

A POTENTIAL FRAMEWORK FOR THE DEBATE ON GENERAL COGNITIVE ABILITY
AND TESTING

BY

STEVEN J. COHN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Fritz Drasgow

ABSTRACT

Following up on Murphy, Cronin, & Tam's (2002) study, the dimensionality of the debate on general cognitive ability (GCA) and GCA testing was reexamined. First, a review of a recent literary debate on topics related to GCA was conducted. Then, a six-factor model of beliefs was hypothesized. One factor represented the belief in the primary importance of GCA among trait predictors of performance. Another reflected the view that the predictive validity of GCA is criterion-dependent. The third factor stood for the belief that there are significant cognitive alternatives to GCA. The fourth factor characterized the opinion that GCA tests are racially biased. The belief that GCA tests are socially fair was the fifth factor. The final factor represented the position that there are tradeoffs between fairness and predictive validity in the choice to use GCA tests. This model was tested empirically. It showed reasonable fit to appropriate survey data, though the independence of some of the factors may be questioned. While a broader "debate" over intelligence is seen as a competition between paradigms, a framework for the debate over GCA and GCA testing based on the six-factor model is recommended. To illustrate the utility of this framework in clarifying positions, it is applied to the scholarly debate on GCA which served as the initial literary review for this study. Argumentation methods which may sharpen future debate are also recommended.

For Patrick Laughlin and Paul Suerken
whose memories inspire me to finish what I start, always

ACKNOWLEDGMENTS

I would like to thank Dr. Fritz Drasgow, my adviser, whose patience, guidance and tolerance were of fundamental importance through several months of revisions. I would also like to recognize Dr. Daniel A. Newman, who not only provided me with idea for the thesis but also procured the data for me, and Dr. James Rounds, who gave his time to read this thesis on short notice and provided valuable edits. I appreciate all that I learned at University of Illinois, my brilliant cohorts and teachers, and the opportunities that I, too, was given to teach students. I appreciate Ashley Ramm's assistance, without whom this paper surely would not have made the deadline. I am deeply indebted to Dr. Kevin Murphy, Dr. Brian Cronin and Anita Tam for their groundbreaking work, without which the present study would not have been possible. Above all others, I appreciate those who inspire me to work to the best of my ability and never to give up: my parents, Stuart and Carol Cohn; my wife, Aleta Cohn; my dear friends and mentors, Paul M. Suerken and Patrick Laughlin; and God, through whom all things good are possible.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
1.1 Background.....	1
1.2 The General Cognitive Ability (GCA) and GCA Testing Debate.....	4
CHAPTER 2: THE PRESENT STUDY.....	9
2.1 Introduction to the Six-Factor Model.....	9
CHAPTER 3: METHODS.....	35
3.1 Focus on Construct Validity.....	35
3.2 Procedure.....	37
CHAPTER 4: RESULTS.....	43
4.1 Model Fit.....	43
4.2 Item Factor Loadings.....	44
CHAPTER 5: DISCUSSION.....	50
5.1 The Model.....	50
5.2 Further support for construct validity.....	50
5.3 Value of the Model.....	51
CHAPTER 6: LIMITATIONS.....	53
CHAPTER 7: RECOMMENDATIONS FOR FUTURE RESEARCH.....	55
REFERENCES.....	57
APPENDIX A.....	86

CHAPTER 1

INTRODUCTION

1.1 Background

Since the birth of differential psychology, perhaps no individual trait difference has inspired as much research as intelligence. However, a formal debate about the nature of intelligence may not be possible, as debate fundamentally is pointless without clear, recognized definitions (Branham, 1991, p. 38). Definitions of intelligence seem to be contingent not only on schools of thought (Guion, 1997) but also on the specific researcher (Goldstein, Zedeck, & Goldstein, 2002). These definitions may be quite different. Some scholars emphasize the cognitive processes that underlie intelligence (Humphreys, 1979; Das, 1986; Haggerty, 1921; Thurstone, 1921, 1924; Berry, 1986). For example, Humphreys (1979) defines intelligence as "...the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills" (p. 115). Similarly, Das (1986) states, "Intelligence, as the sum total of all cognitive processes, entails planning, coding of information and attention arousal" (p. 55). Other scholars largely characterize intelligence as learning and problem solving (e.g., Bingham, 1937; Minsky, 1985; Gardner, 1983/2000). Gottfredson (1997a) provides an elaborate definition of this type: "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (p. 13). Still others define intelligence in practical terms, as an ability to adapt to situations and environments (e.g., Baron, 1986; Simonton, 2003; Sternberg, 2003; Wechsler, 1944). Looking back on two major 20th century symposia on the topic of intelligence, Sternberg observes that there were as many definitions as

experts (Sternberg, 1987, p. 376). Prominent psychologists even have suggested that the many definitions of the term “intelligence” have rendered it essentially meaningless (e.g., Spearman, 1927). Indeed, Boring’s (1923) “narrow definition” of intelligence (i.e., “Intelligence is what the tests test”) (p. 35) illustrates the depth of this semantic evisceration.

Aside from myriad working definitions, many theories have been applied to intelligence. Although these definitions vary greatly, they could be assigned to two broad groups. The first includes theories which lack a single unifying general factor, instead proposing separate primary abilities (e.g., Kelley, 1928; Thurstone, 1938) or “multiple intelligences” (e.g., Gardner, 1983/2003). In the second group are theories that include a general factor, such as Spearman and Holzinger’s models (Spearman, 1904, 1923, 1927; Holzinger, 1934, 1935), the mature form of Vernon’s ABC theory, (1950/1961), and Horn and Cattell’s Gf-Gc models (Horn, 1965, 1976, 1994), which culminate in Carroll’s (1993) three-stratum model. In the top stratum of Carroll’s model is the general intelligence factor (“*g*”, also known as “general cognitive ability”). General cognitive ability was first described by Spearman (1904) to account for a large amount of variance in mental test scores (Kamphaus, Winsor, Rowe, & Kim, 2005), as evidenced by what industrial-organizational psychologists call “positive manifold” [by which they mean the consistently positive correlations among all mental tests (e.g., Goldstein, Scherbaum, & Yusko, 2010)].¹

However, even among those who advocate for models with a general factor, this factor’s relationship with “intelligence” remains unclear, further complicating a broad debate. Some have tried to draw a distinction between general cognitive ability (GCA) and intelligence. Schmidt (2002) laments industrial-organizational psychologists’ equivalent uses of the terms “intelligence” and “GCA”, which he considers bad public relations, due to a common lay belief that intelligence

¹ Spearman (1904) used “positive manifold” to refer to vanishing tetrad differences, which cognitive tests approximate but do not fully achieve.

is genetic potential. Eysenck (1990/1997) argues that psychologists mean three different yet related things when they use the term “intelligence”: psychometric intelligence (GCA), biological intelligence and social intelligence (or “practical intelligence”). Other proponents of *g* theories have disregarded a priori ideas of intelligence, instead describing intellect in terms of cognitive abilities (e.g., Spearman, 1904, pp. 249-250) and their overall factor structure (e.g., Carroll, 1993; Horn & Cattell, 1967; D. Lubinski, personal communication, July 23, 2003). Some scholars avoid using “intelligence” in scientific contexts, instead referring only to GCA, which can be expressed mathematically (e.g., D. Detterman, personal communication, August 23, 2002).

Regardless of their beliefs about the relationship between GCA and intelligence, scholars frequently use the terms interchangeably (Deary, Penke, & Johnson, 2010), including even those who draw a distinction between them. For instance, in Schmidt and Hunter (1998), the phrase “[GMA] (i.e., intelligence or general cognitive ability” (p. 262). appears. Some view the distinction as trivial, favoring GCA as a working definition of intelligence (Gottfredson, 2002; Jensen, 1980, p. 249).

Whatever the relationship between GCA and intelligence may be, it is fair to state that:

1. Many definitions of “intelligence” have been proposed;
2. Various competing theories of intelligence have also been put forward;
3. General cognitive ability has been implicated in this competition; and,
4. While the factor structure of cognitive ability may be debated, there is a common understanding among scholars that GCA refers to the general factor, for which “positive manifold”² serves as evidence.

² See footnote 1, page 4

While Sternberg and Berg (1986) compare historical and modern definitions of intelligence, it appears that no one has advanced a framework for debate. As previously noted, debate requires clear definitions of terminology, of topics of discussion, and of positions on those topics. Thus, the competition between theories of intelligence cannot be a “debate,” *qua definitione*. In fitting with Kuhn (1962/1970), this competition might better be regarded as paradigmatic in nature. Assuming that the present era is one of normal science—a bold assumption, perhaps, given the potential for disconfirmatory evidence from other fields [e.g., neuropsychology (e.g., Hampshire, Highfield, Parkin, & Owen, 2012)]—the application of Kuhn’s desiderata as a framework for research may guide the field of industrial-organizational psychology (and perhaps the broader field of psychology) toward theory choice, even in the absence of formal debate. While paradigmatic analysis is beyond the bounds of this study, it would seem to be worthy of an undertaking.

1.2 The General Cognitive Ability (GCA) and GCA Testing Debate

From their origins to the present, general cognitive ability (GCA) and GCA testing have aroused controversy. Outside of the field of psychology, many point to a “dark past” (Horn, 1993, p. 167) in which GCA testing was advanced as part of the Nazi eugenic agenda (Mehler, 1997). Some industrial-organizational psychologists, too, believe that the past continues to haunt GCA testing. Viswesvaran and Ones (2002) believe that the association between GCA tests and notorious social movements has “stigmatized the societal view of [GCA]” (p. 221), undermining GCA as a source of equal opportunity. Others (e.g., Outtz, 2002) assert that the huge race differences produced by GCA tests relative to other selection tools have made them controversial. In contrast, Schmidt (2002) suggests that, even in the absence of adverse impact, those who believe

that a variety of traits all make major contributions to success would continue to reject research findings on GCA. Salgado and Anderson's (2002) observation that negative reactions to GCA occur even in countries where adverse impact is not an issue may support Schmidt's claim. Tenopir (2002) raises another possible reason for skepticism: To laypeople, GCA does not appear to account fully for intelligence.

Controversy makes for contention. Taylor (1991) states, "There is perhaps no other assessment issue as heated, controversial, and frequently debated as that of bias in cognitive assessment" (p. 3). While a consensus theoretical definition of general cognitive ability has not been reached, a term may be considered "adequately defined" when those who use it know and agree upon the referent (Pedhazur & Schmelkin, 1993, p. 167). Such is the case for GCA. Further, formal debate on GCA and GCA testing should be possible by virtue of the common understanding of the evidence for its existence and of typical empirical definitions.

A primary goal of this study is to examine the dimensionality of prior informal debate and to provide an empirically-tested model to serve as a framework for both formal and informal future debate wherein the factors represent beliefs. A search has identified only one study that advanced such a model (i.e., Murphy, Cronin, & Tam, 2003). In this study, Murphy et al. (2003) propose two general themes of the ongoing debate on cognitive ability testing. These themes are called "*Societal Concerns*" (p. 660) and the "*Primacy or Uniqueness of General Cognitive Ability*" (p. 661). Among societal concerns, the authors mention the unfair "substantial adverse impact" (p. 660) on minority subgroups that may result from the use of GCA tests. The issue of a tradeoff between fairness to groups and efficiency, too, is characterized as a societal concern. The authors touch on test bias, which may occur when, as Murphy et al. (2003) put it, tests "overestimate performance differences between groups (i.e., because groups differ on the test more than they do

on performance)” (p. 661). They also note the potential for test bias to create the racial stratification of society. Among concerns of primacy or uniqueness, the authors mention the “the validity and utility of cognitive ability tests” (p. 661) and the importance of other cognitive abilities.

Review of the methods and results of the Murphy, Cronin, & Tam (2002) study has inspired a different direction for research in the present study. In recapitulation, the researchers conducted a survey on beliefs about the role of cognitive ability and cognitive ability testing in organizations. Two sources of items were used: “The Millennial Debate on ‘*g*’ in I-O Psychology” (Ones & Viswesvaran, 2000), and “The Role of General Mental Ability in Industrial, Work, and Organizational Psychology” (which appeared in a special double issue of *Human Performance*, 2002), the latter of which was composed of articles by Millennial Debate participants (as well as by several other scholars) that delineated their positions on general cognitive ability and GCA tests. The researchers identified a total of 275 assertions, which were narrowed to 49 categorical exemplars. These exemplars were then stated as survey items. Murphy et al. (2003) note that these items were sorted into five general topics: “(1) the importance or uniqueness of cognitive ability tests, (2) how well these tests measure cognitive ability, (3) links between ability tests and job requirements, (4) alternatives or additions to ability testing, and (5) the societal impact of cognitive ability testing” (p. 662). An online survey tool was constructed that asked respondents to express levels of agreement or disagreement with these 49 item statements on a five-point Likert scale. Respondents were asked also to indicate their age, gender, race, and primary employment, as well as the racial makeup of their workplaces. Members and fellows of the Society for Industrial and Organizational Psychology (SIOP) were invited (by email) to participate. 703 completed surveys were submitted. Murphy et al. (2003) used the data from these responses to determine the sample’s degree of consensus or controversy on the 49 belief items, as well as to test a hypothetical

two-factor model of beliefs based on the two sets of issues outlined above (i.e., “*Societal Concerns*” and “*Primacy or Uniqueness of General Cognitive Ability*”). The results from testing the model—results that were the impetus for the present investigation—showed good fit (i.e., “goodness-of-fit index [GFI] = .920, root-mean-square error analysis [sic] [RMSEA] = .034”), in accordance with authoritative sources on “fit” (e.g., Hu & Bentler, 1999). Further, this two-factor model appeared to be parsimonious, with the Parsimony Goodness-of-Fit Index [PGFI] = .85. 46 of 49 hypothesized loadings were statistically non-zero (at $p < .05$). The factor pattern held for males and females, as well as for academics and nonacademics. Item factor loadings and $R^2_{factors}$ were reported. Eighteen of the R^2 values were .25 or larger, which the researchers claimed typically are considered large. However, the factors accounted for a small to moderate amount of the variance in responses to 31 items (where $0.10 \leq R^2 < 0.25$ was considered a conventional indication of a moderate relationship.) To the researchers, this suggested that the model was insufficient for accounting for all of the variance in the data. Still, they state that the model “provide[s] a simple and straightforward way of conceptualizing responses to this survey that is highly consistent with existing literature over the role of cognitive ability testing” (Murphy et al., 2003, p. 667).

After examining the methods and the factor loadings derived from confirmatory factor analysis, it was concluded that present study should be undertaken, for several reasons. First, although model fit and parsimony were reasonably good, the reported factor loadings were quite low. Given the large size of the sample and the relatively relaxed requirements of the field of industrial-organizational psychology, I compared Murphy et al.’s factor loadings against a liberal heuristic (i.e., retain only loadings ≥ 0.3 or ≤ -0.3) relative to the requirements of other sciences (typically, retain only loadings ≥ 0.7 or ≤ -0.7). Only ten out of 49 Murphy et al.’s (2003) items achieved loadings sufficient to retain, only two of which are ≥ 0.4 or ≤ -0.40 . Also, under general

circumstances, the factor loadings and R^2 values should be functions of each other. In Murphy et al. (2003), they are tremendously discrepant. The moderate to high R^2 and correspondingly low loadings indicate that the variance of the factors was enormous relative to the variance of the items and error, which is unusual. Taken in sum, the results (i.e., model fit and many loadings) indicate that the model is measuring two factors, but the individual items are not informative as to the underlying nature of the constructs.

