ABSTRACT

ALLRED, CHRISTOPHER MICHAEL. Applying a Metaheuristic Algorithm to a Multi-Objective Optimization Problem Within Personnel Psychology. (Under the direction of Dr. Adam W. Meade).

Standardized assessments of general cognitive ability (GCA) consistently outperform alternative predictors of job performance, making their inclusion within employment selection batteries valuable to organizations. Unfortunately, GCA test scores typically produce observed racioethnic group differences that can result in different selection rates for majority and minority group members. Organizations seeking to maximize productivity via cognitively loaded selection procedures while simultaneously attempting to build a diverse workforce are faced with a challenging dilemma. The present study addresses this diversity-validity dilemma by applying methods typically employed outside of psychological science. Specifically, predictor and criterion data are simulated under a variety of job complexity conditions to explore the utility of applying a metaheuristic algorithm to differentially weight items such that criterion-related validity is maximized while mean racioethnic group differences are minimized. Findings indicate that a metaheuristic algorithm can lower group differences in GCA test scores, but that these reductions are accompanied by reductions in criterion-related validity. Limitations and future directions are provided.
Applying a Metaheuristic Algorithm to a Multi-Objective Optimization Problem Within Personnel Psychology

by

Christopher Michael Allred

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

Psychology

Raleigh, North Carolina 2019

APPROVED BY:

_______________________________
Dr. Adam W. Meade
Committee Chair

_______________________________
Dr. S. Bartholomew Craig

_______________________________
Dr. Samuel Pond, III

_______________________________
Dr. Mark. A. Wilson
Christopher Michael Allred completed his undergraduate studies in May of 2009 earning a Bachelor of Arts in Psychology with a Minor in Religious Studies from North Carolina State University. He then obtained his Master of Arts in General/Theoretic Psychology with a concentration in Industrial/Organizational Psychology from East Carolina University in May of 2012. Returning to North Carolina State University in August of 2012, he enrolled in the Industrial/Organizational Psychology program where after many years he received his Doctor of Philosophy degree.
ACKNOWLEDGMENTS

For their support and guidance, I would like to first thank the members of my committee: Drs. Adam Meade, Bart Craig, Bob Pond, and Mark Wilson. I have benefited enormously over the years having had the opportunity to study and work alongside you all. Additional thanks to my advisor, Adam Meade, whose expertise and effort were invaluable in both the origin and completion of this project. A special thank you to my mother, Carolyn, and late father, Mike, whose never-ending encouragement gave me the confidence to persevere during this long journey. I am forever grateful to be their son. Lastly, I would like to thank my wife, Jenn, who has been my most steadfast partner. Only with your patience and unwavering support was this a successful endeavor. I look forward to this new chapter in our lives when only our children are in school!
# TABLE OF CONTENTS

LIST OF TABLES .................................................................................................................. ix

Applying a Metaheuristic Algorithm to a Multi-Objective Optimization Problem Within Personnel Psychology ................................................................. 1

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Cognitive Ability</td>
<td>3</td>
</tr>
<tr>
<td>Measurement of general cognitive ability</td>
<td>4</td>
</tr>
<tr>
<td>Criterion-related validity of general cognitive ability</td>
<td>5</td>
</tr>
<tr>
<td>Group differences on general cognitive ability measures</td>
<td>6</td>
</tr>
<tr>
<td>Group differences and the U.S. legal environment</td>
<td>9</td>
</tr>
<tr>
<td>The Diversity-Validity Dilemma</td>
<td>11</td>
</tr>
<tr>
<td>Current efforts to address the diversity-validity dilemma</td>
<td>12</td>
</tr>
<tr>
<td>Alternative predictor methods</td>
<td>13</td>
</tr>
<tr>
<td>Assessing full range of KSAOs</td>
<td>14</td>
</tr>
<tr>
<td>Banding</td>
<td>15</td>
</tr>
<tr>
<td>Reduce verbal ability requirements</td>
<td>16</td>
</tr>
<tr>
<td>Pareto-optimal composites</td>
<td>16</td>
</tr>
<tr>
<td>Ant Colony Optimization</td>
<td>18</td>
</tr>
<tr>
<td>ACO applied to the diversity-validity dilemma</td>
<td>19</td>
</tr>
<tr>
<td>Empirical Keying</td>
<td>22</td>
</tr>
<tr>
<td>Item-level correlational methods of empirical keying</td>
<td>24</td>
</tr>
<tr>
<td>Method</td>
<td>25</td>
</tr>
<tr>
<td>Data Simulation</td>
<td>25</td>
</tr>
<tr>
<td>Population values</td>
<td>25</td>
</tr>
</tbody>
</table>
Limitations and Future Directions................................................................. 38
Conclusion........................................................................................................... 39
REFERENCES ...................................................................................................... 41
LIST OF TABLES

Table 1  Comparison of Criterion-Related Validity and Group Differences Across Methods: Low Job Complexity Condition ................................................................. 56

Table 2  Mean of the Differences Between Methods: Unit Weighting and Ant Colony Optimization ........................................................................................................ 57

Table 3  Comparison of Criterion-Related Validity and Group Differences Across Methods: Moderate Job Complexity Condition .................................................. 58

Table 4  Comparison of Criterion-Related Validity and Group Differences Across Methods: High Job Complexity Condition ......................................................... 59

Table 5  Mean of the Differences Between Methods: Point-Biserial Correlation and Ant Colony Optimization .................................................................................. 60
Applying a Metaheuristic Algorithm to a Multi-Objective Optimization Problem

