership (millions)		Percentage of adults reading each paper in Year 3	
	Year 1	**Year 2**	**Males**	**Females**
The Daily Chronicle	3.6	2.9	7	6
Daily News	13.8	9.3	24	18
The Tribune	1.1	1.4	4	3
The Herald	8.5	12.7	30	23
Daily Echo	4.8	4.9	10	12

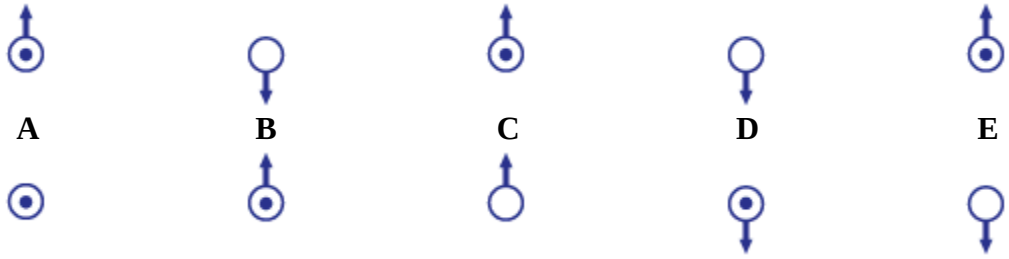
Question 1: Which newspaper was read by a higher percentage of females than males in Year 3?

A	B	C	D	E
The Tribune	The Herald	Daily News	Daily Echo	The Daily Chronicle

Example Inductive Reasoning Item:

In each example given below, you will find a logical sequence of five boxes. Your task is to decide which of the boxes completes this sequence. To give your answer, select one of the boxes marked A to E. You will be told whether or not your answer is correct.

Question 1



Example Deductive Reasoning Item:

Review the facts below:

- *Jane drives a red car.*
- *Susan drives a blue car.*
- *There are no red cars in City A.*
- *Blue cars get 33 miles per gallon of gasoline.*

Based on the information above, which of the following MUST be true?

A. Jane lives in City A.

B. Susan lives in City A.

C. Red cars get 36 miles per gallon of gasoline.

D. Susan's car gets 33 miles per gallon of gasoline.

E. Jane and Susan live in the same city.

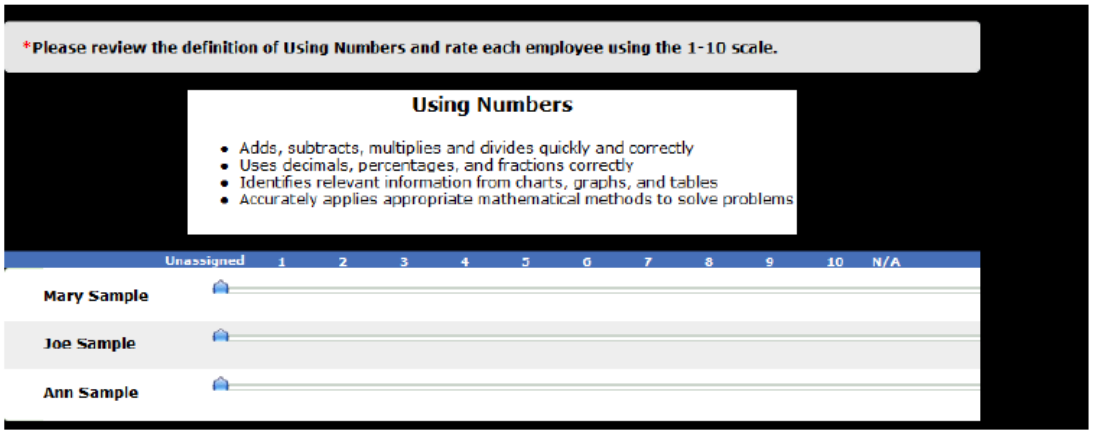
The correct answer is ***D****. Since blue cars get 33 miles per gallon of gasoline, the fact that Susan drives a blue car means that her car gets 33 miles per gallon of gasoline.*

Appendix B

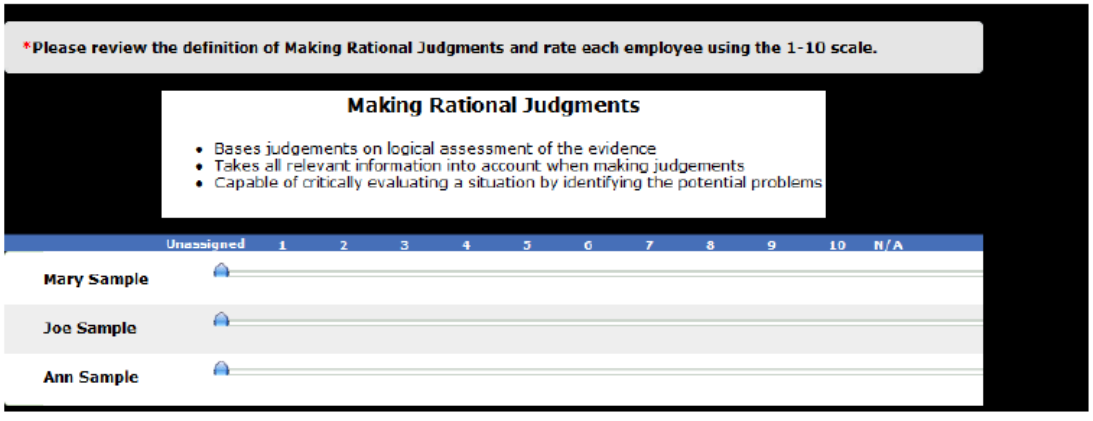
Example Criterion Measures

Performance Area Ratings – Sample items:

Poor to Below Average Performance			Average Performance				Above Average to Outstanding Performance		
1	2	3	4	5	6	7	8	9	10



Poor to Below Average Performance			Average Performance				Above Average to Outstanding Performance		
1	2	3	4	5	6	7	8	9	10



Key Performance Indicator Ratings – Sample items:

1. Compared to others, this employee performs mathematical operations
    - Much slower than most others
    - Slower than most others
    - As fast as most others
    - Faster than most others
    - Much faster than most others
    - Cannot Rate
2. Compared to others, this employee uses proper grammar and appropriate vocabulary in written communication
    - Much less than most others
    - Less than most others
    - As much as most others
    - More than most others
    - Much more than most others
    - Cannot Rate
3. Compared to others, this employee's ability to work on multiple tasks at the same time is
    - Below average
    - Average
    - Above Average
    - Well above average
    - One of the best
    - Cannot Rate
4. Compared to others, this employee asks questions, performs research, and refers to provided materials to answer difficult client questions
    - Much less than most others
    - Less than most others
    - As much as most others
    - More than most others
    - Much more than most others
    - Cannot Rate

Overall Job Performance Ratings:

1. The overall match between this employee's abilities and the job requirements is:
    - Below Average
    - Average
    - Above Average
    - Well Above Average
    - One of the Best
    - Cannot Rate
2. If you had your choice of job candidates, would you hire this employee again?
    - Definitely not
    - Probably not
    - Unsure
    - Probably yes
    - Definitely yes
    - Cannot Rate
3. The employee's productivity level is?
    - Below Average
    - Average
    - Above Average
    - Well Above Average
    - One of the Best
    - Cannot Rate