Second, Murphy et al.'s (2003) approach may not have been ideal for deriving a model for debate. The scholars based their hypothetical model on their a priori knowledge of historical positions on GCA and GCA testing, rather using a deductive approach to generate this model (e.g., examining an *actual* debate). This seems odd, given that they had sorted the questionnaire's 49 items—items that were based on opinions from two informal debates—into “five broad groups”, which might have served as factors for the model. A potential five-factor solution was not mentioned in the article. Moreover, per se, historical beliefs cannot be construed properly as debate, since debates take place within a specific context with a specific audience (Ehninger & Brockriede, 2008). What Murphy et al. (2003) considered debate was, in truth, their own representation of *historical beliefs that were neither necessarily offered in opposition to, nor in response to, others' beliefs, and without a specific context and audience*. For the purposes of this study, an effort was made to arrive deductively at a model. The data from the Murphy et al. (2003) survey was deemed appropriate for examining a model, as it reflects respondents' agreement or disagreement with specific assertions in a specific context and for a specific audience, capturing the essence of debate (if not meeting the definitive requirements of formal debate.)

CHAPTER 2

THE PRESENT STUDY

2.1 Introduction to the Six-Factor Model

This study builds upon Murphy et al.'s (2003) groundbreaking research by exploring a more nuanced model that could be more useful for debate.

Through reviewing the aforementioned assertions from the special issue of *Human Performance* (2002), ultimately, a six-factor model was deduced. The articles in the “special issue” were considered to represent informal debate satisfactorily, as they summarize previous debate positions and address opposing arguments in a single forum. [Note: The deductive process employed in this study is elaborated later, in the Methods section.] The six factors were conceived of as “Primacy,” “Criterion Dependence,” “Alternatives,” “Test Bias,” “Social Fairness” and “Tradeoffs”. In following subsections, each belief factor is defined. Opposing positions are presented, as well as evidence for these beliefs. While historical beliefs were *not* considered during the deductive process for identifying the model, some of these beliefs are now placed *post hoc* on the respective factors. The beliefs held by the special issue scholars are also included. At the end of the “The Present Study” section, a table is provided that shows areas in which these scholars agree and disagree, to suggest one practical use of the model.

2.1.1 Primacy

Primacy: General cognitive ability is the most important trait predictor of performance. Some studies have argued that general cognitive ability (GCA) should hold primary position

among trait predictors of performance (e.g., Schmidt & Hunter, 1998). In the special issue of *Human Performance* (2002), seven out of twelve articles—those written by Gottfredson (2002), Murphy (2002), Ree and Carretta (2002), Salgado and Anderson (2002), Schmidt (2002), Viswesvaran and Ones (2002), and Reeve and Hakel (2002)—tend to support this argument. As evidence, all seven claim that no other predictor matches its efficiency. Indeed, many studies have advocated for the high predictive validity of GCA in predicting training outcomes (e.g., Hunter, 1980c; Jensen, 1996; Levine, Spector, Menon, Narayanan, & Cannon-Bowers, 1996; Ree & Earles, 1991b; Thorndike, 1986) and job performance (Hunter & Hunter, 1984; Pearlman, Schmidt, & Hunter, 1980; Schmitt, Gooding, Noe, & Kirsch, 1984; Schmidt & Hunter, 1998). Also, support is offered for the belief that the predictive validity of GCA generalizes across job families, with variance largely explained by artifactual errors, such as sampling error (Hartigan & Wigdor, 1989; Hunter & Hunter, 1984; Levine et al., 1996; McDaniel, Schmidt, & Hunter, 1988b; Pearlman et al., 1980; Schmitt et al., 1984; Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman, & Shane, 1979). Schmidt (2002), Murphy (2002), and Gottfredson (2002) also point out that, where differences in coefficients are observed, they are lawfully moderated by job complexity, an assertion put forth by various researchers (e.g., Hartigan & Wigdor, 1989; Hunter, 1983a; Hunter & Hunter, 1984; Hunter, Schmidt, & Le, 2006). Further, Salgado and Anderson (2002) present evidence of generalization not only across job families but also across geographic and cultural boundaries.

Given their beliefs in the predictive validity and generalizability of GCA, several scholars indicate that it is the *best* predictor of performance (e.g., Gottfredson, 2002; Murphy, 2002). Gottfredson (2002) declares that only the factor that represents conscientiousness and integrity is as generalizable, but that it tends not to influence performance as much as GCA. Schmidt and

Hunter's (1998) oft-cited summary of research has been held up as support for this belief. While Hunter and Schmidt's (1998) summary suggests that tests of job knowledge may be nearly as valid as general cognitive ability, the authors remark that these tests may simply serve as proxy measures for GCA, given that cognitive ability is instrumental in acquiring knowledge. Schmidt (2002) also claims that GCA has a direct effect on performance, and is not merely mediated by job knowledge, citing the results of Hunter and Schmidt (1996) and Schmidt and Hunter (1992) for support. Indeed, several articles in the special issue suggest that the value of other predictors may be gauged by their incremental predictive validity over GCA (e.g., Gottfredson, 2002; Schmidt, 2002). Implicitly, they accord to GCA the primary position among predictors. Further demonstrating its importance, Le, Oh, Shaffer and Schmidt (2007) say that GCA is three times more effective in predicting performance than conscientiousness. Murphy (2002) sums up the efficiency argument well: "[T]here is a large body of research showing that cognitive ability is unique in terms of its relevance for predicting performance on a wide range of jobs. No other single measure appears to work so well, in such a range of circumstances, as a predictor of overall job performance as a good measure of general cognitive ability" (p. 174). According to Schmidt (2002), so valuable is GCA testing that it not only offers great utility to organizations but also has broader ramifications for the economy (Hunter & Schmidt, 1996).

Other authors question whether general cognitive ability should be granted special status. Goldstein et al. (2002), Sternberg and Hedlund (2002) and Tenopir (2002) all suggest that GCA may not account adequately for all intelligent behavior. Goldstein et al. (2002) further argue that the nature and definition of GCA itself remain unclear, and that the industrial-organizational psychology community cannot afford to adopt a "case closed" (pp. 123) position until alternative predictors have been thoroughly examined. Several authors question the efficiency of GCA as a

predictor. Both Goldstein et al. (2002) and Sternberg and Hedlund (2002) believe that GCA leaves a large amount of variance in job performance unaccounted for; as evidence, they reference results from the very studies that others use in support for primacy (e.g., Schmidt and Hunter, 1998). Thus, it appears that validity data are highly subject to interpretation. Goldstein et al. (2002) further speculate that existing validity estimates may be inflated due to the cognitive overloading of selection systems, with similar kinds of testing on both predictor and criterion (e.g., training performance) potentially leading to common method bias and, in turn, inflated validity estimates. Goldstein et al. (2002) and Outtz (2002) also note that the validity coefficient in operational use in organizations is considerably lower than the corrected coefficient. Some authors also question the generalizability of GCA. Goldstein et al. (2002) observe that GCA does not generalize particularly well to interactive fields, such as sales and law enforcement. Sternberg and Hedlund (2002) believe that GCA may not predict managerial performance as well as other factors do, given the importance of specific procedural knowledge. Finally, doubt about the utility of GCA is expressed. Outtz (2002) wonders how extravagant, speculative claims of utility can be used in support of a predictor with fairly low validity, particularly when the adverse impact it causes is tangible. Kehoe (2002) is suspicious of the value of GCA tests in organizations, given that GCA testing alone may offer little incremental predictive validity over other approaches that yield less adverse impact.

Interestingly, among all of the articles in the special issue, only one (i.e., Reeve & Hakel, 2002) emphasizes the necessity yet insufficiency of GCA for performance in organizations. At first glance, this appears to be a middle-ground position; fundamentally, however, it is but a nuanced presentation of the pro-primacy position, with other factors cast as supplementary to GCA. Still, it is not entirely different than the argument Goldstein et al. (2002) make by invoking Vroom's formulation of performance [i.e., $P = F(M \times A)$] (Vroom, 1964).

2.1.1.1 Criterion Dependence

Criterion dependence: The predictive validity of general cognitive ability depends on the criterion of interest.

Scholars generally seem to recognize that cognitive and noncognitive factors have special influence in different domains of work behavior. While the importance of voluntary helping behavior in organizations has long been recognized (e.g., Katz, 1964), more recently, scholars have defined the dimensions of this behavior (Bateman & Organ, 1984; Borman, Motowidlo, & Hanser, 1983; Dozier & Miceli, 1985; Mowday, Porter & Steers, 1982; Organ, 1977; Smith, Organ, & Neer, 1983; Staw, 1984). For Example, Organ (1988) refers to these behaviors collectively as “organizational citizenship behavior” (OCB); Brief and Motowidlo (1986) call them “prosocial organizational behavior” (POB). More recently, Borman and his colleagues (Borman, Hanson, & Hedge, 1997; Borman & Motowidlo, 1993b, 1997; Motowidlo, Borman, & Schmit, 1997) have identified a broad domain called “contextual performance,” which includes a wider range behaviors that pertain to the social and psychological context of the organization yet do not contribute directly to core performance.

Various studies assert that noncognitive factors predict non-core performance behavior better than general cognitive ability (GCA) (e.g., Arvey & Murphy, 1998; Borman et al., 1997; Borman & Motowidlo, 1997; Borman, Penner, Allen, & Motowidlo, 2001; Campbell, McHenry & Wise, 1990; Crawley, Pinder & Herriot, 1990; Hatrup & Jackson, 1996; Day & Silverman, 1989; McHenry, Hough, Toquam, Hanson & Ashworth, 1990; Motowidlo & Van Scotter, 1994) or its direct consequences (e.g., experience, job knowledge) do (e.g., Borman, Motowidlo & Schmit, 1997). Among personality traits, conscientiousness is believed to correlate highest with

citizenship behavior in a number of reports (e.g., Hogan, Rybicki, Motowidlo & Borman, 1998; Miller, Griffin & Hart, 1999; Neuman & Kickul, 1998; Organ & Ryan, 1995), though agreeableness and positive affectivity, too, are said to share significant positive relationships with OCB (Organ & Ryan, 1995). For some interactive occupations where the relationship between GCA and performance may be relatively weak (Hirsh, Northrop & Schmidt, 1986), it is suggested that personality variables may be more influential in overall job success (Goldstein, 2002). Even among scholars who believe in the pervasive influence of GCA, some feel that non-cognitive factors may serve as superior predictors in non-core domains (e.g., Ree & Carretta, 2002). Some also recognize that the GCA-linked “complexity factor” is not always the most influential in all job attributes (e.g., Gottfredson, 2002). However, they may still doubt the legitimacy of a separate “contextual performance” domain, considering it to be a catch-all for behaviors that defy specification in a job description (Schmidt, 1993). Instead, there are those who contend that the criterion space should be divided by the level of performance that is required. For example, they might state that cognitive ability predicts maximum performance (i.e., “can do”) best (Cronbach, 1949; DuBois, Sackett, Zedeck & Fogli, 1993) and that non-cognitive factors best predict typical performance (i.e., “will do”) (Gottfredson, 2002).

In the special double issue of *Human Performance* (2002), ten out of twelve articles (i.e., Goldstein et al., 2002; Gottfredson, 2002; Kehoe, 2002; Outtz, 2002; Murphy, 2002; Ree and Carretta, 2002; Reeve & Hakel, 2002; Sternberg & Hedlund, 2002; Tenopir, 2002) specifically acknowledge criterion dependence, mostly by focusing on the difference in the predictive validity of GCA and noncognitive predictors for core and contextual performance. In fact, among these ten articles, only three (i.e., Sternberg and Hedlund, 2002; Ree and Carretta, 2002; Schmidt, 2002) make no distinction between these domains of performance. The importance of motivation for

predicting typical performance (Tenopyr, 2002) and sustained performance (Goldstein et al., 2002) is also mentioned. Construct definition and measurement issues that may affect the predictive validity of GCA are raised (i.e., by Goldstein et al., 2002; Gottfredson, 2002; Outtz, 2002; Reeve & Hakel, 2002; Tenopyr, 2002). For example, Gottfredson (2002) notes that GCA predicts performance measured objectively (e.g., by job knowledge or work samples) better than performance measured subjectively (e.g., by supervisor ratings). Only Salgado and Anderson (2002) and Viswesvaran and Ones (2002) take no positions whatsoever on criterion dependence. Even Schmidt (2002) notes a case in which the measurement of criteria may affect GCA validity. With near consensus among the authors—as well as among many members and fellows of SIOP (Murphy et al., 2003)—“criterion dependence” may be a settled matter.

However, one difference between the authors is the emphasis which they place upon studying differences in criterion-specific predictive validity. Where Schmidt (2002), Viswesvaran and Ones (2002), and Ree and Carretta (2002) seem to have little interest, other authors suggest that further research should be pursued. Also, four articles point out the potential value of as-yet-ignored outcome variables. Goldstein et al. (2002) suggest that all current performance criteria are at the individual level of analysis, and that the relationships between GCA and group- and organizational-level variables remain unclear. Kehoe (2002) speculates that “overall worth or merit” (p. 105) to the organization may be an interesting criterion. Finally, Outtz (2002) and Murphy (2002) observe that diversity and fairness may be important to organizations. To those who believe that GCA tests produce adverse impact (e.g., Goldstein et al., 2002; Kehoe, 2002; Murphy, 2002; Outtz, 2002), the validity of GCA as a predictor of diversity and group-level equity would be expected to be quite different than its validity in predicting traditional efficiency criteria.

In future debate, the relationship between GCA and unusual criteria might be discussed, where the broader issue of criterion dependence should probably be disregarded.

2.1.1.1.1 Alternatives

Alternatives: There are cognitive alternatives to general cognitive ability that are significant determinants of intelligent behavior.

Not all scholars agree that general cognitive ability (GCA) is the only, or even primary, cognitive determinant of performance. Some even question construct validity altogether. Guilford (1956, 1967, 1977, 1985), Gardner (1983/2003) and Das and his colleagues (e.g., Das, Kirby & Jarman, 1975; Das, Naglieri & Kirby, 1994) have developed theories of interactive cognitive processes or separate kinds of intelligence. Sternberg acknowledges the existence of GCA, but it is not a unifying factor in his model and is deemphasized. Sternberg's triarchic theory of human intelligence (Sternberg, 1985, 1988, 1996, 1997, 1999) posits three broad cognitive factors—analytical intelligence (Sternberg's equivalent of GCA), practical intelligence and creative intelligence—which are used jointly to achieve successful intelligence (Sternberg, 1997). Sternberg claims that practical intelligence enables one to acquire tacit knowledge (TK), an unarticulated form of procedural knowledge that can be gained without formal instruction and is useful in solving everyday problems (Wagner, 1987; Sternberg & Hedlund, 2002). Sternberg and his colleagues cite studies that may indicate correlations between TK and promotion (Sternberg, 1997), managerial performance, and success in academe (Wagner & Sternberg, 1985), and they claim that TK accounts for substantial variance over and above that for which GCA accounts (Sternberg & Hedlund, 2002). Further, they state that, in one study (i.e., Sternberg, et al., 2001), a negative correlation was found between TK and GCA.

Much early intelligence research placed greater emphasis on the influence of specific cognitive abilities (e.g., Burt, 1949; Hull, 1928; Kelley, 1928; Thorndike, 1921; Thurstone, 1938) than on Spearman's general factor. Some of these were proposed as intelligences or "primary mental abilities" in theories that rejected GCA (e.g., Thurstone, 1938), only to be reconciled with it later (e.g., Thurstone, 1947). In several cases, what were once considered to be uncorrelated factors were incorporated into Carroll's (1993) hierarchical model, where GCA is the single factor in the highest stratum. Other early intelligence scholars (e.g., Hull, 1928) recognized GCA but believed that specific abilities could compensate, or even substitute, for it. Even Spearman (1927) believed that positive manifold might be lower in high-ability individuals, allowing their strongest abilities to exert disproportionate influence in performance. Several researchers claim empirical support for changes in positive manifold across ability levels (e.g., Detterman & Daniel, 1989; Legree, Pifer, & Grafton, 1996). According to some, these findings may have practical implications for academic and vocational interest counseling (Dawis, 1992) as well as for personnel classification (Murphy, 1996), and particularly for differential assignment theory (DAT) (Scholarios, Johnson, & Zeidner, 1994; Zeidner & Johnson, 1994; Zeidner, Johnson, & Scholarios, 1997). Some scholars also believe that tests should draw on abilities that are relevant to the job (e.g., Goldstein, Yusko, Braverman, Smith, & Chung, 1998).