Within Personnel Psychology

Organizations across all industries strive to identify and hire those who are most likely to thrive on the job. Increasingly, organizations are also focused on creating a diverse workforce. Selection practitioners attempting to optimize both the performance and racioethnic diversity of selected applicants find themselves in a difficult position. To increase the likelihood that top applicants are identified and selected, organizations rely on a variety of pre-employment assessments. These assessments are used to measure job-related individual difference variables so that between-applicant comparisons can be made and future job performance can be estimated. The most notable predictor of any number of organizationally relevant criteria, job performance included, is undoubtedly general cognitive ability (GCA). While widely recognized as a robust predictor of performance (Hunter, 1980; Hunter, 1986; Hunter & Hunter, 1984; Pearlman, Schmidt, & Hunter, 1980; Ree & Earles, 1992; Schmidt & Hunter, 1981; Schmidt, Hunter, & Pearlman, 1981; Schmitt, Gooding, Noe, & Kirsch, 1984), the use of cognitively loaded tests for personnel selection purposes is not without controversy (Cronbach, 1975; Herrnstein & Murray, 1994; Jacoby & Glauberman, 1995; Murphy, Cronin, & Tam, 2003; Neisser et al., 1996; Wigdor & Garner, 1982). Criticisms of GCA measures stem from observed mean group-level differences across majority and minority racial groups (Bobko & Roth, 2013; Hartigan & Wigdor, 1989; Jensen, 1980; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Sackett & Wilk, 1994; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). Consequently, hiring policies that make use of a top-down selection procedure where GCA is included as a heavily-weighted predictor often result in a disproportionate selection rate favoring the majority racial group. Thus, the ability for organizations to utilize more predictive selection tools while
developing a diverse workforce has been considered an “intractable problem” (Ryan & Ployhart, 2014) with Pybrun, Ployhart, and Kravitz (2008) labeling the situation as the diversity-validity dilemma.

There have been several proposed methods to address unequal selection ratios for majority and minority group members. To date, some of the more effective strategies involve abandoning GCA tests for less valid predictors with smaller subgroup differences (Hough, Oswald, & Ployhart, 2001; Schmitt, Clause, & Pulakos, 1996). Given the amount of research linking GCA to organizationally relevant criteria, more work is needed to explore how cognitive predictors can be employed within personnel selection systems while minimizing racial group differences. The current research study addresses this problem, seeking to investigate the utility of applying a metaheuristic algorithm to the diversity-validity dilemma. Specifically, I examine whether ant colony optimization (ACO; Dorigo & Stützle, 2004) can be applied to a general cognitive ability measure to differentially weight items such that predictive validity is maximized while mean racial group differences are minimized.

To address this objective, I will begin by providing an overview of the cognitive ability-job performance relationship, noting observed majority-minority group differences. I will then discuss the current U.S. legal framework guiding the use of cognitive ability measures in pre-employment selection. Building upon this discussion I will introduce currently in use strategies by which organizations strive to reduce mean group-level differences between racioethnic groups. I will then present optimization algorithms and how they can be used to address known challenges within personnel selection. As part of this overview I will focus primarily on the ant colony optimization algorithm. I will then proceed to address the main objective of this study, to determine if generating differential item weights via ACO is a viable method to simultaneously
reduce racioethnic subgroup differences in cognitive ability predictors while maintaining the highest predictive capability possible under such constraints.

**General Cognitive Ability**

General cognitive ability as an individual difference variable first received attention from psychologists beginning with Spearman’s (1904) research introducing his proposed model of human intelligence. Observing that there were often strong positive correlations among unique measures designed to assess cognitive ability, Spearman began to develop the view that underlying these positive correlations was a common factor. The first factor of Spearman’s two-factor theory of intelligence was labeled $g$, representing GCA. As a dominant factor asserting its influence across all performance domains, $g$, in tandem with a secondary domain-specific factor, was hypothesized to account for this correlated, though imperfect, relationship in performance across various measures. While models other than Charles Spearman’s have been proposed to describe the structure of cognitive ability (e.g., Carroll, 1993; Thurstone, 1938), there have been several extensive investigations (e.g., McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Ree & Earles, 1991; Ree, Earles, & Teachout, 1994) illustrating that in most instances, there is a minimal amount of incremental validity to be had in predicting performance by including specific abilities above and beyond that which is explained via GCA. The connective component amongst numerous aptitudes and specific abilities is the GCA construct.

As with the predictive validity of GCA, there is also consensus as to what the construct represents. In Gottfredson (1997), fifty-two experts in human intelligence across a variety of disciplines agreed that “intelligence is a very broad general mental capacity that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience” (p. 13). It is reflective of a deep-seated
capability to “understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought” (Neisser et al., 1996, p. 77). Although intelligence and GCA are often used interchangeably, GCA is used throughout this study to emphasize that the construct of interest is a developed ability as opposed to one’s genetic potential (Schmidt, 2002). At its core, GCA as used within the industrial/organizational psychology literature (e.g., Hunter, 1986; Hunter & Schmidt, 1996) refers simply to one’s ability to learn.

**Measurement of general cognitive ability.** Over the past century, standardized assessments of GCA have become quite prevalent. As the most common method for measuring cognitive abilities, standardized assessments offer both ease of administration and scoring as well as desirable levels of reliability, resulting in their wide use across military, civilian, and educational settings (Ones, Dilchert, Viswsvaran, & Salgado, 2010). GCA can, and often is, assessed with tests of varying forms and lengths. Measures may use different formats of administration (e.g., paper-and-pencil, computer-based) and often consist of verbal, mathematical, or spatial-reasoning items. Some employ a single item type (e.g., Raven’s Progressive Matrices; Raven, Court, & Raven, 1938), while many include a variety of item types, both verbal and nonverbal (Neisser et al., 1996). As a testament to the generality of GCA, while many assessments may look quite different, they nevertheless can measure the same thing (Gottfredson, 2002). The spread of GCA test scores throughout the population approximates a normal distribution (Gottfredson, 1997). Test scores are often converted to a scale with 100 as the mean and with a standard deviation of fifteen (Neisser et al., 1996). Under this framework, much of the population, approximately 95%, fall within two standard deviations of the mean (e.g., test scores between 70-130).
Criterion-related validity of general cognitive ability. More than a century of research establishes the criterion-related validity of GCA. More so than any other individual difference variable, researchers have argued for GCA’s unique status amongst other predictor variables when comparing predictors of performance (Gottfredson 2002; Schmidt, 2002). Data collected from millions of job applicants and employees across thousands of primary validation studies and summarized within hundreds of meta-analytic reviews clearly show that no other predictor construct consistently produces validities as high as GCA (Ones & Dilchert, 2004). Meta-analytic estimates for the correlation between GCA and overall job performance range between .51 and .58 for jobs of moderate to high complexity, and between .23 to .38 for low-complexity occupations (Hunter, 1980; Hunter & Hunter, 1984; Schmidt & Hunter, 1998). These validity coefficients are moderated by job complexity due to GCA’s central role in learning. As jobs require more complex learning for successful performance, job knowledge acquisition and job mastery become increasingly dependent upon GCA (Hunter, 1986). These estimates of the GCA-performance relationship are especially notable given that correlations above .50 in the social sciences are relatively rare and are considered large (Cohen & Cohen, 1988). Moreover, Salgado, Anderson, Moscoso, Bertua, and de Fruyt (2003) have provided evidence that similar validities are present in organizations located outside of the United States.

The correlation between GCA and performance both here in the United States and abroad are especially important because they can be interpreted as measures of effect size (Chernyshenko, Stark, & Drasgow, 2011). In the context of employee selection, a correlation between a predictor variable and an outcome variable indicates the expected performance improvement when a typical applicant’s predictor score increases. For example, a correlation of .50 between a test of GCA and future job performance would mean that by selecting an applicant
who is 1 standard deviation higher on the cognitive ability test would result in .50 standard deviation improvement in job performance. This is a sizable performance gain and illustrates the utility of GCA in predicting relevant organizational criteria.

**Group differences on general cognitive ability measures.** Racioethnic group differences on standardized assessments of cognitive ability have been observed for decades (Roth et al., 2001). This contrasts with reports showing generally negligible, or nonexistent differences between the sexes (Ployhart & Holtz, 2008). When making between-group comparisons such as these, group differences are often captured with the Cohen’s $d$ statistic. Here, the $d$ statistic represents group differences in standard deviation units, allowing for meaningful comparisons of transformed raw scores into standardized values. It is calculated by dividing the difference between two groups’ means by the pooled standard deviation for the two groups. A $d$ value of 0.0 indicates that there are no group differences. Traditionally, $d$ values of .20, .50, and .80 are said to represent small, medium, and large effects, respectively (Cohen, 1977). The larger the $d$ statistic, the greater relative difference there is between groups. Greater differences make it unlikely that minority group members will be selected when scores are rank ordered and applicants are chosen starting with the highest scoring.

While differences in GCA have been examined between multiple groups (e.g., sex and age), undoubtedly the comparison that receives the most attention in research and practice are race and ethnic group differences. On average, Black-White differences of about one standard deviation are commonly reported, with Blacks scoring lower (Hartigan & Wigdor, 1989; Hunter & Hunter, 1984; Neisser et al., 1996; Ones et al., 2010; Roth et al., 2001; Sackett & Shen, 2010). Mean group-level differences ranging from $d = .48$ to .72 have also been reported between Hispanics and Whites, with Hispanics scoring lower (Schmitt et al., 1996; Roth et al., 2001).
Systematic data comparing Whites to other racial and ethnic groups within the United States are difficult to obtain and infrequently reported. Gottfredson (1997) did, however, summarize what little information there was on Asian-White differences by stating that Asians tend to score slightly higher, though there is little definitive evidence supporting the magnitude of this difference.

One point to consider when discussing racial group differences on GCA test scores is that just as with the criterion-related validity of GCA, job complexity also acts as moderator variable for racioethnic group differences. Whereas the relationship between GCA and job performance is strongest for increasingly complex jobs, observed Black-White differences in GCA are lower for highly complex jobs ($d = .63$) than moderate ($d = .72$) or low complexity jobs ($d = .86$; Roth, et al., 2001). Wilk and Sackett (1996) have speculated that the changes observed in group differences across levels of job complexity may be due to individuals with lower GCA self-selecting out of jobs that have prohibitively high training requirements or require high levels of education.

Additionally, it is important to understand that the group differences on GCA test scores reported above are of no inherent interest in and of themselves. Differences on any predictor within a selection context are only relevant in that they exhibit a relationship with some criteria of interest (e.g., job performance). Moreover, when discussing mean group-level differences on predictor variables within a personnel selection context, it is important to remember that those differences cannot be uniformly applied to the level of the individual (i.e., committing an ecological fallacy). Members of all racioethnic groups are distributed across the GCA continuum. Knowing that certain groups may center on different means of GCA tells us nothing about the standing of the individual applicant.
Though there have been many hypotheses as to why observed racioethnic group differences in GCA exist, the exact causal determinants are largely unknown (Campbell, 1996). Method of measurement, the influence of culture, coaching, applicant perceptions, and criterion issues have all been shown to contribute at least somewhat to group differences (Hough et al., 2001). Controversially, genetic-based explanations have been offered by some as the primary source of mean racioethnic group differences (e.g., Jensen, 1969; Jensen, 1972; Eysenck, 1971). Such explanations have been rendered less likely following advances in human genomic research that have failed to identify specific genes contributing to normal variation of GCA (Nisbett, Blair, Dickens, Flynn, Halpern, & Turkheimer, 2012). Further evidence relying upon racial admixture studies and racial crossing studies consisting of interracial parents or adoptees cast further doubt that the Black-White gap in GCA is primarily attributable to genetic factors (Colman, 2016). Arguments counter to the genetic account include environmental factors such as discrimination against minority group members as well as socioeconomic factors such as nutrition, family income, parental care, and educational quality (Neisser et al., 1996). Recently, Cottrell, Newman, and Roisman (2015) proposed an empirically supported model illustrating that Black-White mean group differences in GCA test scores arise from several racially disparate conditions related to maternal disadvantages and parenting factors. Today there is little controversy that genes have at least some effect on GCA, though few would argue that environmental factors are not also worthy of investigation and likely play a significant role in observed group differences.

In addition to both nature and nurture explanations of group differences, in the past it has been claimed that the GCA tests themselves have been the source of group differences via biased tests. A test can be considered biased when an individual’s observed test score does not match
their true score on the latent trait being tested (Meade & Fetzer, 2009). As a systematic error, test bias can render between-group comparisons problematic. There is, however, little support that test bias plays a significant role in observed racial group differences in GCA test scores (Campbell, 1996; Gottfredson, 1997; Jensen, 1980; Neisser et al., 1996; Reynolds & Brown, 1984; Wigdor & Garner, 1982). The consensus today is that psychometrically sophisticated assessments of GCA used within both educational and employment settings do not result in differential item functioning across racial groups (Bartlett, Bobko, Mosier, & Hannan, 1978; Hartigan & Wigdor, 1989; Linn, 1973; Young, 2001) and that regardless of group membership, individuals exhibiting similar standing on the latent trait of interest, in this case GCA, have the same probability of providing observed responses to test items. Moreover, research examining differential prediction, the between-group equality of test-criteria relationships, has shown that there is no underprediction of minority group performance relative to their majority group counterparts (Cole, 1981; Jensen, 1980; Sackett, Borneman, & Connelly, 2008). Tests of GCA are generally valid for predicting a wide variety of criteria, job performance included. While group differences in test scores should be treated seriously and investigated to rule out test bias as a source, the presence of group differences in and of itself should not be considered an indicator of bias.