Emotional intelligence (EI), a potential extra-*g* intellect, is defined by Mayer, Salovey, & Caruso (2004) as "the capacity to reason about emotions, and of emotions to enhance thinking" (p. 197). Their four-branch model of EI includes the abilities to: 1) perceive emotion; 2) use emotion to enable thought; 3) understand emotions; and, 4) manage emotion (Mayer & Salovey, 1997). They believe that EI overlaps with Gardner's "intrapersonal intelligence" factor but can be distinguished both from social intelligence and GCA (Mayer & Salovey, 1993). Mayer, Caruso

and Salovey (1999) attempt to show that EI could be operationalized as a set of intercorrelated, measurable abilities that capture unique variance. EI has captured public attention (Mayer, Salovey & Caruso, 2004), touted as important for competencies relevant “in almost any job” (Chemiss, 2000, p. 10) and potentially twice as important as IQ (Goleman, 1998). Until recently, evidence for EI seems to have been limited. A study by Joseph and Newman (2010) supports a cascading model of EI through meta-analysis and provides evidence for the incremental validity of EI over Big Five personality traits and cognitive ability (and particularly when self-report mixed measures of EI are utilized). Recent research by MacCann, Joseph, Newman and Roberts (2014) suggests that EI is not independent of *g* but rather is a second stratum factor in the CHC model.

By contrast, classicists reject extra-GCA intelligences theories, often declaring that they simply identify abilities similar to factors in hierarchical models [e.g., Carroll’s (1993) critique of Gardner’s intelligences]. They note that multiple intelligence theories have come and gone (e.g., Guilford, 1956) when empirical verification was applied and that MI test scores correlate highly with GCA scores (Brand, 1996; Kuncel, Hezlett, & Ones, 2004). Visser, Ashton and Vernon (2006) found that GCA correlates highly with Gardner’s cognitive intelligences but shares a weaker positive relationship with “intelligences” that appear to represent motor, sensory or purely non-cognitive factors (e.g., personality). Classicists also believe that Gardner and his colleagues’ rejection of psychometrics—coupled with their failure to link the intelligences or define their components—has made it difficult to create valid MI measures (Allix, 2000; Waterhouse, 2006). According to Waterhouse (2006), while cognitive science and neuroscience have been enthusiastic about GCA, there has been no interest in MI theory (contrary to claims by Gardner, 2004).

Sternberg’s theory of practical intelligence (PI) has also attracted skeptics, with several asserting that it is encompassed by GCA (Gottfredson, 1988; Messick, 1992), that the “tacit

knowledge” it predicts is essentially redundant with job knowledge (Tenopyr, 2002), and that job knowledge already has well-researched relations with other constructs (e.g., job performance) (Schmidt & Hunter, 1993). Classicists also charge Sternberg and his colleagues with committing a host of methodological errors, such as the use of anecdotal evidence, selective reporting, small sample sizes, and highly range restricted participants, as well as ignoring the lack of discriminant validity between its three intelligences, and failing to account for differences in experience (Gottfredson, 2002, 2003; Jensen, 1993; Kuncel, Campbell, & Ones, 1998; Kuncel, Hezlett & Ones, 2001; Ree & Earles, 1993; Viswesvaran and Ones, 2002). They also challenge Sternberg’s claims of strong empirical support, noting the small number of studies on PI/TK (Gottfredson, 2002; Jensen, 1993; Ree & Earles, 1993; Schmidt & Hunter, 1993). Gottfredson (2003) offers especially trenchant criticism, accusing Sternberg and colleagues of hypothesizing after results are known, by using low intercorrelations as evidence of the “domain specificity” of TK tests and using high intercorrelations as evidence that they measure a common factor (i.e., practical intelligence).

Concerns regarding emotional intelligence come not only from classicists, but also EI theorists themselves. Mayer and his colleagues note that the myriad conceptualizations and definitions of emotional intelligence (Mayer, Caruso & Salovey, 1999) and how the influence of popular culture has shaped the construct (Mayer, Salovey & Caruso, 2004). Critics observe these same issues, leading them to look upon EI as an “elusive concept” (Davies, Stankov & Roberts, 1998, p. 989), resistant to measurement (Becker, 2003) or too broad to be tested (Locke, 1995). Scholars frequently point out fundamental problems, such as poor discriminant validity among EI constructs (Lopes, Salovey, Cote, & Beers, 2005), and the inadequacy of EI measures to tap into anything more than personality and general intelligence (Davies et al., 1998; Matthews, Zeidner,

& Roberts, 2005; Schulte, Ree & Carretta, 2004). Despite the claims of Goleman and others (Waterhouse, 2006), some believe that the evidence of predictive validity is scant, particularly for real-world success, one of the key claims of EI research (Waterhouse, 2006). In fact, Collins (2001) has found that EI competencies account for no variance in job success over cognitive ability and personality. Barchard (2003) has shown evidence that EI does predict academic success but not incrementally over other traits. Critics also state that there is no supporting evidence from cognitive psychology, neuroscience or other psychology disciplines (Matthews, Zeidner, & Roberts, 2002). Rather, in their opinions, EI studies rely on anecdotal evidence and self-report (Zeidner Matthews, & Roberts, 2004). Landy (2005) noted that (at the time) much of the most critical EI data was held in privately-owned databases and could not be examined. Others believe that publicly-available data has shown little evidence that EI is a single construct or set of specific abilities (e.g., Waterhouse, 2006). This has led some to conclude that EI is more myth than science (Matthews, Zeidner, & Roberts, 2002). Notably, however, most criticism of EI was made prior to recent empirical studies that offer stronger support (e.g., Joseph and Newman, 2010).

Finally, classicists cite many studies that have shown that specific abilities add negligible predictive validity over GCA (e.g., Besetsny, Earles, & Ree, 1993; Besetsny, Ree, & Earles, 1993; Hunter, 1983b; Olea & Ree, 1994; Ree & Earles, 1991a, 1991b, 1992; Ree, Earles, & Teachout, 1994; Thorndike, 1986). Recent research by cognitive psychologists claims to support these findings (e.g., Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004). Some have presented evidence that certain aptitudes are almost perfectly predicted by GCA (e.g., working memory: Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004).

The positions of most special issue authors are fairly clear. Goldstein et al. (2002) refer to GCA as a “general composite score,” (p. 128), without any clear meaning in and of itself. They

encourage the investigation of unorthodox theories such as multiple intelligences, emotional intelligence and practical intelligence, as part of the ongoing effort to explore alternatives to GCA. Sternberg and Hedlund (2002) stand behind their theory of practical intelligence and tacit knowledge. While Tenopir (2002) is skeptical of tacit knowledge, due to construct definition problems, she acknowledges that Sternberg's work holds some promise. Reeve and Hakel (2002) contend that specific abilities account for unique, reliable variance and may be useful in predicting specific performance criteria. Kehoe (2002) suggests that testing specific abilities may offer advantages in decreasing group differences while maintaining predictive validity, though he warns that this approach may result in differential prediction.

Gottfredson (2002), Murphy (2002), Ree and Carretta (2002), and Viswesvaran and Ones (2002) all consider the evidence for multiple intelligences to be weak. Gottfredson (2002) states that there is little evidence for practical intelligence, appearing in only six studies (several of which were unpublished at the time of the special issue) in only five job families. She considers practical intelligence to be little more than an expression of experience in some domain. Gottfredson (2002), Murphy (2002), Ree and Carretta (2002) and Schmidt (2002) believe that specific abilities offer little predictive value over GCA. Schmidt (2002) points out an important theoretical reason: Given positive manifold, the more aptitudes that are tested, the more that GCA is represented.

In the special issue, Only Outtz (2002) and Salgado and Anderson (2002) refrain entirely from stating their positions on alternative cognitive predictors.

2.1.1.1.1.1 Test Bias

Test Bias: Cognitive ability tests are biased.

Scholars who believe that general cognitive ability (GCA) tests are unbiased acknowledge that mean scores differ across groups, but they attribute disparity to real differences in ability

(Gottfredson, 1988; Gottfredson, 2002; Hunter & Hunter, 1984; Jensen, 1980; Kuncel & Hezlett, 2010; Neisser et al., 1996; Schmidt, 1988; Schmidt, 2002). They reject the view that GCA tests are subject to measurement bias, that they unfairly favor Whites (Jensen, 1980, 1987; Reeve & Hakel, 2002; Reynolds & Jensen, 1983) and that they *create* group differences, rather than simply measuring real group differences. Findings from item response theory may support this position. While studies have identified differential item functioning (DIF) across subgroups, effect sizes have been found to be small and with no aggregate bias across items (Drasgow, 1987; Sackett, Schmitt, Ellingson and Kabin, 2001). Sackett, Schmitt, Ellingson and Kabin (2001) consider many DIF studies fundamentally flawed (e.g., Medley and Quirk, 1974) or inconclusive (e.g., Scheuneman & Gerritz, 1990; Schmitt & Dorans, 1990). Hough, Oswald & Ployhart (2001) also cite several studies that show that Black-White DIF on general cognitive ability tests may favor either group, is not always found for hypothetically culture-linked items, and does not always favor the “right group” for a given presumed culture-related item (e.g., Raju, van der Linden, & Fler, 1995; Waller, Thompson, & Wenk, 2000). Further, they state that attempts to create general cognitive ability test items related to both Black and White culture or to remove content potentially biased against Blacks have not narrowed mean differences between groups (e.g., DeShon, Smith, Chan & Schmitt, 1988; Hunter & Hunter, 1984; Jensen and McGurk, 1987). Some scholars believe that “culture-free” tests have reduced differences but have not eliminated them (Vernon & Jensen, 1984). Sackett, Schmitt, Ellingson and Kabin (2001) observe that when studies designed to test whether group differences can be reduced by changes in test format have appeared to succeed, they may have measured constructs that are not heavily GCA loaded (e.g., Chan & Schmitt, 1997). Finally, equating groups on job performance instead of overall test score when analyzing item

passing rates has been discredited as an attempt at using an approach similar to performance-based fairness models (Schmitt, Hattrup, & Landis, 1993).

Predictive bias has also been scrutinized. Under Cleary's model (Cleary, 1968; Cleary & Hilton, 1968), bias between test and criterion scores is established when group regression lines are not identical. The conventional belief among industrial-organizational psychologists is that general cognitive ability accurately predicts meaningful race differences on measures of job performance (Bobko, Roth, & Potosky, 1999; Hartigan & Wigdor, 1989; Hunter, 1981b; Jensen, 1980; Neisser et al., 1996; Schmidt, 1988; Schmidt & Hunter, 1999; Schmitt, Clause, & Pulakos, 1996; Wigdor & Garner, 1982). Many researchers feel that GCA tests do not underpredict for minority group performance (e.g., Jensen, 1980; Neisser et al., 1996; Sackett & Wilk, 1994). Some even believe that the regression line intercept tend to be lower for Blacks and Hispanics. Thus, while they may technically be Cleary-unfair, differences in slopes are minor and GCA may slightly underpredict for the majority group (Bartlett, Bobko, Mosier, & Hannan, 1978; Gottfredson, 1986b; Hartigan & Wigdor, 1989; Houston & Novick, 1987; Humphreys, 1986; Hunter & Hunter, 1984; Hunter, Schmidt & Rauschenberger, 1984; Kuncel & Sackett, 2007; Linn, 1978; Schmidt, 1988; Rotundo & Sackett, 1999; Rushton & Jensen, 2005; Sackett, Schmitt, Ellingson Kabin, 2001; Sackett & Wilk, 1994; Schmidt & Hunter, 1981, 1998; Schmidt, Pearlman, Hunter, 1980). Hunter and Schmidt (1976) also point out that, while an unreliable test still might seem biased against Blacks, in fact, "the bias created by random error works against more applicants in the white group" (p. 1056). Hunter and Schmidt (1977) point out that, as GCA is only one determinant of performance, the difference on the predictor should be greater than on the criterion. Those who believe that GCA is predictively unbiased believe that, with a 1.0 SD Black-White difference in GCA, the .51 validity coefficient properly predicts the 0.50 SD B-W difference on measures of

job performance (Hunter & Hunter; 1984; Schmidt, 2002). Further, some say that Blacks and Hispanics who score low perform poorly, no differently than low-scoring Whites (Bartlett et al., 1978; Campbell, Crooks, Mahoney, & Rock, 1973; Gael & Grant, 1972; Gael, Grant, & Ritchie, 1975a, 1975b; Grant & Bray, 1970; Jensen, 1980; Schmidt, Pearlman, & Hunter, 1980; Schmidt, 1988; Schmidt, 2002).

Even scholars who believe that GCA tests are unbiased seem to admit that selection based on actual performance would promote more diversity than selection based on general cognitive ability scores. The one standard deviation B-W mean difference across job families on the predictor (Herrnstein & Murray, 1994; Hunter and Hunter, 1984; Jensen, 1980; Loehlin, Lindzey and Spuhler, 1975; Neisser et al. 1996; Reynolds, Chastain, Kaufman, & McLean, 1987; Sackett & Wilk, 1994; Williams & Ceci, 1997) has been shown to be much larger than even the largest differences found on measures of job performance (Hunter & Hirsh, 1987; Hunter & Hunter; 1984; Schmidt, 2002; Schmidt and Hunter, 1998). Some consider this disparity between d on the predictor and d on the criterion a moot point: As we do not have access to future performance scores, we must rely on imperfect predictors (e.g., Murphy, 2002). Cole (1973) suggests that there may be a way to overcome this challenge. When historical data with distributions of both test and performance scores are available, conditional probability (i.e., probability of selection given future success) could be used in the form of a ratio of true positives to false negatives that could be applied across subgroups. Hunter & Schmidt (1976) argue against this model and its precursor, Darlington's fairness Definition 3 (1971), on the grounds that they essentially reverse the predictor and criterion, even in concurrent validation studies (Schmitt, Hattrup & Landis, 1993).

Some critics maintain that GCA tests are biased by the nature of the measure. "Culturalists" (Helms, 1992) argue that culture cannot be removed from tests. They believe that Black-relevant

content would result in fairer assessments and, similarly, that acculturation and assimilation might lead to improvement in Black persons' performance on current tests. Functional and linguistic equivalence between cultures is questioned. Thus, it is posited that if GCA functions differently across cultures and cultures communicate differently, yet current tests measure a single shared general intelligence factor divorced from cultural considerations, discrimination results from testing (Helms, 1992). There may be evidence of DIF on verbal subtests, suggesting that they are biased against Black test-takers (Ironson & Suboviak, 1979). It has been hypothesized that, since the written language of a GCA test might favor those who share the culture of its designers, non-verbal measures might be more culture-fair. However, recent research (e.g., Cronshaw, Hamilton, Onyura, & Winston, 2006) is presented as evidence that supposedly culture-free tests are actually strongly biased against Blacks. It is suggested that the role of race in determining test scores is not fully understood. Some scholars even believe that we cannot accurately determine what race differences in intelligence are until we know what intelligence is and can find the source of differences (Sternberg, Grigorenko, & Kidd, 2005). Helms (1993) cautions against the use of GCA tests, which measure what she views as a "hypothetical construct" (p. 1086) based on intercorrelations between tests "of similar content, format and perhaps cultural loadings" (p. 1086). She suggests that the nature of GCA itself might be affected by the kinds of tests used to assess it. Helms (1992) also questions whether non-psychometric issues, such as differences in visible racial and ethnic group (VREG) test-taking strategies, might affect performance on both predictor and criterion (e.g., success in training), inflating the validity of GCA testing.

Darlington (1971) believes that performance on both test and criterion is determined by various abilities. If a test measures abilities irrelevant to performance but on which races have unequal standing, bias exists that may affect selection. In his Definition 3 of test fairness, bias is

assessed by partialing out the criterion and observing whether a non-zero correlation between race and test remains. In the early 1970s, scholars felt that aptitude tests might measure constructs correlated with race but uncorrelated with performance (e.g., Wallace, 1973). Recently, there has been renewed interest in examining GCA tests within frameworks similar to Darlington's. With job performance held constant, Newman, Hanges and Outtz (2007) reported a 0.42 semipartial correlation between race and GCA test score, indicating that, independent of performance, roughly 18% of the variance in GCA test scores can be accounted for by B-W group differences. Outtz and Newman (2010) believe not only that intelligence tests measure ability, but also that part of the variance in measurement can also be accounted for by differences in previously-learned material, in turn accounted for by status and privilege. Indeed, there may be evidence that academic knowledge—a construct some claim is related to socioeconomic status—could be responsible for large mean differences between Black and White test takers (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). A several researchers have created intelligence tests specifically designed to minimize the influence of prior learning as much as possible. These include tests of problem-solving skills and creativity [e.g., Rainbow Project (Sternberg, 2006)], as well as of processing, inferential reasoning, decision-making and knowledge integration skills [e.g., Siena Reasoning Test. (Yusko, Goldstein, Oliver, & Hanges, 2010).] Fagan and Holland (2002, 2007) have shown evidence that the B-W group difference on items that can only be solved with acquired specific knowledge (of a kind that they consider typical of GCA tests) is considerably greater than the difference on items that can be solved with information available to both races or with newly acquired knowledge. They consider this evidence that cultural differences in information and items contribute to race differences in GCA scores and that there is no real difference in GCA. Other scholars accept that there may be differences between groups, but that these differences

require better explanations than “race” alone (Helms, 1993). Outtz and Newman (2010) suggest that industrial-organizational psychologists study the psychological, sociological and economic factors that contribute to performance.

Subgroup contamination on criterion measures has also been of concern. Lefkowitz (1994) has observed that, when differences on the measured criterion are biased, analysis of test bias is confounded. Over the last 20 years, studies have suggested differences in performance of varying magnitude (Kraiger & Ford, 1985; McKay & McDaniel, 2006; Pulakos, White, Oppler, & Borman, 1989; Wendelken & Inn, 1981), such that mean differences between Blacks and Whites may be smaller than once thought (Roth, Huffcutt & Bobko, 2003; Bobko, Roth, & Potosky, 1999; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). Group differences are believed to vary based on rater-ratee race interaction (Kraiger & Ford, 1985; Stauffer & Buckley, 2005), on criterion type and cognitive loading (McKay & McDaniel, 2006), and on type of measure (e.g., subjective versus objective within category level, sample versus abstract, etc.) (Ford, Kraiger, & Schechtman, 1986; Pulakos et al., 1989). With perceptions of bias on the predictor, on the criterion measure or on both, some believe CGA tests function differently across races, with potential predictive bias.

In the special issue, only Outtz (2002) strongly questions whether GCA tests are unbiased. He notes that group differences are much larger on the predictor than on the criterion, and, thus, the rate of false negatives among minority group members may be higher than that of the majority. He raises the possibility that GCA tests may measure constructs irrelevant to successful job performance. Also, he observes that differences in performance may be less than some believe, due to the types of criteria considered and the way in which they have been measured (e.g., where rater-ratee race differences may interact). Goldstein et al. (2002) suggest that the way in which selection systems are cognitively overloaded may have negative effects on minority subgroups.

Eight of the remaining ten authors consider GCA tests unbiased, though their positions differ slightly. Gottfredson (2002), Kehoe (2002), Reeve and Hakel (2002), and Schmidt (2002) are of the opinion that there are real group differences in intelligence that GCA tests measure. Reeve and Hakel (2002) believe that stereotype threat does not account for major differences in group means. They also attempt to refute the more general argument that GCA tests are “White,” observing that Jews and Asians consistently score higher than non-Jewish Whites. Predictive bias is also addressed frequently in the special issue. Gottfredson (2002), Murphy (2002), Reeve and Hakel (2002) and Schmidt (2002) share a belief that the predictive validity of GCA does not differ across groups; if anything, some believe that there is evidence that GCA scores are biased *in favor* of minority group members (Gottfredson, 2002; Reeve and Hakel, 2002; Salgado, 2002; Schmidt, 2002). Schmidt states that equally low-scoring Whites and Blacks share equally low chances of failure on the job. Countering one of Outtz’s arguments, Murphy (2002), Schmidt (2002) and Viswesvaran and Ones (2002) all remark that the group mean differences in GCA scores are expected to be larger than differences in performance scores, as GCA is an imperfect predictor. Kehoe (2002), Schmidt (2002) and Tenopir (2002) caution that redesigning tests to reduce group differences may introduce predictive bias against the majority. Tenopir (2002) also states that issues of rater-ratee race may not be as serious as they were once thought to have been.

2.1.1.1.1.1 Social Fairness

Social Fairness: General cognitive ability tests produce fair social outcomes.

As conceptualized in the six-factor model, “social fairness” beliefs are related to the effects that general cognitive ability tests produce, rather than to measurement or predictive bias. Some scholars believe that a selection tool is socially fair if it is equitable to individuals; however, they point out that “fairness,” as a statement of values, is non-psychometric in nature. These scholars

believe that general cognitive ability (GCA) tests do not create workforce discrimination; rather, they reveal disparities in ability that may be due to societal factors (e.g., Hunter & Hunter, 1984; Schmidt & Hunter, 1981). Murphy (2002) identifies the widely-held belief that discrimination due to GCA testing may be considered “fair” if GCA tests can be shown to be predictively valid (i.e., if performance ratings covary with test scores) (e.g., Hunter & Hunter, 1984; Sharf, 1988). Those who hold this belief make little distinction between social fairness and Cleary-model fairness (Arvey & Faley, 1988; Gottfredson, 2002), such that “relational equivalence” remains central to this argument.

Those who feel that GCA tests are fair seem to believe that discarding them would *create* social injustice. Schmidt and Hunter (1976) caution that rejecting highly valid tests (e.g. the general cognitive ability test) leads to the use of less valid approaches to selection (e.g., interviews) resulting in greater unfairness, potentially for both groups and individuals. Also, as suggested previously, some believe that ignoring the group differences that CGA tests reveal or rushing to apply Title VII are tantamount to ignoring the social conditions that underlie these differences (e.g., Hunter & Hunter, 1984; Schmidt & Hunter, 1981; Sharf, 1988). Further, according to Ryanen (1988), preferential treatment for any group imperils the principle of equality (Ryanen, 1988). By throwing away GCA tests, Ryanen (1988) believes that society would move away from a merit-based system to a patronage system, in which racist social policy would dictate that race should be considered in selection. Preferential treatment would then become institutionalized, as disadvantaged subgroups would resist giving up an unfair advantage to which they would feel entitled (Ryanen, 1988). Gottfredson (1986b) suggests that we do not even know the costs to society associated with promoting group equality by throwing out tests.

Then, there are also those who consider general cognitive ability tests socially unfair. In addition to the aforementioned “culturalists”, Gottfredson (1986b) defines two groups that question the fairness of intelligence testing in selection: “functionalists” and “revisionists.” According to Gottfredson, functionalists consider education and training to be proximal factors of performance and believe that intelligence is a distal factor that has no direct relationship with criteria. Thus, they view the heavy focus on group differences in intelligence as diverting attention from disparities in skill acquisition. Gottfredson says that revisionists believe that intelligence and education are not critical in performance, and that knowledge, skills, and abilities can be acquired on the job. They believe that socioeconomic status is the precursor to GCA and educational credentials and, thus, they are manifestations of the privilege that creates discrimination when included in selection systems (Gottfredson, 1986b). Other scholars agree that test fairness is not merely a psychometric issue, but a complex question that includes considerations of test administration, score interpretation and application, as well as social, legal and organizational concerns (Hough et al., 2001; Sireci & Geisinger, 1998; Willingham, 1999). Some think that a test may be “unbiased” yet lead to unfair outcomes (Darlington, 1971; Thorndike, 1971).

As defined by Gottfredson (1986b), functionalists, revisionists and culturalists do not feel that rejecting cognitive testing would create social injustice. Rather, various scholars argue that GCA tests are but one class of many biased selection procedures that cause discrimination and should be rejected. Several strong views of this kind are shared in a special issue of the *Journal of Vocational Behavior* in 1988. Seymour (1988) contests the scientific basis for validity generalization and provides a blueprint for the way in which such studies can be challenged. Goldstein and Patterson (1988) believe that validity generalization is a step backward that could undermine the social advances associated with affirmative action. In addition to predictive validity,

various scholars question the use of utility to justify the use of ability tests, stating that evidence of utility is not clear enough to do so (Goldstein & Patterson, 1988; Levin, 1988; Seymour, 1988).

Bishop (1988) presents a milder functionalist view: While conventional tests may offer predictive validity and utility, the use of selection procedures that include material from high school curricula might increase social justice by encouraging prospective workers to pursue education. Thus, while GCA tests might add value in selection, they can be supplemented or replaced by tests that measure a more proximal factor (e.g., education). Many who encourage the abandonment of general cognitive ability tests also believe that various employment tests limit opportunities for members of minority groups. Seymour (1988) states that tests “can be an engine for the exclusion of racial minorities more permanent and thorough than individual ill will” (Seymour, 1988). Goldstein and Patterson (1988) view validity generalization and social fairness (i.e., as embodied in affirmative action) as forces locked in an all-or-nothing battle, where a victory for psychometricians would be a tremendous loss of prospects for Blacks. It is worth observing that those who believe in the social fairness of general cognitive ability tests agree with those who find them socially unfair in at least one respect: They believe that the world is full of disparities in opportunity. Those who believe that GCA tests are unfair feel that general cognitive ability tests create these disparities, whereas those who believe in their social fairness feel that tests correctly sort workers into job families of disparate attractiveness.

In the special issue, several authors’ points of view can be characterized as “pro-fairness”. Schmidt (2002) and Gottfredson (2002) believe that GCA tests offer great utility and that their abandonment would be unfair to society and to individuals (Schmidt, 2002). Some authors, in fact, seem to believe that GCA tests are instruments of social justice. Reeve and Hakel (2002) state, “Indeed, no other social intervention has been as successful at breaking down class-based privilege

and moving underprivileged, intelligent youth into higher education than the standardized assessment of intelligence” (p. 50). Ones and Viswesvaran (2002) support this view, adding that, despite the stigma placed on GCA tests due to the eugenics movement, “[GCA] can be used to create a society of equal opportunity for all.” Schmidt (2002) and Gottfredson (2002) concur. For this and other reasons, some “pro-fairness” authors object to the term “adverse impact” itself, a term that they claim implies that ability tests cause differences rather than reveal them (e.g., Schmidt, 1988; Schmidt, 2002), some replacing it with “disparate impact” (e.g., Gottfredson, 2002). Schmidt (1988) has recast “adverse impact” as a label for a lower passing rate for Blacks; Gottfredson (2002) feels that claims of adverse impact are driven by social policy.

Some special issue authors believe that sole reliance on GCA tests is unfair (e.g., Outtz, 2002; Murphy, 2002) and that the other tests may be fairer (e.g., Goldstein et al., 2002; Outtz, 2002; Sternberg & Hedlund, 2002). Outtz (2002) and Goldstein et al. (2002) suggest that selection systems may be cognitively overloaded and stacked against Blacks, with differences overestimated in selection and performance. While Outtz (2002) questions whether GCA tests are fair at individual level, both he and Murphy (2002) note the undesirable inequity that they cause at group and societal levels.

2.1.1.1.1.1.1 Tradeoffs

Tradeoffs: There are tradeoffs between predictive validity and fairness when deciding whether to use general cognitive ability tests.

Some scholars accept that there are real group differences in general cognitive ability (GCA), and that the choice to employ these tests is a decision to trade efficiency for adverse impact. Murphy (2002) suggests that this choice is not scientific but value-based, and that both academics and human resource professionals “must.... take a stand on where their values lie” (p. 182). He

recommends that organizations utilize policy-capturing studies to derive utility functions that frame the trade-offs for decision-makers. De Corte, Lievens and Sackett also recognize the competing goals of predictive validity and adverse impact (De Corte, Lievens & Sackett, 2007; DeCorte, Lievens & Sackett, 2008; De Corte, Sackett & Lievens, 2011; Sackett, De Corte, & Lievens, 2008) and advance an approach to designing Pareto-optimal selection systems.

Six of the twelve special issue papers directly address tradeoffs. Five of these (i.e., Goldstein et al., 2002; Gottfredson, 2002; Kehoe, 2002; Murphy, 2002; and Outtz, 2002) appear to acknowledge that predictive validity and minority group passing rate trade against each other in the context of GCA testing. All six papers (including Schmidt, 2002) also state that this tradeoff is not inherent in selection systems, as other predictors may be introduced that simultaneously increase validity and reduce mean group differences in scores. Among the group, all consider the consequences of the tradeoff undesirable, though Schmidt (2002) and Gottfredson (2002) believe that it will continue until great changes in society occur, and not changes in testing. There is a distinct difference in the positions of the six “tradeoffs” authors. Kehoe (2002), Murphy (2002) and Outtz (2002) believe that when organizations choose to use GCA tests, they are accepting the adverse impact that results from their use. Schmidt (2002) and Gottfredson (2002) do not accept the theory of “adverse impact” and, thus, do not perceive that the decision to utilize GCA tests trades with social unfairness. Kehoe (2002), Murphy (2002) and Outtz (2002) suggest that organizations that employ GCA tests may be trading away value to the organization, depending on how value is defined (e.g., diversity). By contrast, Schmidt (2002) believes that *not* using the most valid predictor comes at the cost of productivity.

The usefulness of the six factors in clarifying areas of agreement and disagreement between scholars is demonstrated in Table 1.

Table 1. Scholars' areas of agreement and disagreement in the special edition, by factor.

	Primacy	Criterion Dependence	Alternative Cognitive Predictors	Test Bias	Fairness	Tradeoffs
Yes	• Gottfredson • Murphy • Ree and Carretta • Salgado and Anderson • Schmidt • Reeve and Hakel • Viswesvaran and Ones	• Goldstein, Zedeck and Goldstein • Gottfredson • Kehoe • Murphy • Outtz • Ree and Carretta • Reeve and Hakel • Schmidt • Tenopyr	• Goldstein, Zedeck and Goldstein • Sternberg and Hedlund • Tenopyr • Reeve and Hakel	• Goldstein, Zedeck and Goldstein • Outtz	• Gottfredson • Kehoe • Reeve and Hakel • Schmidt • Viswesvaran and Ones	• Goldstein, Zedeck and Goldstein • Gottfredson • Kehoe • Murphy • Outtz • Schmidt
No	• Goldstein, Zedeck and Goldstein • Kehoe • Outtz • Sternberg and Hedlund • Tenopyr		• Gottfredson • Kehoe • Murphy • Ree and Carretta • Schmidt • Viswesvaran and Ones	• Gottfredson • Kehoe • Murphy • Reeve and Hakel • Salgado and Anderson • Schmidt • Tenopyr • Viswesvaran and Ones	• Goldstein, Zedeck and Goldstein • Murphy • Outtz • Sternberg and Hedlund	• Sternberg and Hedlund
Unclear		• Salgado and Anderson • Sternberg and Hedlund • Viswesvaran and Ones	• Outtz • Salgado and Anderson	• Ree and Carretta • Sternberg and Hedlund	• Ree and Carretta • Tenopyr • Salgado and Anderson	
Notes	• Kehoe's position is moderate • Reeve and Hakel's position is nuanced	• Schmidt only indicates that GCA validity might be <i>higher</i>, due to measurement issues. • Schmidt, and Ree and Carretta, place little emphasis on criterion dependence	• Reeve and Hakel only support the utility of aptitudes • Viswesvaran and Ones consider alternatives to be public relations • Only Goldstein et al. and Sternberg and Hedlund strongly urge more research	• Kehoe suggests that differential prediction might be better analyzed outside of validation studies, with overall value to the organization as a criterion	• Murphy tests fair to individuals level but unfair to groups • Gottfredson and Schmidt challenge the notion of "adverse impact" • Only Outtz strongly asserts unfairness	• Goldstein et al. consider tradeoffs overtaken • While there is near consensus, scholars opinions on <i>inevitability</i> may differ

CHAPTER 3

METHODS

3.1 Focus on Construct Validity

The goal of the present study was to identify a *useful framework* for the debate over cognitive ability (GCA) and GCA testing. This framework could only be considered worthy of use if a corresponding model were meaningful and could be supported empirically. A prior study (i.e., Murphy et al., 2003), summarized in a previous section of this article, had proposed two dimensions of the debate and, thus, tested a two-factor model believed to represent them. While the results of confirmatory factor analysis demonstrated that this two-factor model fit the data from a survey of beliefs about GCA and GCA testing, all lambda values were moderately to extremely low. As Edwards (2003) notes, “CFA is the hallmark of modern approaches to construct validation” (p. 332). In confirmatory factor analysis, a formula that indicates the relationship between a measure and a construct of interest is $X_i = \lambda_i \xi + \delta_i$. In this formula, X_i signifies the *i*th measure, ξ signifies a factor corresponding with the construct, the λ_i signify the “factor loadings” of the X_i , and the δ_i signify “the uniquenesses of the X_i ” (where the δ_i consist of measurement specificity and random measurement error) (p. 341). Thus, the higher that lambda (and, relatedly, R^2) values are, the stronger the relationship between responses and respective factors should be, given unchanged delta values. When a series of items load uniquely, this serves to validate the existence of the factors, and when lambda values are high, it shows that the items meaningfully represent those factors.