**Group differences and the U. S. legal environment.** Concern with incorporating tests of GCA into employment selection systems is because of observed racioethnic group differences in cognitive ability test scores. In the United States, Title VII of the 1964 Civil Rights Act prohibits discrimination in personnel decisions based on race, color, national origin, religion, or sex. Thus, psychological tests used for personnel selection are often subject to a high level of scrutiny to ensure that discrimination against a protected group is not present. Discrimination
within the context of employment opportunities is typically posited as taking two separate forms. Disparate treatment occurs when evidence is presented showing clear intent to discriminate via an employment decision. In contrast, adverse impact does not require intentional discrimination. First accepted by the Supreme Court in Griggs v. Duke Power (1971), adverse impact occurs when there is a “substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group.” (Equal Employment Opportunity Commission [EEOC], 1978, Section 60-3.16). As discrimination need not be purposeful for adverse impact to occur, the selection rates for majority and minority groups become a key indicator of whether employment opportunities of racioethnic minorities and women are being disproportionately limited. The rule of thumb used to determine if the selection rates are evidence of adverse impact is known as the four-fifths rule (EEOC, 1978, Section 60-3.4). The four-fifths rule is applied by comparing the subgroup with the highest selection ratio (the majority group, often Whites) to subgroups with lower selection ratios (the minority group). Adverse impact is present if the selection ratios for the minority subgroup are found to be less than four-fifths, or 80%, of the majority subgroup. For example, if a pre-employment test of GCA resulted in the hiring of 70% of White job candidates and 50% of Black job candidates, the adverse impact ratio would be .71, indicating the presence of adverse impact. Given that many valid tests of GCA exhibit large effect size differences by race, majority and minority racial applicant groups will often observe different selection rates.

The situation described above illustrating the presence of adverse impact using a single predictor is unlikely to resemble the actual hiring practices within contemporary organizations in the United States. The extent to which groups differ in regard to the underlying trait distributions assessed, the predictor constructs in the selection system and their interrelationships, the manner
in which they are combined, along with the overall selection ratio all interact to determine hiring rates for minority group members (Ones et al., 2010; Sackett & Roth, 1996). The presence or absence of adverse impact results from an employee selection system in its entirety and not an individual predictor score. With that, the practical implications of utilizing a single valid predictor exhibiting unequal score distributions on rates of adverse impact should not be overlooked.

**The Diversity-Validity Dilemma**

The term diversity-validity dilemma coined by Pyburn et al. (2008) refers to the inherent conflict that occurs when organizations strive to maximize organizational productivity via cognitively loaded selection procedures while simultaneously attempting to build a diverse workforce. These are often competing goals given that highly predictive selection tools such as tests of GCA exhibit mean group-level differences that can result in vastly different selection rates for the majority and minority group. For organizations, the potential impact of the diversity-validity dilemma begins during the hiring process but can spread to the greater workforce. Consider that when racial groups achieve different group-level means on predictor scores used for organizational entry, it may very well lead to differential rates of promotion and career advancement for each group later on. Selection procedures not only exhibit a proximal impact on who is and who isn’t hired but a more distal impact whereby society as whole is affected due to broad disparities between racial group representation across occupational level (Cottrell et al., 2015).

In addition to the social responsibility objectives that many organizations feel compelled to strive towards, there is a business case to be made for championing organizational diversity. Organizations who wish to remain competitive would likely do well to emphasize diversity not
only for its own sake but for the business impact it can have. As the workforce increasingly consists of individuals from diverse racioethnic backgrounds, organizations that are equipped to attract top talent from both majority and minority racial groups will be best positioned to achieve a competitive advantage by utilizing all of the available talent pool (Cox & Blake, 1991). Additionally, organizations that employ a diverse workforce may be better suited in addressing the needs of increasingly diverse clients (Konrad, 2003). In fact, there is evidence supporting the claim that increased gender and racioethnic diversity is related to an organization’s financial performance (Hunt, Layton, & Prince, 2015). Today’s world is becoming increasingly connected on a global scale. It is understandable that diverse organizations would be best situated to achieve higher levels of performance in such a setting.

Current efforts to address the diversity-validity dilemma. Managers and selection professionals within the United States are legally compelled to hire and encourage a diverse workforce (EEOC, 1978). Moreover, building a selection system that is both highly predictive of future applicant success while simultaneously resulting in a high rate of selection for minority group members has proven to be quite challenging. To address this current state, organizations may respond in a variety of ways based on internal value judgements. Some organizations may choose to implement a selection system that maximizes prediction at the expense of racioethnic diversity. Conversely, organizations may place a high value on a diverse workforce and be more willing to tolerate less optimal predictive capabilities to ensure that there is a more equal pass rate of majority and minority racial groups. For organizations seeking a balance between validity and diversity maximization, there are a variety of currently in use strategies that may lead to a reduction in predictor racioethnic group differences, and consequently, adverse impact, with a minimal decrement in validity. In their review of 16 such strategies, Ployhart and Holtz (2008)
found that two broad categories of methods were the most effective in addressing the diversity-validity dilemma. The first, prioritizes the use of predictors with small or nonexistent subgroup differences; thus, excluding measures of GCA. The second, combines or manipulates both cognitive and noncognitive predictor scores. Out of the 16 specific strategies examined, the authors found four to be generally effective, but none completely satisfactory. Most strategies used today by organizations offer only small or inconsequential reductions in subgroup differences accompanied by losses in validity.