Low lambda values do not, in themselves, “prove” that a construct of interest does not exist; rather, they do not provide strong evidence for it, as the corresponding items remain uninformative. Also, while a good χ^2 statistic and good fit index scores suggest that the model summarizes the data well, fit alone is not sufficient evidence for construct validity. As mentioned previously, Murphy et al.’s (2003) two-factor model fit the data, and the large number of items loading (albeit with quite low lambda values) uniquely provided some evidence of its two factors. However, with low loadings, the items shed little light on the nature of the “g-ocentric” and “Societal” belief factors. It is worth pointing out, however, that the factor model was probably primarily intended to create a descriptive framework, and the emphasis on construct validity in this kind of taxonomic approach could be somewhat lower than in a generative approach. It was not the researchers’ goal to create a test that would produce meaningful, interpretable factor scores. Deep insights into latent variables that were not part of a reflective model might not have been deemed critical, and particularly given that these variables (e.g., belief in the “Social Fairness” of GCA tests) were of little interest beyond the current research. Nonetheless, even if the approach to this research were narrowly psycholexical [e.g., Goldberg’s (1990) investigation of personality terminology] or socioanalytic (Hogan, 1983, 1996)—where factor analysis had been undertaken not necessarily to understand the broader debate but simply to ascertain the dimensionality of the questionnaire or of the nature of the respondents—the factors still would only be considered useful with some support for their hypothetical meanings. The Murphy et al. (2003) paper provides little of this kind of support.

Without directly pitting it against Murphy et al.’s two-factor solution, the present study sought to find a model that fit the data, with high factor loadings. It was believed that such results would allow scholars to use this model with some confidence in framing future debate over GCA

and GCA testing, with items not only informing the factors but providing specific points of argument.

3.2 Procedure

It was conjectured that the reason that Murphy et al.'s (2003) factors were not deeply informed by the indicators was due to the inductive approach by which the model was identified. That is: The researchers' a priori conceptualization of historical disagreement may have prevented them from examining the structure of an actual debate, upon which their model could have been based. In the present study, the framework was to be generated deductively, the dimensions of which would then be tested empirically in the form of a factor model. Given the emphasis on deduction, several rules were established prior to investigation. First, research would begin with a literature review of an actual (and recent) debate and not through grouping the items in the Murphy et al. (2003) survey. In part, this rule was set to limit the researcher to deduction; in part, it was designed to avoid sorting the items neatly into Murphy et al.'s (2003) "five broad groups," which, as noted previously, might have served as factors in an alternative to their two-factor model. Once the literature review was completed, and presumably with a better understanding of dimensions of the debate, survey items would be reviewed and grouped according to Thurstone's technique. Based on these groupings, a model of a certain number of factors would be selected. Then, a characterization of the model would be subjected to factor analysis using Murphy et al.'s (2003) survey data. Bearing in mind that this factor analysis would be for heuristic (and not absolute) purposes, if the fit statistic and index scores were good and factor loadings were

relatively high, the model might be valid, informative and useful in debate, even if it could not be used for measurement.

For the literature review, the special issue of *Human Performance* (2002) was selected, and for numerous reasons: 1. Most of the authors had participated previously in the “Millennial Debate on *g* in I-O Psychology” and were aware of each other’s positions, such that they were able to defend their viewpoints and counter others’ arguments, either directly [e.g., Schmidt’s (2002) section entitled “Responses to Issues in the Debate”] or indirectly; 2. These authors could all fairly be considered experts on the topics of general cognitive ability (GCA) and GCA testing; 3. The publication was recent enough to point toward an *actual* factor model useful for contemporary purposes; 4. Murphy et al.’s survey items were, in part, based upon the authors’ assertions; and, 5. The investigator did not have access to the transcripts of any actual debate, but felt that these articles served as reasonable surrogates. [Note: Recordings or transcripts of the “Millennial Debate on *g* in I-O Psychology”, which might have served the present purposes better, could not be accessed.]

At the end of this review, an exhaustive list of the assertions advanced by the various authors was created. A grid was constructed that displayed these assertions, showing areas in which these authors agreed or disagreed. The arguments first appeared to sort into seven categories, in which the authors’ assertions seemed to be either “for” or “against” some general position. These seven categories were considered as factors for a model. Items from the Murphy et al. (2003) survey were then grouped, with the expectation that they might load onto factors in a subsequent model. Again, there appeared to be seven groups. However, upon further consideration of the grid, two of the initial groups (i.e., “Fairness” and “Limit to Potential”) appeared to collapse into one (i.e., “Social Fairness”.) Thus, it was suggested that a slightly more parsimonious six-factor

model should be examined in factor analysis, with the open possibility of examining a seven-factor model. Four of these six factors seemed to be concise statements of Murphy et al.'s (2003) item groupings: "Primacy" (which mirrored the "importance or uniqueness of cognitive ability tests" group), "Criterion Dependence" (which resembled the "links between ability tests and job requirements" group), "Alternative Cognitive Predictors" (which was similar to the "alternatives or additions to ability testing"), and "Social Fairness" (which resembled the "societal impact of cognitive ability testing" group.) The remaining factors were named "Predictive Bias" and "Tradeoffs". These factors did not appear in Murphy et al.'s (2003) conceptualized groups of items.

While these six factors were identified through the use of the aforementioned grouping technique, a number of ways in which items could be placed on these factors were identified. Ultimately, one characterization would be selected, which could be considered a "fair guess". Rather than using all of 49 Murphy et al.'s (2003) total items, 29 were picked. Realizing a clear, simple and easily interpretable model was considered to be more important than simply accounting for all of the previous study's 49 items. Some items appeared to be clearly related to the factors (e.g., #37, "General cognitive ability tests are fair" seemed to measure "Social Fairness"); others had no direct relationship with GCA at all (e.g., #21, "Diversity in the workplace gives organizations competitive advantage") and were easily eliminated. Further, items that are not clearly related to the factors could be expected to load lower. They run greater risk of misassignment and may increase the chance of poor fit without adding considerable explanatory value. As previously mentioned, good fit and high loadings would be strong indications of construct validity, and providing evidence for these hitherto unexplored constructs was an

important objective. Thus, for a variety of reasons, an efficient model seemed preferable to an all-encompassing model.

Confirmatory factor analysis (CFA) was selected for testing the characterization of the six-factor model on Murphy et al.'s (2003) data. CFA is the “hallmark” for construct validation (Edwards, 2003), in part by accounting for δ_i [as opposed to, for example, clustering (e.g., Revelle, 1979), which could have been used if the objective were merely to see how actual responses group together.] In the present study, first, an initial CFA was run on the “fair guess” characterization, utilizing the Murphy et al. (2003) data. Thus, while the approach could be considered methodologically exploratory, the actual analysis was technically confirmatory. However, since a “fair guess” version of the six-factor model was unlikely to be exactly correctly specified, some noise might have been fitted. As a first CFA run might be over-fitted—hence the parameter estimates (and, in turn, the p-values and fit) too liberal in the first run—a subsequent CFA should be run for the purpose of confirmation/validation. This is what was done. Typically, the whole sample is split into two subsamples. The first subsample is used as training data; the second subsample is held over for confirmation/validation. However, in this case, having already tested the model on the whole sample (in this case, the training data), a CFA on the first subsample was performed for confirmation/validation. A CFA on the second (holdout) subsample was used to test a very slightly different model, in which any items with very low lambda and R^2 values (thus, little informing the model) in the previous CFA were dropped. This new model could then be examined.

Indices of fit and parsimony were selected. Murphy et al. (2003) utilized three indices (i.e., the Goodness-of-Fit Index, the Parsimony Goodness-of-Fit Index, and Root Mean Square Error of Approximation) to gauge how well their model summarized the survey data, as well as to judge its

parsimony. Root Mean Square Error of Approximation (RMSEA) (Steiger & Lind, 1980) indicates how well the model fits the population covariance matrix (Byrne, 1998), with sensitivity to the number of parameters, such that parsimony is favored. It is considered “one of the most informative fit indices” (Diamantopoulos & Siguaw, 2000, p. 85). The Goodness-of-Fit Index (GFI) (Jöreskog & Sorbom, 1981) indicates how much variance is accounted for by population covariance (Tabachnick & Fidell, 2007), such that the lower the variance that cannot be explained by the model, the higher the statistic. GFI is downwardly biased when there are a large number of degrees of freedom relative to sample size (Sharma, Mukherjee, Kumar, & Dillon, 2005), and upwardly biased with large samples (Bollen, 1990; Miles & Shevlin, 1998) and as the number of parameters increases (MacCallum & Hong, 1997). Accordingly, the Parsimony Goodness-of-Fit Index (PGFI) (Mulaik et al, 1989) was created to adjust for the loss of degrees of freedom. Due to its sensitivity, recommendations against the use of GFI have been made (Kenny, 2014; Sharma et al, 2005). While Murphy et al. (2003) utilized GFI and PGFI, they were dropped from the present study.

While RMSEA is considered a fine indicator of model fit, other fit indices have been included in the present study. The ratio of chi square to degrees of freedom (i.e. χ^2/df) is an old test of fit (Kenny, 2014). Small values indicate goodness-of-fit and large values indicate badness-of-fit. However, there is no established standard for what constitutes good and bad fit (Kenny, 2014). Also, as the χ^2 test tends produce Type I errors when sample sizes are large (Bentler & Bonnet, 1980; Jöreskog & Sorbom, 1981) or when the distribution is non-normal (Kenny, 2014), it is not the best indication of fit for the present study. RMSEA, based largely on χ^2/df (Kenny, 2014), is preferred for reasons already outlined. The χ^2 test is presented, but mainly as a matter of convention.

In addition to absolute fit indices, several relative fit indices are reported. The Normed Fit Index (NFI) (Bentler & Bonnet, 1980) compares the hypothesized model χ^2 to the null model χ^2 , such that its improvement in fit is indicated. However, as NFI may overestimate fit for models with more parameters, the Non-Normed Fit Index (NNFI) may be somewhat preferable for the present study (Kenny, 2014). Similar to NFI, the Incremental Fit Index (IFI) compares χ^2 of the hypothesized model to χ^2 of the null model, but in a manner that also accounts for degrees of freedom. Thus, IFI is fairly insensitive to the size of the sample (Kenny, 2014).

Given the specifics of the present study (i.e., the number of model parameters, the large size of the sample, and the high number of degrees of freedom), RMSEA, NNFI and IFI should be the best indicators of fit. The Parsimony Normed Fit Index (PNFI) score is also included, despite the lack of a widely-accepted cutoff heuristic.

CHAPTER 4

RESULTS

4.1 Model Fit

Confirmatory factor analysis used under the exploratory approach. Under Lisrel 8.71, a six-factor model with a specified pattern of fixed-at-zero and free-to-be-estimated loadings was tested, per the selected “fair guess” characterization. This model appeared to fit the data reasonably well, particularly given that somewhat lower fit index cutoffs (other than for RMSEA) may be used for complex models (e.g., models with more factors, with more degrees of freedom, etc.) than for simpler models (Cheung & Rensvold, 2002). [Root-Mean-Squared Error Analysis (RMSEA) = 0.0594; Normed Fit Index = 0.882; Non-Normed Fit Index (NNFI) = 0.925; Incremental Fit Index (IFI) = 0.934]. χ^2 was expectedly high, given the large sample size ($N = 703$), and significant (771.387, $P = 0.00$); there were 362 degrees of freedom. $\frac{\chi^2}{df} = 2.13$. The model also seemed to be parsimonious [Parsimony Normed Fit Index (PNFI) = 0.787, though there is no generally-accepted cutoff for PNFI.]

Subsequent confirmatory factor analyses. Each subsample of the data was analyzed using Lisrel 8.71. For the first subsample, model fit was slightly better than in the first CFA. (RMSEA = 0.0495; NFI = 0.883; NNFI = 0.937; IFI = 0.944). Again, χ^2 was expectedly high and significant (651.349, $P = 0.00$); there were 362 degrees of freedom. $\frac{\chi^2}{df} = 1.799$. This model was also found to be reasonably parsimonious (PNFI = 0.787; again, there is no accepted cutoff.) Thus, the six-factor model’s fit was validated. A similar model (with only item #43 dropped, for reasons explained in the Item Factor Loadings section below) was tested on the second subsample. Fit:

(RMSEA = 0.051; Minimum Fit Function Chi-Square = 621.610, P = 0.00; 335 degrees of freedom; $\frac{\chi^2}{df} = 1.855$; NFI = 0.887; NNFI = 0.937; IFI = 0.945). Parsimony: (PNFI = 0.786). Thus, the “new” model achieved nearly identical fit. Fit statistics are summarized in Table 2.

Table 2. Summary of fit statistics

	<i>N</i>	χ^2	<i>df</i>	χ^2/df	RMSEA	NFI	PNFI	NNFI	IFI
6-factor CFA (full sample)	703	771.39	362	2.13	0.059	0.88	0.79	0.93	0.93
6-factor CFA (1 st subsample)	360	651.35	362	1.80	0.050	0.88	0.79	0.94	0.94
6-factor CFA (2 nd subsample)	343	621.61	335	1.86	0.051	0.89	0.79	0.94	0.95

4.2 Item Factor Loadings

Confirmatory factor analysis used in exploratory approach. In factor analysis, items with small lambda values may be considered to have little informative value and may negatively affect fit. Thus, decisions to retain or drop these items may be made on the basis of loadings and R^2 . Retention heuristics were provided in a prior section of this article. In recapitulation, given the nature of the present research and the size of the sample, a liberal heuristic for retaining items based on lambda values (i.e., factor loadings) was utilized (i.e., retain $\lambda \geq 0.3$ or ≤ -0.3). $R^2 > 0.25$ were considered large, where $0.10 \leq R^2 < 0.25$ was considered a conventional indication of a moderate relationship. Corresponding with the lambda heuristic, $R^2 < 0.10$ tends to suggest a weak relationship, such that the item should generally be discarded. In Lisrel 8.71, a total of 29 items loaded onto the six-factors. Only (i.e., item #43) failed to meet both retention heuristics.

Table 3. Factor Pattern of CFA Run on Full Sample.