The most effective strategy Ployhart and Holtz (2008) found was to use alternative predictor methods such as interviews and assessment centers in lieu of traditional multiple-choice measures. Assessing the full range of knowledge, skills, and abilities required for successful job performance by including valid predictors other than GCA were also found to be generally effective. Lastly, under specific circumstances, the use of banding and reducing verbal ability requirements were found to have some merit. All four of these strategies are accompanied by potentially significant drawbacks in implementation and real-world impact.

*Alternative predictor methods.* Replacing measures of GCA with alternative predictor methods (e.g., assessment centers) may reduce racioethnic group difference scores, but the amount of reduction can be quite variable based on what replacement predictor is used. If the reduction is minimal then the investment of both time and money to develop, administer, and score the alternative predictor will be needlessly costly (Ployhart & Holtz, 2008). Additionally, when a method is used as a predictor, it becomes impossible to isolate the variance attributable to the underlying constructs assessed via the method (Campbell & Fiske, 1959). This limits the ability of the selection professional to both justify why the alternative predictor method was included in the selection battery as well as determine the relative incremental validity of
measured constructs (Arthur & Villado, 2008). Furthermore, it is difficult to imagine a job where performance is not laden with cognitive ability to some degree. The complete exclusion of a traditional GCA measure would rarely seem justified even when encouraging diversity is a priority.

Assessing full range of KSAOs. Using alternative predictor constructs in combination with GCA has also shown some promise in addressing the diversity-validity dilemma, though there is reason to believe that the reduction in adverse impact from adding noncognitive predictors is more modest than what is reported throughout the literature (Potosky, Bobko, & Roth, 2005). Practitioners hope that by combining noncognitive predictors that exhibit small subgroup differences (e.g., personality) with GCA, a reduction in the difference between majority and minority groups for the selection battery composite score can be achieved. This is an intuitively appealing idea, though not a panacea for the diversity-validity dilemma. Consider that when regression-based weights are used, the predictor variable with the highest validity coefficient, usually GCA, exerts the greatest influence on majority-minority composite score differences (Schmitt, et al., 1997). Even in situations where uncorrelated predictors are equally valid and weighted the same in a composite, Sackett and Wilk (1994) showed that when one predictor exhibiting racioethnic group differences equal to 1 \( SD \) is combined with a second predictor exhibiting no racioethnic group differences, the composite predictor unexpectedly exhibits racioethnic group differences equal to 0.71 \( SD \). Second, one can expect diminishing returns in reduced group differences as more and more predictors are added to the composite (Sackett, Schmitt, Ellingson, & Kabin, 2001). It is also entirely possible that a composite predictor made up of two moderately correlated variables may increase observed group differences by removing error from the measure of both variables (Sackett et al., 2001). Overall,
varying levels of predictor validity, predictor subgroup differences, composite weighting schemes, and predictor intercorrelations all contribute to how adverse impact will be reduced (Sackett & Roth, 1996; Sackett & Ellingson, 1997; Schmitt, et al., 1997). As with using alternative predictor methods, the magnitude of reduction in group differences is dependent upon which noncognitive predictors are incorporated into the selection battery (e.g., Ryan, Ployhart, & Friedel, 1998). Finally, just as with employing alternative predictor methods, one may risk incurring a significant time and financial burden with a limited payoff if numerous noncognitive predictors are administered on a large scale.

**Banding.** As opposed to rank-order selection, banding involves establishing a defined range of predictor scores and treating them as equivalent (Campion et al., 2001). The rationale underlying this strategy is that predictor scores are never perfectly reliable. Thus, small differences in predictor scores are not meaningfully representative of applicants true standing on the construct of interest and should not be used to differentiate between applicants within the established band. Banding addresses the diversity-validity dilemma because bands can be created such that lower-scoring minority group members are included. Selection within the band can then be done based on criteria or factors that exhibit no racioethnic group differences. Unfortunately, the success of banding in limiting racioethnic group differences is often dependent upon systematic preferential treatment of minority group members, a procedure that is generally deemed illegal (Barrett, Doverspike, & Arthur, 1995; Gutman & Christiansen, 1997). If selection practitioners were to incorporate banding into their selection systems without placing minority group status as a selection criteria within bands, they would not run afoul of the law, though they would likely see little to no reduction in group differences either (Sackett & Roth, 1991). A situation that lowers the practical utility of banding immensely.
Reduce verbal ability requirements. Under certain circumstances it may be appropriate to minimize the verbal ability requirements of predictors to reduce group differences. This is traditionally done by employing a video-based or audio-only mode of assessment to reduce reading level requirements. This addresses the diversity-validity dilemma by removing variance in predictor scores due to verbal ability and not the intended construct shown to be related to the criteria of interest. This can be an effective strategy in reducing racioethnic group differences but only when a job analysis supports the decision (i.e., the predictor test in use has inappropriately high reading and writing requirements relative to the job requirements; Sackett et al., 2001).

Much like using alternative predictor methods and assessing the full range of KSAOs, developing and scoring predictors that are less reliant upon verbal ability can be quite expensive.

Pareto-optimal composites. One approach that has received increased attention since Ployhart and Holtz (2008) completed their review of strategies for reducing racioethnic group differences in selection is the concept of Pareto-optimality applied to the formation of predictor composites. Pareto-based weighting schemes are a form of multi-objective optimization where one outcome cannot be improved upon without negatively impacting an additional outcome (De Corte, Lievens, & Sackett, 2007; 2008). An example of a Pareto-optimal tradeoff within the context of the diversity-validity dilemma would be a specific predictor weighting scheme that achieves a certain validity coefficient, but does so with the smallest racioethnic group differences possible relative to all other set of weights resulting in the same validity coefficient (De Corte et al., 2008). Conversely, one may start out with an acceptable level of racioethnic group differences and seek to identify the composite weighting scheme that maximizes criterion-related validity relative to all other weighting schemes for that specific group difference value. There
can be several predictor composites that balance validity and adverse impact at different levels of both, with the ideal Pareto-optimal solution determined by what an organization deems desirable.

Pareto-optimality as described by De Corte and colleagues (2007) is a two-step procedure. In the beginning stage, a set of predictor weights are identified that maximizes the correlation between the predictor composite score and the criteria of interest. Next, a second set of predictor weights is identified that maximizes the correlation between the predictor composite score and the maximum possible adverse impact ratio. Once these two weighting schemes are generated, representing both ends of the Pareto-optimal tradeoff, a second stage determines other unique predictor weighting schemes that correspond to new Pareto-optimal tradeoffs falling between the originally generated maximally obtainable validity and diversity objectives.

To date, while Pareto-based weighting schemes have advanced our understanding of racioethnic diversity and validity within selection contexts (e.g., De Corte, Sackett, & Lievens, 2010; 2011; Druart & De Dorte, 2012), optimization methods resulting in Pareto-optimal composites have only been applied to selection batteries. This has resulted in a situation where the sole focus has been on how best to combine multiple predictor measures using only test-level weights. Within a single predictor such as GCA, item-level optimized weighting schemes have been left unexplored. This is unfortunate given that selection batteries made up of multiple tests are often more financially costly and time intensive than single measure predictors. Additionally, no study investigating how best to combine multiple predictors has yet to identify a clear solution to the diversity-validity dilemma (Potosky, Bobko, & Roth, 2008). Examining weighting schemes at a more granular level relative to pareto-optimal composites, especially with measures highly predictive of relevant criteria are needed.
There have been many personnel selection strategies designed to minimize racioethnic group differences. Their effectiveness measured in terms of validity, minority group selection ratio, and costs have resulted in a situation where further research is warranted (Ployhart & Holtz, 2008). More work is needed to identify different strategies that increase minority representation within organizations but do so with an analytically derived minimal loss of predictive capability. Methods typically employed outside of psychological science may be suited to address the diversity-validity dilemma. Below is a discussion of one such method.

**Ant Colony Optimization**

The main difficulty underlying the diversity-validity dilemma is one of optimization. Optimization problems are centered on the maximization or minimization of an objective, given certain constraints (De Corte et al., 2007). The diversity-validity dilemma can be thought of as a multi-objective optimization problem with two objectives; to maximize predictive validity while minimizing racioethnic group differences. Algorithmic solutions to address optimization problems have been developed and applied within both industry and a variety of academic disciplines, though are less common within psychological research. There are two categories of algorithms applied to optimization problems, exact and heuristic algorithms (Prakasam & Savarimuthu, 2016). Exact algorithms are guaranteed to identify the best solution to a problem whereas heuristic algorithms are approximate methods designed to provide good, but not the very best solution to problems in a reduced time frame (Blum, 2005). Due to the complexity associated with many combinatorial optimization problems like the diversity-validity dilemma, exact algorithms are impractical for application, resulting in heuristic algorithms receiving widespread attention from researchers and practitioners alike (Prakasam & Savarimuthu, 2016). One of the most studied heuristic algorithms is the ant colony optimization technique.
Ant colony optimization is an example of a metaheuristic, an algorithmic framework that can be applied to different types of optimization problems to identify near-optimal solutions in an acceptable amount of time (Osman & Laporte 1996). Like many algorithms designed to address optimization problems, the inspiration for the design of the ACO framework was derived from the problem-solving behavior of animals found in nature. Specifically, the foraging behaviors of Argentine ants. When ant colonies locate a food source they can typically identify the shortest route between the colony and the food source. This colony-level intelligence is achieved through the collective efforts of individual ants following relatively simple strategies. As ants first leave the colony in search of food, they do so in a random fashion, pursuing different paths. When ants locate food, they evaluate it in terms of both quality and quantity (Dorigo & Blum, 2005). During their search, the ants leave a pheromone trace along their path, the intensity of which may be based in part on the quality of food located. Shorter paths from the colony result in decreased travel times for the ants, resulting in less evaporation of the pheromone deposits and an intensification of the signal. Subsequently, the ants who come along after are more likely to follow the path with the strongest concentration of pheromone compared to other available paths (Deneubourg, Pasteels, & Verhaeghe, 1983). As more and more ants select the pheromone heavy path, a positive feedback mechanism occurs where the probability of alternative paths being selected decreases and the most efficient path is eventually settled on. It is this stochastic search procedure in combination with feedback that informs the ACO algorithm.

**ACO applied to the diversity-validity dilemma.** The ACO technique can address the multi-objective problem inherent in the diversity-validity dilemma by deriving an item weighting algorithm for a measure of GCA such that both validity and racioethnic diversity are maximized in a Pareto-optimal like fashion. To be clear, this does not mean that the application of the ACO
algorithm to predictor weighting within personnel selection will miraculously achieve the absolute maximal predictive validity while eliminating racioethnic group differences all together. Rather, a near Pareto-optimal solution can be achieved such that given a set of values (e.g., balancing validity and diversity concerns) results in a weighting scheme whereby the predictive validity coefficient cannot be meaningfully improved without negatively impacting the objective of limiting racioethnic group differences. Thus, the application of ACO to personnel selection not only allows for the use of a highly valid predictor, GCA, but does so while simultaneously ensuring that there are no significant group differences and that in doing so, the highest possible validity is achieved.

Traditionally, scale scores of multiple-choice tests of GCA are determined by evaluating each item as correct or incorrect and compiling item scores in a unit weighted fashion (i.e., all items receive equal weights). ACO methods can be used to generate differential item weights for multiple-choice tests of GCA in the following manner. First, pseudo-randomly generated weights are assigned to each item making up the GCA measure. Weight assignment is based on an underlying probability distribution. In the beginning of the ACO implementation, the probability of selecting each potential weight is equal. Over time, the probabilities change based on the performance of the item weights generated. GCA scores are computed as the product of the item and the associated weight summed across items. Performance is evaluated based on how well the newly generated item weights can predict the criteria of interest (e.g., job performance) with a predetermined constraint in place restricting possible solutions such that racioethnic group differences are minimal to nonexistent and unlikely to result in an undesirable adverse impact ratio. It may be that ACO is unable to find any combination of item weights that dramatically reduce racioethnic group differences without a substantial loss in validity. On the other hand, it
may be that the ACO results in a solution where racioethnic group differences are greatly improved with a minimal loss in validity. In any case, ACO would provide beneficial information to selection practitioners about what is and is not possible when addressing the diversity-validity dilemma via predictor weighting schemes. Thus, allowing for a more informed response when deciding how to balance organizational concerns of performance and diversity.