Corresponding item content is specified in Appendix A

Item #	Primacy	Criterion Dependence	Alternatives	Test Bias	Social Fairness	Tradeoffs	R^2
1	0.59						0.35
2	0.75						0.59
10	0.73						0.54
11	0.42						0.18
14	-0.53						0.28
28	0.56						0.31
29	0.56						0.31
45	0.48						0.23
16		0.53					0.28
17		0.52					0.28
33		0.31					0.10
4			0.42				0.17
5			-0.52				0.27
9			0.49				0.24
15			0.65				0.43
30			-0.38				0.14
31			0.66				0.44
13				0.43			0.18
39				0.50			0.25
49				0.60			0.36
26					0.69		0.48
27					0.58		0.34
36					0.74		0.55
37					0.88		0.69
38					-0.52		0.27
19						0.56	0.31
22						0.47	0.23
*43						0.26	0.07
47						0.56	0.32

* Item failed to meet retention heuristic

The degree to which factors covary is an indication of their independence (or lack thereof). In the factor correlation matrix, one value is > 0.70 and another is < -0.70 . These are somewhat large. Thus, while some factors do not appear to be strongly related (e.g., “Alternatives” and “Tradeoffs”), the independence of others (e.g., “Primacy” and “Social Fairness”) is questionable.

Table 4. Factor Correlation Matrix of CFA Run on Full Sample.

	Primacy	Criterion Dependence	Alternatives	Test Bias	Social Fairness	Tradeoffs
Primacy	1.00					
Criterion Dependence	-0.48	1.00				
Alternatives	-0.59	0.60	1.00			
Test Bias	-0.59	0.49	0.45	1.00		
Social Fairness	0.73	-0.29	-0.50	-0.72	1.00	
Tradeoffs	-0.10	0.26	0.12	0.32	-0.33	1.00

Subsequent Confirmatory Factor Analyses. Confirmatory factor analysis (CFA) has been characterized as the gold standard for construct validation. In the present study, the same heuristics applied to lambda values and R^2 was used in all factor analyses. Of the 29 items retained from the first CFAs, three (items #11, 17, and 43) fell outside the CFA lambda value retention range, but only item #43 failed to meet both heuristics (i.e., for lambda value and R^2), indicating a particularly weak relationship. While item #43 was included in the first subsample CFA, it was dropped in the second. As noted above, almost no change in fit resulting from removing this item. However, given its limited informative value, however, it should be eliminated (in order to remove what is a probably fruitless point of debate.) While the CFA results supported the meaningfulness of “Primacy”, “Alternatives” and “Social Fairness” beliefs factors, “Criterion Dependence”, “Tradeoffs” and particularly “Test Bias” were not as well informed their respective items.

Table 5. Factor Pattern of CFA Run on First Subsample.

Corresponding item content is specified in Appendix A

Item #	Primacy	Criterion Dependence	Alternatives	Test Bias	Social Fairness	Tradeoffs	R^2
1	0.55						0.30
2	0.67						0.45
10	0.69						0.48
*11	0.29						0.09
14	-0.46						0.21
28	0.57						0.33
29	0.54						0.29
45	0.54						0.29
16		0.61					0.37
*17		0.24					0.06
33		0.49					0.24
4			0.42				0.18
5			-0.45				0.20
9			0.44				0.19
15			0.60				0.37
30			-0.50				0.25
31			0.47				0.22
13				0.32			0.10
39				0.61			0.37
49				0.77			0.59
26					0.63		0.40
27					0.46		0.21
36					0.65		0.43
37					0.77		0.59
38					-0.57		0.33
19						0.33	0.11
22						0.45	0.20
*43						0.18	0.03
47						0.65	0.42

* Item failed to meet retention heuristic

Table 6. Factor Pattern of CFA Run on Second Subsample (Testing the “New” Model).

Corresponding item content is specified in Appendix A

Item #	Primacy	Criterion Dependence	Alternatives	Test Bias	Social Fairness	Tradeoffs	R^2
1	0.55						0.30
2	0.67						0.45
10	0.69						0.48
*11	0.29						0.09
14	-0.46						0.21
28	0.57						0.33
29	0.54						0.29
45	0.54						0.29
16		0.60					0.37
*17		0.24					0.06
33		0.49					0.24
4			0.42				0.18
5			-0.45				0.20
9			0.44				0.19
15			0.60				0.37
30			-0.50				0.25
31			0.47				0.22
13				0.32			0.10
39				0.61			0.37
49				0.77			0.59
26					0.63		0.40
27					0.46		0.21
36					0.65		0.43
37					0.77		0.59
38					-0.57		0.33
19						0.31	0.10
22						0.43	0.19
47						0.64	0.41

* Item failed to meet retention heuristic

As expected, in light of the CFA run on the full sample, some factors were expected to be strongly related and others weakly related in the subsample CFAs. Indeed, correlations were somewhat higher than before. The independence of “Primacy” and “Social Fairness” may be questioned, as may the independence of “Test Bias” and “Social Fairness.” As expected, correlation matrices were also nearly identical between the two CFAs run on subsamples.

Table 7. Correlation Matrix of CFA Run on First Subsample.

	Primacy	Criterion Dependence	Alternatives	Test Bias	Social Fairness	Tradeoffs
Primacy	1.00					
Criterion Dependence	-0.33	1.00				
Alternatives	-0.62	0.62	1.00			
Test Bias	-0.58	0.48	0.43	1.00		
Social Fairness	0.78	-0.31	-0.56	-0.73	1.00	
Tradeoffs	-0.21	0.43	0.24	0.45	-0.58	1.00

Table 8. Correlation Matrix of CFA Run on Second Subsample (Testing the “New” Model).

	Primacy	Criterion Dependence	Alternatives	Test Bias	Social Fairness	Tradeoffs
Primacy	1.00					
Criterion Dependence	-0.33	1.00				
Alternatives	-0.62	0.62	1.00			
Test Bias	-0.57	0.48	0.43	1.00		
Social Fairness	0.78	-0.31	-0.56	-0.73	1.00	
Tradeoffs	-0.25	0.45	0.26	0.47	-0.62	1.00

CHAPTER 5

DISCUSSION

5.1 The Model

The six-factor model fit the training data, and fit was then supported in subsequent analyses. The final model achieved good fit, like Murphy et al.'s (2003). Discarding the low loading item #43 should be beneficial, as the assertion it makes might no longer be suggested as a point of debate. Items #1 and #17 also showed low loadings; in another characterization, they might not be included.

However, there are lingering questions about the factors. As previously noted, high correlations between factors (as shown in Tables #4, #7 and #8) cast doubt upon their independence. The difference between Test Bias and Social Fairness might be clear to those who debate GCA and GCA testing, but they may have been confused in some way among the broader SIOP membership that responded to the survey.

5.2 Further support for construct validity

The survey results may indirectly support the meaningfulness (and, thus, usefulness) of the model in other ways. Murphy et al. (2003) grouped the 49 items as “consensus items”, “polarizing opinions” and “neither controversy nor consensus”, based on percentage of respondents who agreed or disagreed with them. Ultimately, by Murphy et al.'s (2003) heuristics, only ten items were “polarizing”. Four of these ten items loaded onto the present study's “primacy” factor. It is

not surprising, thus, that primacy seems to have been the most fertile area of debate in the special issue of *Human Performance* (2002). Too, it is not surprising that “Primacy” was the best validated factor, as those on each side should be expected to provide similar responses. By contrast, all three items which loaded onto the present study’s “criterion dependence” factor were referred to as “consensus” in the Murphy et al. (2003) study. It should be evident in Table 3 that, in the special issue, “Criterion Dependence” was not contentious. In an uncontentious area of debate, sides are less distinct. [Note: Indeed, based on argumentation theory (as explained in this study’s suggestions for future research), scholars may consider eliminating (or at least revising) “criterion dependence” in future debates. Perhaps it can be retained if group- and organizational-level criteria—e.g., Kehoe’s (2002) “overall worth or merit”—are introduced.]

Further, in the special issue, only Outtz strongly questioned whether GCA tests are socially fair and unbiased; correspondingly, none of items that loaded onto the present study’s “Predictive Bias” and “Social Fairness” factors fell into Murphy et al.’s (2003) “polarizing” category. Tradeoffs between efficiency and fairness were not featured prominently in the special issue. The six-factor model’s “Tradeoffs” factor was not strongly validated by high factor loadings and R^2 .

5.3 Value of the model

Darlington (2004) proposes two uses for factor analysis. The absolute use of factor analysis aims at fully accounting for the relationships between variables; the heuristic use is simply intended to provide the best summarization of the data. The six-factor solution was not intended to serve as a measurement model. It summarizes the data well and with higher loadings than Murphy et al.’s (2003) two-factor model. The meaning of the two-factor model is open to great

interpretation and offers little insight into what specific points of debate might be. The six-factor model should be more informative. It cannot be construed as a *better* model, but only as *another* model; however, it is the only of the two models that offers utility for future debate. Table 1 illustrates this utility as a structure for debate by showing where scholars agreed and disagreed in the special issue.

CHAPTER 6

LIMITATIONS

One of the objectives this study was to arrive at a model deductively, through examination of an actual debate, rather than the inductive approach adopted in Murphy et al.'s (2003) study. However, without access to a recording or transcript of a prior debate, follow-on articles to the "Millennial Debate" were used in proxy. While written arguments can be considered debate (Ehninger & Brockriede, 2008), only Viswesvaran and Ones (2002) had the opportunity to support or rebut the assertions of the various articles.

Further, because it was not possible to survey SIOP members and fellows a second time on the same beliefs, research could only be based on data from a preexisting survey. This survey had at least two content problems that cast doubt upon the reliability of responses. As a result, it may be that no model can be validated properly by testing it on the survey data.

The first problem is a matter of the way in which item content was stated. While 22 items specifically used the term "general cognitive ability," 20 referred to "cognitive ability." In some cases, the interchangeable use of the terms "general cognitive ability" and "cognitive ability" might have led respondents to differentiate between the two, thinking the latter term referred to aptitudes. In some cases, it is conceivable that this ambiguity could have had some impact on interpretation, however slight. For example, the assertion in item #34 reads, "In jobs where cognitive ability is important, Blacks are likely to be underrepresented." A respondent who believes that the use of general cognitive ability tests necessarily leads to adverse impact might have expressed agreement if the term "cognitive ability" was interpreted as GCA. However, if that same respondent believes that tests of specific abilities or other intelligences *might not*

necessarily lead to adverse impact [as Kehoe (2002) suggests], she or he might have expressed disagreement.

The second problem may be more important. The meanings of certain “Tradeoffs” items were equivocal. An example of this problem arises in item #26, which reads, “The use of cognitive ability tests in selection leads to more social justice than their abandonment.” A respondent who believes that “social justice” refers to fairness toward groups might answer “strongly disagree”; however, if that same respondent were instructed to consider “fairness” at individual level, she or he might have answered “strongly agree”. Note that this item is also subject to the same potential for multiple interpretations of the first type identified.

Finally, the survey results were obtained from a broad sample of SIOP members, including both academics and practitioners. Typically, only experts are involved in debates and panel discussions, but many of those who responded may not have been deeply knowledgeable about GCA and GCA testing. While covariation may not have differed greatly among experts and nonexperts, further research should be conducted to examine whether there is evidence that the model holds for experts before using it in debate.

CHAPTER 7

RECOMMENDATIONS FOR FUTURE RESEARCH

While discussion forums such as the “Millennial Debate on *g*” may provoke intellectual exchange and stimulation, they do not qualify as formal debate. This distinction is more than semantic. It is important to recognize what “debate” is—its purpose, its rules and its value. Debate, fundamentally, is a form of argumentation (Patterson & Zarefsky, 1983; van Eemeren et al., 1996) in which parties cooperatively investigate disputed subjects by making appeals to a third-party judge, by whose decision they agree to abide (Ehninger & Brockriede, 2008). Pera (1994) states that, in scientific debate, rules of conduct and judgment should be formalized, and the soundness of arguments must be established within the context of that debate. This kind of debate might be highly unusual for industrial-organizational psychologists, who appear to allow the weight of evidence simply determine the direction of the field over periods of time. According to Pera (1994), however, it is through debate that arguments are sharpened and, ultimately, that valid science can be determined. Industrial-Organizational psychologists need not fear a polarizing effect in proper debate. Debate should be a cooperative effort with the objective of illuminate issues, rather than winner-take-all (Ehninger & Brockriede, 2008). The model advanced in this study might serve as a framework for such a formal debate on GCA, the purpose of which would be to advance a common paradigm.

One of Murphy et al. (2003) demographic characteristics was “Primary Employment”, segmented into “Academic”, “Consulting”, “Industry”, and “Other”. It might be interesting to explore whether the six-factor model holds for human resources practitioners, or if another model is preferable. This could be initially tested through Murphy et al.’s (2003) data, by using those

characteristics. Also, members of the Society for Human Resources Management (SHRM) seem not to have been surveyed on their beliefs in GCA in the past; they are natural practitioner respondents. The Murphy et al. (2003) survey (or a modified form thereof) might be administered to them. Another approach might employ surveys of the type that Naess (1966) describes (i.e., pro-et-contraria and pro-aut-contraria), allowing evidence to be judged by respondents. A Likert scale could supplant the typical “yes/no” response format, providing data upon which factor models could be tested. This might be more informative than a traditional study, as it may contribute a practitioner view of the weight of evidence that Murphy et al. (2003) were not able to identify, even among the practitioners in their sample.

Finally, other models and other characterizations should be explored. Intuitively, the present study’s “Primacy”, “Alternatives” and “Criterion Dependence” factors may be related to Murphy et al.’s (2003) “g-ocentric” factor. The present study’s “Test Bias”, “Social Fairness” and “Tradeoffs” factors may be related to Murphy et al.’s (2003) “Societal” factor. A second-order model, with the present study’s factors loading onto the Murphy et al. (2003) factors according to these relationships, might achieve better fit and could be more informative.

REFERENCES

- Allix, N. M. (2000). The theory of multiple intelligences: A case of missing cognitive matter. *Australian Journal of Education, 44*(3), 272-288.
- Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees*. Reading, MA: Addison-Wesley.
- Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49*, 141-168.
- Barchard, K. A. (2003). Does emotional intelligence assist in the prediction of academic success?. *Educational and Psychological Measurement, 63*(5), 840-858.
- Baron, J. (1986). Capacities, dispositions, and rational thinking. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 29-33). Norwood, NJ: Ablex.
- Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31*(2), 233-241.
- Bateman T. S. & Organ D. W. (1984). Job satisfaction and the good soldier: The relationship between affect and employee "citizenship". *Academy of Management Journal, 26*, 587-595.
- Becker, T. (2003). Is emotional intelligence a viable concept?. *The Academy of Management Review, 28*(2), 192-195.
- Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

- Besetsny, L. K., Earles, J. A., & Ree, M. J. (1993). Little incremental validity for a special test of abstract symbolic reasoning in an Air Force intelligence career field. *Educational and Psychological Measurement*, 53(4), 993-997.
- Besetsny, L. K., Ree, M. J., & Earles, J. A. (1993). Special test for computer programmers? Not needed: The predictive efficiency of the Electronic Data Processing Test for a sample of Air Force recruits. *Educational and Psychological Measurement*, 53(2), 507-511.
- Berry, J.W. (1986). A cross-cultural view of intelligence. In R. Sternberg and D. K. Detterman (Eds). *What is intelligence?: Contemporary viewpoints on its nature and definition*, (pp. 35-38). Norwood, NJ: Ablex.
- Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper & Brothers.
- Bishop, J. H. (1988). Employment testing and incentives to learn. *Journal of Vocational Behavior*, 33(3), 404-423.
- Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, 52(3), 561-589.
- Boring, E. G. (1923). Intelligence as the Tests Test It. *New Republic*, 36. 35-37.
- Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. Chapter in J. T. Spence (Ed.). *Annual Review of Psychology* (pp. 299-337). Palo Alto, CA: Annual Reviews, Inc.
- Borman, W. C., & Motowidlo, S. J. (1993b). Expanding the criterion domain to include elements of contextual performance. In W. C. Borman (Ed.), *Personnel Selection in Organizations* (pp. 71-98). New York, NY: Jossey-Bass.