Currently there are several metaheuristic algorithms that could be used to address the diversity-validity dilemma. ACO was chosen for this study due to its many beneficial qualities, originally outlined by Prakash and Savarimuthu (2016). First, the biological-basis in which the algorithm identifies solutions is easily understandable and communicated. Second, among other metaheuristics such as Tabu Search, simulated annealing, and genetic algorithms, ACO tends to outperform its rivals in identifying solutions to optimization problems. Last, while many metaheuristic concepts have been proposed within various academic literatures, few provide the actual algorithms to implement in novel situations.

There are limited examples of ACO and other metaheuristic algorithms being implemented in such a way that they directly address challenges within industrial/organizational psychology. While ACO has been used to inform the development of shorter psychological scales (e.g., Leite, Huang, & Marcoulides, 2008; Olaru, Witthöft, & Wilhelm, 2015) and model specification searches within structural equation modeling (e.g., Marcoulides & Drezner, 2003), further efforts exploring other applications of ACO are warranted. As such, the purpose of the current study is to investigate whether ACO can be used to generate differential item weights for a multiple-choice measure of GCA such that items are weighted to maximize predictive validity while ensuring minimal racioethnic group differences on the predictor. If successful, ACO could
be considered a complimentary approach to other strategies employed by selection practitioners
to manage the diversity-validity dilemma.

The principal inquiry that needs to be explored is whether the ACO algorithm has the
capability to generate optimal item weights for a multiple-choice measure of cognitive ability to
maximize both validity and diversity under a variety of realistic conditions, including variations
in job complexity. As such, the following research questions are proposed:

*Research Question 1*: How does the ACO generated scoring algorithm compare to a
traditional unit weighted scoring strategy regarding criterion-related validity in environments
representative of (a) low job complexity (b) moderate job complexity and (c) high job
complexity?

*Research Question 2*: How does the ACO generated scoring algorithm compare to a
traditional unit weighted scoring strategy regarding racioethnic group differences in
environments representative of (a) low job complexity (b) moderate job complexity and (c) high
job complexity?

**Empirical Keying**

While employing ACO to generate scoring algorithms is a modern strategy, efforts to
empirically derive predictor weights considering predictor-criterion relationships are not new.
One such strategy is empirical keying, a class of methods whereby weights are derived at either
the item or response option-level to maximize predictive validity (Stokes & Searcy, 1999).
Empirical keying is most often associated with scoring biographical data (Hogan, 1994), though
there are no mathematical barriers when applying the method to other types of predictor
measures. A common criticism levied against empirical keying is that one loses the ability to
interpret the constructs underlying the predictor scores because empirical relationships are
valued above all else (Hough & Paullin, 1994; Mael & Hirsch, 1993). This should be much less of a concern when a measure of GCA is empirically keyed. Whereas a biographical measure can consist of a hodgepodge of potentially relevant life events and experiences, a psychometrically sound measure of GCA is one where items and associated response options were rationally derived based on decades of established theory.

A second criticism of empirical keying is that the scoring keys generated are often too sample specific, limiting their generalizability to other organizations. This is a legitimate concern when employing empirical keying methods but can be addressed with appropriate developmental design. First, ensuring that one has a sufficiently large and representative sample to develop stable differential weights strengthens the confidence in the generality of the empirically derived scoring key. Here, the representativeness of the sample refers to the degree to which the sample used in the development of the empirical key is similar to the sample in which the deployment of the key is intended (Hogan, 1994). Additionally, the effectiveness of any empirically derived scoring key in generalizing across samples is dependent upon the adequacy of the criterion in capturing the performance domain of interest (Thayer, 1977). Lastly, when relying on empirical relationships to develop item weights, it is often the case that the weights generated will capitalize on idiosyncratic elements of the sample in which they were derived. The use of a cross-validation sample or an appropriate statistical strategy to estimate the shrinkage in criterion-related validity is necessary to ensure its utility (Browne, 2000; Cascio & Aguinis, 2005). Though criticisms of empirical keying have been levied, when maximizing prediction is of primary importance, empirical keying can be a worthwhile pursuit when conducted properly (Mitchell & Klimoski, 1982). While there are several empirical keying procedures of varying complexities available, there does not appear to be a single method that receives universal
recognition as a superior keying strategy under all conditions (Cucina, Caputo, Thibodeaux, & Maclane, 2012). One class of methods that have been shown to yield comparable criterion-related validities to other methods while being less cumbersome are correlational methods (Cucina, et al., 2012).

**Item-level correlational methods of empirical keying.** Correlational methods of empirical keying involve generating weights for response options based on the relationship between response alternatives and the criterion variable. The point-biserial correlation is commonly employed to calculate the correlation between a dichotomously scored item and the criterion variable score, resulting in a raw weight value. Raw weights can be used to generate total scores or they can be converted into unit weights based on either the magnitude of the correlation or the significance level.

Past research has shown that in general, the point-biserial correlation is a legitimate option for practitioners who wish to maximize prediction. Furthermore, the item-level version of the method is less susceptible to shrinkage than its option-level counterpart as there are fewer parameters estimated, providing a coarser, though effective method of empirical keying. Because the point-biserial correlation method incorporates the empirical relationship between predictor components and criteria when developing weights and resulting scoring keys, it provides a benchmark by which the ACO scoring key can be evaluated in terms of prediction. Although traditional empirical keying methods do not consider racioethnic group differences, they can act as a worthwhile reference point to the ACO scoring algorithm. As such, the following research questions are proposed:

*Research Question 3:* How does the ACO generated scoring algorithm compare to the point-biserial correlation method regarding criterion-related validity in environments
Research Question 4: How does the ACO generated scoring algorithm compare to the point-biserial correlation method regarding racioethnic group differences in environments representative of (a) low job complexity (b) moderate job complexity and (c) high job complexity?

Method

Data Simulation

Simulated data were modeled to explore the utility of applying the ACO algorithm to the diversity-validity dilemma under a variety of conditions relevant to both researchers and practitioners alike. The simulation technique relied upon both archival data and meta-analytic estimates of real-world relationships between study variables to determine population correlations among variables and mean differences among racioethnic groups. These population parameters were then used to produce simulated sample predictor and criterion scores. 100 sample replications were conducted per job complexity level resulting in a total of 300 samples each containing 1,000 respondents.