- Borman, W. C., & Motowidlo, S. J. (1997). Introduction: Organizational citizenship behavior and contextual performance. Special issue of *Human Performance*, Borman, W. C., & Motowidlo, S. J. (Eds.). *Human Performance*, 10, 67-69.
- Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983). *A model of individual performance effectiveness: Thoughts about expanding the criterion space*. Paper presented at part of symposium, Integrated Criterion Measurement for Large Scale Computerized Selection and Classification, the 91st Annual American Psychological Association Convention. Washington, DC.
- Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment*, 9, 52-69.
- Brand, C. (1996). *The g factor: General intelligence and its implications*. Chichester, UK: Wiley.
- Branham, R. J. (1991). *Debate and critical analysis: the harmony of conflict*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review*, 11 (4), 710-725.
- Burt, C. (1949). The structure of the mind. *British Journal of Educational Psychology*, 19(3), 176-199.
- Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). An investigation of sources of bias in the prediction of job performance: A six-year study. (Final Project Report PR-73-37). Princeton, NJ: Educational Testing Service.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

- Campbell, J. P., McHenry, J. J. and Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, 43, 313-333.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.
- 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(1), 143-159.
- Chemiss, C. (2000). *Emotional intelligence: What it is and why it matters*. Paper presented at the Annual Meeting for the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit statistics for testing MI. *Structural Equation Modeling*, 9, 235-255.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2), 115-124.
- Cleary, T., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, 28. 61-75
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum. (Retrieved from <http://www.lrdc.pitt.edu/schneider/P2465/Readings/Cohen,%201988%20%28Statistical%20Power,%20273-406%29.pdf>)
- Cole, N. S. (1973). Bias in selection. *Journal of Educational Management*, (10)4, 237-255.
- Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by *g*. *Intelligence*, 32(3), 277-296.

- Collins, V. L. (2001). Emotional intelligence and leader success. Unpublished doctoral dissertation, University of Nebraska. Lincoln, NE. [Retrieved from DigitalCommons@University of Nebraska-Lincoln]
- Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment centre dimensions, personality and aptitudes. *Journal of Occupational Psychology*, 63(3), 211-216.
- Cronbach, L.J. (1949). *Essentials of psychological testing*. New York, NY: Harper.
- Cronshaw, S. F., Hamilton, L. K., Onyura, B. R., & Winston, A. S. (2006). Case for non-biased intelligence testing against Black Africans has not been made: A comment on Rushton, Skuy, and Bons (2004). *International Journal of Selection and Assessment*, 14(3), 278-287.
- Darlington, R. B. (1971). Another look at “cultural fairness”. *Journal of Educational Measurement*, 8(2), 71-82.
- Darlington, R. B. (2004). *Factor Analysis*. Retrieved from <http://www.unt.edu/rss/class/mike/Articles/FactorAnalysisDarlington.pdf>
- Das, J. P. (1986). On definition of intelligence. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?: Contemporary viewpoints on its nature and definition* (pp. 55-56). Norwood, NJ: Ablex.
- Das, J. P., Kirby, J., & Jarman, R. F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychology Bulletin*, 82(1), 87-103.
- Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Needham Heights, MA: Allyn & Bacon.
- Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology*, 75(4), 989-1015

- Dawis, R. V. (1992). The individual differences tradition in counseling psychology. *Journal of Counseling Psychology*, 39(1), 7–19.
- Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology*, 42(1), 25-36.
- De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, 92(5), 1380-1393.
- De Corte, W., Lievens, F., & Sackett, P. R. (2008). Validity and adverse impact potential of predictor composite formation. *International Journal of Selection and Assessment*, 16(3), 183-194.
- De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology*, 96(5), 907-926.
- Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, 11, 201–211.
- DeShon, R. P., Smith, M. R., Chan, D., & Schmitt, N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context?. *Journal of Applied Psychology*, 83(3), 438-451.
- Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, 13(4), 349-359.
- Dozier, J. B., & Miceli, M. P. (1985). Potential predictor or whistle-blowing: A prosocial behavior perspective. *Academy of Management Review*, 10, 823-836.

- Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72(1), 19-29.
- DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and White-Black differences. *Journal of Applied Psychology*, 78(2), 205-211.
- Edwards, J. R. (2003). *Construct validation in organizational behavior research*. In J. Greenberg (Ed.), *Organizational behavior: The state of the science* (pp. 327-331). Mahwah, NJ: Erlbaum.
- Ehninger, D., & Brockriede, W. (2008). *Decision by debate*. New York, NY: International Debate Education Association.
- Eysenck, H. J. (1990/1997). *Rebel with a cause: The autobiography of Hans Eysenck*. New Brunswick, NJ: Transaction Publishers.
- Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence*, 30(4), 361-387.
- Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, 35(4), 319-334.
- Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin*, 99(3), 330-337.
- Gael, S., & Grant, D. L. (1972). Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology*, 56(2), 135-139.

- Gael, S., Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology*, 60(4), 411-419.
- Gael, S., Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology*, 60(4), 420-426.
- Gardner, H. (1983/2003). *Frames of mind. The theory of multiple intelligences*. New York, NY: Basic Books.
- Gardner, H. (2004). Audiences for the theory of multiple intelligences. *The Teachers College Record*, 106(1), 212-220.
- Grant, D. L., & Bray, D. W. (1970). Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology*, 54(1 pt. 1), 7-14.
- Goldstein, B. L., & Patterson, P. O. (1988). Turning back the title VII clock: The resegregation of the American work force through validity generalization. *Journal of Vocational Behavior*, 33(3), 452-462.
- Goldstein, H. W., Scherbaum, C. A., & Yusko, K. P. (2010). Revisiting *g*: Intelligence, adverse impact, and personnel selection. In J. Outtz (Ed.). *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 95-134). San Francisco: Jossey-Bass.
- Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology*, 51(2), 357-374.
- Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). *g*: Is this your final answer? *Human Performance*, 15(1/2), 123-142.
- Goleman, D. (1998). *Working with emotional intelligence*. New York, NY: Bantam Books.

- Gottfredson, L. S. (1986b). Societal consequences of the g factor in employment. *Journal of Vocational Behavior*, 29, 379-410.
- Gottfredson, L. S. (Ed.) (1986). The g factor in employment. *Journal of Vocational Behavior*, 29 (3), 293-296.
- Gottfredson, L. S. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior*, 33(3), 293-319.
- Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24(1), 13-23.
- Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, 15(1/2), 25-46.
- Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence*, 31(4), 343-397.
- Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, 64, 1029-1041.
- Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, 53(4), 267-293.
- Guilford, J.P. (1967). *The Nature of Human Intelligence*. New York, NY: McGraw Hill Book Company.
- Guilford, J.P. (1977). *Way beyond the IQ*. Buffalo, NY: Creative Education Foundation, Inc.
- Guilford, J.P. (1985). The structure-of-intellect model. In B.B. Wolman (Ed.), *Handbook of intelligence* (pp. 225-266). New York, NY: Wiley.
- Guion, R. M. (1997). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Haggerty, N.E. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, 12(4), 212-216
- Hampshire A., Highfield R. R., Parkin B. L., Owen A. M. (2012). Fractionating human intelligence. *Neuron*, 76, 1225–1237
- Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.
- Hatrup, K., & Jackson, S. E. (1996). Learning about individual differences by taking situations seriously. In K. R. Murphy (Ed.). *Individual differences and behavior in organizations* (pp. 507-547). San Francisco, CA: Jossey-Bass.
- Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing?. *American Psychologist*, 47(9), 1083-1101.
- Helms, J. E. (1993). I also said" white racial identity influences white researchers.”. *The Counseling Psychologist*, 21(2), 240-243.
- Herrnstein, R. J. & Murray. C. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.
- Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology*, 39(2), 399-420.
- Hogan, R. (1983). A socioanalytic theory of personality. In M. M. Page (Ed.), 1982 *Nebraska Symposium on Motivation* (pp. 55– 89). Lincoln, NB: University of Nebraska Press.
- Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality* (pp. 163–179). New York, NY: Guilford Press.

- Hogan, J., Rybicki, S. L., Motowidlo, S. J., & Borman, W. C. (1998). Relations between contextual performance, personality, and occupational advancement. *Human Performance*, 11(2-3), 189-207.
- Holzinger, K. J. (1934, 1935). *Preliminary report on the Spearman-Holzinger unitary traits study*, No. 1-9. Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.
- Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic and developmental study of the structure among primary mental abilities* (Unpublished doctoral dissertation). University of Illinois, Champaign. [Retrieved from <http://www.iqscorner.com/2009/07/john-horn-1965-doctoral-dissertation.html>]
- Horn, J. L. (1976). Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology*, 27, 437-485.
- Horn, J. L. (1994). The theory of fluid and crystallized intelligence. In R.J. Sternberg (Ed.). *The Encyclopedia of human intelligence* (Vol. 1, pp. 443-451). New York, NY: Macmillan.
- Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, 26, 107-129.
- Horn, R. V. (1993). *Statistical Indicators for the Economic and Social Sciences*. Cambridge, UK: Cambridge University Press.
- 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(1-2), 152-194.
- Houston, W. M., & Novick, M. R. (1987). Race-Based differential prediction in Air Force technical training programs. *Journal of Educational Measurement*, 24(4), 309-320.

- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structural analysis: Conventional versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Hull, C. L. (1928). *Aptitude testing*. Yonkers, NY: World Book Co.
- Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence*, 3(2): 105–120.
- Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology*, 71(2), 327-333.
- Hunter, J. E. (1980c). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Employment Service, U.S. Department of Labor.
- Hunter, J. E. (1981b). *Fairness of the General Aptitude Test Battery (GATB): Ability differences and their impact on minority hiring rates*. Washington, DC: U.S. Employment Service.
- Hunter, J. E. (1983a). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance and measurement theory* (pp. 257–266). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Hunter, J. E. (1983b). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance for the USES* (Test Research Report No. 44). Washington, DC: U.S. Employment Services, U.S. Department of Labor.
- Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper, I. T. Robertson, Ivan T. (Eds.). *International review of industrial and organizational psychology* 1987, (pp. 321-357). Oxford, England: John Wiley & Sons
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72–98.

- Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, 83(6), 1053-1071
- Hunter, J. E., & Schmidt, F. L. (1977). A critical analysis of the statistical and ethical definitions of test fairness. *Psychological Bulletin*, 83, 1053–1071.
- Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law*, 2(2), 324-347.
- Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law*, 2(3-4), 447-477.
- Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, 91, 594-612.
- Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. E. Reynolds and R. T. Brown (Eds.), *Perspectives on bias in mental testing*. New York, NY: Plenum.
- Ironson, G. & M. Suboviak (1979). A comparison of several methods of assessing bias. *Journal of Educational Measurement*, 16(4), 209-226.
- Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.
- Jensen, A. R. (1987). Black-White bias in “cultural” and “noncultural” test items. *Personality and Individual Differences*, 8(3), 295-301.
- Jensen, A. R. (1993). Test validity: *g* versus “tacit knowledge”. *Current Directions in Psychological Science*, 2(1), 9–10.
- Jensen, A. R. (1996). The *g*-factor. *Nature*, 381(6585), 729-729.

- Jensen, A. R., & McGurk, F. C. (1987). Black-white bias in “cultural” and “noncultural” test items. *Personality and Individual Differences*, 8(3), 295-301.
- Johnson, W., Bouchard Jr, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence*, 32(1), 95-107.
- Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Education Resources.
- Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, 95(1), 54-78.
- Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2005). A history of intelligence test interpretation. In D.P. Flanagan and P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed; pp. 23–38). New York, NY: Guilford.
- Katz, D. (1964). The motivational basis of organizational behavior. *Behavioral Science*, 9, 131-133.
- Kehoe, J. F. (2002). General mental ability and selection in private sector organizations: A commentary. *Human Performance*, 15(1/2), 97-106.
- Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differential mental abilities*. Stanford, CA: Stanford University Press.
- Kenny, D. A. (2014). *Measuring Model Fit*. Retrieved from <http://davidakenny.net/cm/fit.htm#RMSEA>
- Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, 70(1), 56-65.
- Kuhn, T. S. (1962/1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

- Kuncel, N. R., Campbell, J. P., & Ones, D. S. (1998). Validity of the Graduate Record Examination: Estimated or tacitly known? *American Psychologist*, 53(5), May 1998, 567-568.
- Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, 19(6), 339-345.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127(1), 162–181.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all?. *Journal of personality and Social Psychology*, 86(1), 148-161.
- Kuncel, N. R., & Sackett, P. R. (2007). Selective citation mars conclusions about test validity and predictive bias: A comment on Vasquez and Jones. *American Psychologist*, 6(2), 145-146.
- Landy, F. J. (2005). Some historical and scientific issues related to research on emotional intelligence. *Journal of Organizational Behavior*, 26(4), 411-424.
- Le, H., Oh, I.-S., Shaffer, J., & Schmidt, F. L. (2007). Implications of methodological advances for the practices of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives*, 21, 6-15.
- Lefkowitz, J. (1994). Race as a factor in job placement: Serendipitous findings of “ethnic drift”. *Personnel Psychology*, 47(3), 497-513.
- Legree, P. J., Pifer, M. E., & Grafton, F. C. (1996). Correlations among cognitive abilities are lower for higher ability groups. *Intelligence*, 23(1), 45-57.
- Levin, H. M. (1988). Issues of agreement and contention in employment testing. *Journal of Vocational Behavior*, 33(3), 398-403.

- Levine, E. L., Spector, P. E., Menon, S., Narayanan, L., & Cannon-Bowers, J. (1996). Validity generalization for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance*, 9, 1–22.
- Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63(4), 507-512.
- Loehlin, J. C., Lindzey, G., & Spuhler, J. N. (1975). *Race differences in intelligence*. San Francisco, CA: Freeman.
- Lopes, P. N., Salovey, P., Côté, S., & Beers, M. (2005). Emotion regulation abilities and the quality of social interaction. *Emotion*, 5(1), 113-118.
- Locke, E. A. (2005). Why emotional intelligence is an invalid concept. *Journal of Organizational Behavior*, 26(4), 425-431.
- MacCann, C, Joseph, D. L., Newman, D. A., Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, 14(2), 358-374.
- Matthews, G., Zeidner, M., & Roberts, R. D. (2002/2004). *Emotional intelligence: Science and myth*. Cambridge, MA: MIT Press.
- Matthews, G., Zeidner, M. & Roberts, R. D. (2004). Emotional intelligence: An elusive ability?. O. Wilhelm and R. Engle (Eds.) *Handbook of understanding and measuring intelligence* (pp. 79-99). Thousand Oaks, CA: Sage Publications.
- Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27(4), 267-298.
- Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence*, 17(4), 433-442.

- Mayer, J. D. & Salovey, P. (1997). What is emotional intelligence? In P. Salovey and D. Sluyter (Eds.). *Emotional development and emotional intelligence: educational implications* (pp. 3-31). New York, NY: Basic Books.
- Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Emotional intelligence: Theory, findings, and implications. *Psychological inquiry*, 15(3), 197-215.
- McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988b). Job experience correlates of job performance. *Journal of Applied Psychology*, 73, 327-330.
- McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, 43(2), 335-354.
- McKay, P. F., & McDaniel, M. A. (2006). A reexamination of black-white mean differences in work performance: more data, more moderators. *Journal of Applied Psychology*, 91(3), 538-554.
- Medley, D. M., & Quirk, T. J. (1974). The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examination. *Journal of Educational Measurement*, 11(4), 235-245.
- Mehler, B. (1997). Beyondism: Raymond B. Cattell and the new eugenics. *Genetica*, 99, 153-163.
- Messick, S. (1992). Multiple intelligences or multilevel intelligence? Selective emphasis on distinctive properties of hierarchy: On Gardner's Frames of Mind and Steinberg's Beyond IQ in the context of theory and research on the structure of human abilities. *Psychological Inquiry*, 3(4), 365-384.
- Miller, R. L., Griffin, M. A., & Hart, P. M. (1999). Personality and organizational health: The role of conscientiousness. *Work & Stress*, 13(1), 7-19.

- Minsky, M. L. (1985). *The society of mind*. New York: Simon & Schuster.
- Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, 10(2), 71-83.
- Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, 79(4), 475.-480.
- Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). Employee-organization linkages: The psychology of commitment, absenteeism, and turnover. New York: Academic Press.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105(3), 430-445.
- 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. R. (2002). Can conflicting perspective on the role of *g* in personnel selection be resolved. *Human Performance*, 15(1/2), 173-186.
- Murphy, K. R., Cronin, B. E., & Tam A. P (2003). Controversy and consensus regarding the use of cognitive ability testing in organizations. *Journal of Applied Psychology*, 88, 660–671.
- Naess, A. (1966). *Communication and argument. Elements of applied semantics*. London, UK: Allen and Unwin. (Published in Norwegian in 1947).
- Naess, A. (1992b). How can the empirical movement be promoted today? A discussion of the empiricism of Otto Neurath and Rudolp Carnap. In E. M. Barth, J. Vandormael and F.

- Vandamme (Eds.). *From an empirical point of view: The empirical turn in logic* (pp. 107-155). Ghent, Belgium: Communication & Cognition.
- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). *Intelligence: Knowns and unknowns. American Psychologist*, 51(2), 77-101.
- Neuman, G. A., & Kickul, J. R. (1998). Organizational citizenship behaviors: Achievement orientation and personality. *Journal of Business and Psychology*, 13(2), 263-279.
- Newman, D. A., Hanges, P. J., & Outtz, J. L. (2007). Racial groups and test fairness, considering history and construct validity. *American Psychologist*, 62, 1082-1083.
- Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology*, 79(6), 845-851.
- Ones, D. S., & Viswesvaran, C. (2000). *The Millennial Debate on "g" in I-O Psychology*. Annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Organ, D. W. (1977). A reappraisal and reinterpretation of the satisfaction-causes-performance hypothesis. *Academy of Management Review*, 2, 46-53.
- Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books/D.C. Heath and Company.
- Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology*, 48(4), 775-802.
- Outtz, J. L. (2002). The role of cognitive ability tests in employment selection. *Human Performance*, 15(1/2), 161-171.

- Outtz, J. L., & Newman, D. A. (2010). "A theory of adverse impact". In J. Outtz (Ed.). *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 53-94). San Francisco, CA: Jossey-Bass.
- Patterson, J. W., & Zarefsky, D. (1983). Contemporary debate. Boston, MA: Houghton Mifflin.
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology*, 65, 373-407.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). Definitions and variables. In E. J. Pedhazur and L. P. Schmelkin (Eds.), *Measurement, design, and analysis: An integrated approach*. (pp. 164-179). Mahwah, NJ: Lawrence Erlbaum.
- Pera, M. (1994). *The discourses of science*. (C. Botsford, Trans.). Chicago, IL: University of Chicago Press. (Original work published in 1991).
- Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*, 74(5), 770.
- Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19(4), 353-368.
- Ree, M. J. & Carretta, T. R. (2002). G2K. *Human Performance*, 15(1/2), 3-23.
- Ree, M. J. & Earles, J. A. (1991a). The stability of convergent estimates of *g*. *Intelligence*, 15, 271-278.
- Ree, M. J., & Earles, J. A. (1991b). Predicting training success: Not much more than *g*. *Personnel Psychology*, 44, 321-332.

- Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science*, 1(3), 86-89.
- Ree, M. J., & Earles, J. A. (1993). g is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science*, 2(1), Feb 1993, 11-12.
- Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology*, 79(4), 518-524.
- Reeve, C. L. & Hakel, M. D. (2002). Asking the right questions about g. *Human Performance*, 15(1/2), 47-74.
- Reynolds, C. R., Chastain, R. L., Kaufman, A. S., & McLean, J. E. (1987). Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology*, 25(4), 323-342.
- Reynolds, C. R., & Jensen, A. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, 3(4), 423-438.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, 54(2), 297-330.
- Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: a new meta-analysis. *Journal of Applied Psychology*, 88(4), 694-706.
- Rotundo, M., & Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology*, 84(5), 815-822.
- Rushton, J. P., & Jensen, A. R. (2005). Wanted: More race realism, less moralistic fallacy. *Psychology, Public Policy, and Law*, 11(2), 328-336

- Ryanen, I. A. (1988). Commentary of a minor bureaucrat. *Journal of Vocational Behavior*, 33(3), 379-387.
- Sackett, P. R., De Corte, W., & Lievens, F. (2008). Pareto-Optimal Predictor Composite Formation: A complementary approach to alleviating the selection quality/adverse impact dilemma. *International Journal of Selection and Assessment*, 16(3), 206-209.
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56(4), 302-318.
- Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, 49(11), 929-954.
- Salgado, J. F., & Anderson, N. (2002). Cognitive and GMA testing in the European Community: Issues and evidence. *Human Performance*, 15(1/2), 75-96.
- Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, 27(2), 109-131.
- Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior*, 33(3), 272-292.
- Schmidt, F. L. (1993). Personnel psychology at the cutting edge. In N. Schmitt & W. Borman (Eds.). *Personnel selection* (pp. 497-515). San Francisco, CA: Jossey-Bass.
- Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, 15(1/2), 187-210.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology* 62, 529-540.

- Schmidt, F. L., & Hunter, J. E. (1981). *Employment testing: Old theories and new research findings. American Psychologist, 36*(10), 1128-1137.
- Schmidt, F. L., & Hunter, J. E. (1992). Causal modeling of processes determining job performance. *Current Directions in Psychological Science, 1*, 89–92.
- Schmidt, F. L. & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2*(1), 8-9.
- 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*(2), 262–274.
- Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183-198.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt–Hunter Bayesian validity generalization model. *Personnel Psychology, 32*, 257–281.
- Schmidt, F. L.; Pearlman, K.; Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33*(4), 705-724.
- Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. R. Cooper and I. T. Robertson (Eds.). *International Review of Industrial and Organizational Psychology, 11*, 115-140. New York, NY: Wiley
- Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*(1), 67-81.

- Schmitt, N., Hatrup, K., & Landis, R. S. (1993). Item bias indices based on total test score and job performance estimates of ability. *Personnel Psychology*, 46(3), 593-611.
- Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407-422.
- 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(5), 719-730.
- Scholarios, D. M., Johnson, C. D., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology*, 79(3), 412-424.
- Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional intelligence: not much more than *g* and personality. *Personality and Individual Differences*, 37(5), 1059-1068.
- Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization". *Journal of Vocational Behavior*, 33(3), 331-364.
- Sharf, J. C. (1988). Litigating personnel measurement policy. *Journal of Vocational Behavior*, 33(3), 235-271.
- Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, 58, 935-943.
- Simonton, D. K. (2003). An interview with Dr. Simonton. In J. A. Plucker (Ed.), *Human intelligence: Historical influences, current controversies, teaching resources*.
http://www.intelltheory.com/simonton_interview.shtml

- Sireci, S. G., & Geisinger, K. F. (1998). Equity issues in employment testing. In J.H. Sandoval, C. Frisby, K.F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.), *Test interpretation and diversity* (pp. 105-140). Washington, D.C.: American Psychological Association.
- Smith, C. A., Organ, D. W., & Near, J. P. 1983. Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, 68, 655– 663.
- Spearman, C. E. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Spearman, C. E. (1923). *The nature of 'intelligence' and the principles of cognition* (2nd ed.). London, UK: Macmillan.
- Spearman, C. E. (1927). *The Abilities of Man*. London: Macmillan.
- Stauffer, J. M., & Buckley, M. R. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology*, 90(3), 586-591.
- Staw, B. M. (1984). Organizational behavior: Review and reformulation of the field's outcome variables. *Annual Review of Psychology*, 35(1), 627- 667.
- Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of factors*. Paper presented and the annual meeting of the Psychometric Society, Iowa City, IA.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.
- Sternberg, R. J. (1987). Intelligence. In R. L. Gregory (Ed.). *The Oxford companion to the mind* (pp. 375-379). Oxford, UK: Oxford University Press.
- Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence*. New York: Viking.

- Sternberg, R. J. (1996). The triarchic theory of intelligence. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.). *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 92-104). New York, NY: Guilford.
- Sternberg, R. J. (1997). *Successful Intelligence*. New York, NY: Plume.
- Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology*, 3, 292-316.
- Sternberg, R. J. (2000). *Handbook of intelligence*. Cambridge: Cambridge University Press.
- Sternberg, R. J. (2003). An interview with Dr. Sternberg. In J. A. Plucker, (Ed.), *Human intelligence: Historical influences, current controversies, teaching resources*.
http://www.intelltheory.com/sternberg_interview.shtml
- Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, 34(4), 321-350.
- Sternberg, R. J., & Berg, C. A. (1986). Quantitative integration: Definitions of intelligence: A comparison of the 1921 and 1986 symposia. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 155-162). Norwood, NJ: Ablex.
- Sternberg, R. J., Grigorenko, E. L., & Kidd, K. K. (2005). Intelligence, race, and genetics. *American Psychologist*, 60(1), 46-59.
- Sternberg, R. J. & Hedlund, J. (2002). Practical intelligence, *g*, and work psychology. *Human Performance*, 15(1/2), 143-160.
- Sternberg, R. J., Nokes, K., Geissler, P.W., Prince, R., Okatcha, F., Bundy, D. A., & Grigorenko, E. L. (2001). The relationship between academic and practical intelligence: A case study in Kenya. *Intelligence*, 29(5), 401-418.

- Taylor, R. L. (1991). Bias in cognitive assessment: Issues, implications, and future directions. *Diagnostique*, 17(1), 3-5.
- Tenopyr, M. L. (2002). Theory versus reality: Evaluation of *g* in the workplace. *Human Performance*, 15(1/2), 107-122.
- Thorndike, E. L. (1921). On the Organization of Intellect. *Psychological Review*, 28(2), 141-151.
- Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement*, 82(2), 63-70.
- Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior*, 29, 322-339.
- Thurstone, L.L. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, 16: 201-207.
- Thurstone, L.L. (1924). *The nature of intelligence*. London: Routledge.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.
- Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of the vectors of the mind*. Chicago, IL: University of Chicago Press.
- Van Eemeren, F. H., Grootendorst, R., Snoeck-Henckemans, F. S., Blair, J.A., Johnson, R. H., Krabbe, E. C. W., Plantin, C., Walton, D. N., Willard, C. A., Woods, J. & Zarefsky, D. (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Mahwah, NJ; Lawrence Erlbaum Associates.
- Vernon, P. E. (1950/1961). *The structure of human abilities*. London, UK: Methuen.
- Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences*, 5(4), 411-423.

- Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006). Beyond *g*: Putting multiple intelligences theory to the test. *Intelligence*, 34(5), 487-502.
- 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), 211-231.
- Vroom, V. H. (1964). *Work and motivation*. New York, NY: Wiley.
- Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology*, 52, 1236–1247.
- Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology*, 49(2), 436–458.
- Wallace, S. R. (1973). Sources of bias in the prediction of job performance: Implications for future research. In L. A. Crooks (Ed.), *An investigation of sources of bias in the prediction of job performance*. Princeton, NJ: Educational Testing Service.
- Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous *and* heterogeneous scales: An illustration with the MMPI. *Psychological Methods*, 5(1), 125-146.
- Waterhouse, L. (2006). Inadequate evidence for multiple intelligences, Mozart effect, and emotional intelligence theories. *Educational Psychologist*, 41(4), 247-255.
- Waterhouse, L. (2006). Multiple intelligences, the Mozart effect, and emotional intelligence: A critical review. *Educational Psychologist*, 41(4), 207-225.
- Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed). Baltimore, MD: Williams & Wilkins.

- Wendelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon?. *Journal of Applied Psychology*, 66(2), 149-158.
- Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies: Part 1. Report of the committee*. Washington, DC: National Academy Press.
- Williams, W. M., & Ceci, S. J. (1997). Are Americans becoming more or less alike? Trends in race, class, and ability differences in intelligence. *American Psychologist*, 52(11), 1226-1235.
- Willingham, W. W. (1999). A systemic view of test fairness. In S. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 213-242). Mahwah, NJ: Lawrence Erlbaum Associates.
- Yusko, K., Goldstein, H. W., Oliver, L. O., & Hanges, P. J. (2010). Cognitive ability testing with reduced adverse impact: Controlling for knowledge. In C. J. Paullin (Chair): *Cognitive Ability Testing: Exploring New Models, Methods, and Statistical Techniques*. Symposium presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.
- Zeidner, J., & Johnson, C. D. (1994). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Walker and J. H. Harris (Eds.). *Personnel selection and classification* (pp. 377-410). Hillsdale, NJ: Erlbaum.
- Zeidner, J., Johnson, C. D., & Scholarios, D. (1997). Evaluating military selection and classification systems in the multiple job context. *Military Psychology*, 9(2), 169-186.
- Zeidner, M., Matthews, G., & Roberts, R. D. (2004). Emotional intelligence in the workplace: A critical review. *Applied psychology: An international review*, 53(3), 371-399.

APPENDIX A

Items selected for model specification are grouped according to the six factors—

Primacy:

- #1 There is no substitute for general cognitive ability.
- #2 General cognitive ability is the most important individual difference variable.
- #10 General cognitive ability should almost always be a part of a personnel selection system.
- #11 General cognitive ability enhances performance in all domains of work.
- #14 The validity of cognitive ability tests show levels of validity too low to justify the negative social consequences of those tests
- #28 Cognitive ability tests should almost always be a part of a personnel selection system
- #29 Tests of noncognitive traits are useful supplements to *g*-loaded tests in a selection battery, but they cannot substitute for tests of *g*.
- #45 Average scores on general cognitive ability tests are related to the effectiveness of an organization.

Criterion Dependence:

- #16 Although cognitive ability tests are the best predictors of technical or core performance, they are not the best predictors of other facets of job performance
- #17 The predictive validity of cognitive ability tests depends on how performance criteria are defined and measured.

- #33 The multidimensional nature of job performance necessitates the use of both cognitive and noncognitive selection measures.

Alternatives:

- #4 There is more to intelligence than what is measured by a standard cognitive ability test.
- #5 General cognitive ability accounts almost totally for the predictive validity of ability tests.
- #9 Different jobs are likely to require different types of cognitive abilities.
- #15 Tacit knowledge is a form of practical intelligence, which explains aspects of performance that are not accounted for by *g*.
- #30 Combinations of specific aptitude tests have little advantage over measures of general cognitive ability in personnel selection.
- #31 Tacit knowledge contributes over and above general cognitive ability to the prediction of job performance.

Test Bias:

- #13 General cognitive ability tests measure constructs that are not required for successful job performance.
- #39 Racial differences produced by cognitive ability tests are substantially higher than racial differences on measures of job performance.
- #49 A workforce selected on the basis of actual performance would be less racially segregated than a workforce selected on the basis of cognitive ability tests.

Social Fairness:

- #26 The use of cognitive ability tests in selection leads to more social justice than their abandonment.
- #27 Tests of general cognitive ability can be used to create equal opportunities for all.
- #36 Professionally developed cognitive ability tests are not biased against members of racial or ethnic minority groups.
- #37 General cognitive ability tests are fair

Tradeoffs:

- #19 There are combinations of noncognitive measures with criterion-related validity comparable to those achieved by cognitive ability tests.
- #22 Choosing to use cognitive ability tests implies a willingness to accept the social consequences of racial discrimination.
- #43 Massive societal changes will be necessary to significantly affect the adverse effects of cognitive ability tests.
- #47 There is a tradeoff between the cost-effective use of cognitive ability tests and social responsibility in selection practices.