Population values. Per Roth et al.’s (2001) meta-analytic review of racioethnic group differences in GCA, the means of the theta score distributions for majority group members were set to .86, .72, .63 across low, moderate, and high job complexity environments, respectively. Means of the theta score distributions across all levels of complexity for minority group members were held constant at 0. For a given sample, the proportion of majority group members to minority group members was set to correspond to current U. S. labor statistics with Whites
and Blacks estimated to make up 78% and 12% of the workforce, respectively (U. S. Bureau of

Predictor-criterion correlations were programmed such that they corresponded to Schmidt
and Hunter’s (1998) reported criterion-related validity estimates for GCA and overall job
performance for low \( (r = .23) \), moderate \( (r = .51) \), and high \( (r = .58) \) job complexity
environments.

**Simulating predictor data.** Similar to the widely used computerized adaptive testing
version of the Armed Forces Qualification Test (AFQT) component of the Armed Services
Vocational Aptitude Battery (ASVAB; Department of Defense, 1984; Segall & Moreno, 1999),
predictor data for a dichotomously scored, 60-item measure of GCA were generated. For all
samples across all job complexity conditions, predictor response vectors were produced using
Penfield’s (2003) IRT-Lab software in the following manner. First, for a given respondent in a
simulated sample, a theta score was randomly assigned according to a normal distribution. Next,
the probability of a correct response on a given item for each respondent was obtained using the
respondent’s theta score and Birnbaum’s (1968) dichotomous three-parameter logistic model.
Item parameter estimates for the simulated 60-item measure were programmed such that they
reflected AFQT parameter estimates as described within Moreno, Wetzel, McBride, and Weiss
(1984). By summing the probabilities of each response option, the cumulative probability was
calculated. Additionally, a uniformly distributed random variate \( (u) \) was generated on the interval
between 0 and 1. With both the cumulative probability and \( u \), the response option with the lowest
value for which the probability was greater than or equal to \( u \) was identified and was selected as
the item response. This process was repeated for all 60 items resulting in a correct (1) or
incorrect (0) score for every item. Sum scores were then calculated representing the unit weighted GCA predictor variable.

Simulating criterion data. Using GCA scores, simulated criterion values representing overall job performance were generated within R (R Core Team, 2018). First, for each desired correlation across complexity conditions, a 2 x 2 matrix was constructed with 1’s along the diagonal and the other elements equal to the desired criterion-related validity coefficient. Next, a Cholesky decomposition was performed on this matrix. Then, a randomly generated vector of criterion scores was created and combined with the GCA scores to form a new matrix. When multiplied by the Cholesky decomposed matrix the GCA scores were held constant and the criterion scores were induced to achieve a predictor-criterion correlation that approximated the desired relationship. Repeated across all samples, this process resulted in a score representing an overall job performance criterion for every respondent.

Research Design

Each sample was split into a calibration sample approximating two-thirds of simulated respondents and a cross-validation sample made up of the remaining one-third. Calibration samples were used for the calculation of point-biserial correlations and the application of the ACO algorithm. Point-biserial correlations were calculated by taking the correlation between each dichotomously scored item and the criterion score. This correlation coefficient was then applied as a raw weight to that item. The ACO algorithm, written in R (R Core Team, 2018) and developed by Meade, Thompson, and Schwall (2016) was modified for the purposes of this study. To address the diversity-validity dilemma, the algorithm was constrained such that only solutions with a $d$ value of .10 or less for racioethnic group differences were produced. Within each calibration sample the ACO algorithm was run 10 independent times so that the candidate
solutions generated were of the highest quality and not stuck in local optima. The item weight solution with the highest criterion-related validity coefficient under the $d = .10$ constraint was then deployed to the cross-validation sample. Similarly, the calibration sample was used to determine point-biserial item weights that were later applied to the cross-validation sample. This process was completed for all samples so that scale scores representing GCA generated from each scoring method could be computed and direct comparisons of criterion-related validity and degree of racioethnic group differences could be made within the cross-validation sample. Cross-validation criterion-related validities and observed group differences for the ACO scoring algorithm are considered more robust than their calibration counterparts. This is because they are less dependent upon the specific sample characteristics by which the weights were originally derived from.

**Analyses.** Within each job complexity condition, criterion-related validity and Cohen’s $d$ were calculated for both the calibration and cross-validation holdouts for every simulated sample using the GCA predictor variables obtained by the unit-weighted scoring strategy, the point-biserial correlation method, and the ACO scoring algorithm. Across all samples within each level of job complexity, the mean criterion-related validity and Cohen’s $d$ for each of the three comparison methods was determined. Research questions were addressed by comparing these mean validity coefficients and group difference values.

**Results**

**Criterion-Related Validity of ACO and Unit Weighting**

Research Question 1 sought to compare the criterion-related validity of ACO generated GCA predictor scores with a traditional unit weighted scoring strategy across three levels of job complexity. Results specific to each complexity level are described below.
**Low job complexity.** The criterion-related validity and Cohen’s $d$ effect size estimates for each method deployed within the low job complexity condition are presented in Table 1. The results in Table 1 show that the ACO produced a lower cross-validation validity coefficient ($r = 0.107$) compared to the unit weighted approach ($r = 0.223$). In contrast, the ACO produced a higher calibration validity coefficient ($r = 0.276$) compared to the unit weighted approach ($r = 0.229$). As expected, the amount of shrinkage observed in the ACO approach (0.169) was greater when compared to the unit weighted approach (0.006) as the rationale behind unit weighting is that it is not sensitive to sample specific properties. Table 2 presents the mean of the differences between the ACO and unit weighted approach. The mean difference in the cross-validation validity between the two methods was $Diff_r = 0.116$. In terms of percent difference, the ACO approach resulted in a validity coefficient 52% lower than the unit weighted approach in the low job complexity cross-validation sample.

**Moderate job complexity.** The criterion-related validity and Cohen’s $d$ effect size estimates for each method deployed within the moderate job complexity condition are presented in Table 3. The results in Table 3 show that the ACO once again produced a lower cross-validation validity coefficient ($r = 0.352$) compared to the unit weighted approach ($r = 0.510$). Within the calibration sample the ACO produced a lower validity coefficient ($r = 0.425$) compared to the unit weighted approach ($r = 0.512$). The amount of shrinkage observed in the ACO approach (0.073) was again greater than what was observed with the unit weighted approach (0.002) which produced a slightly lower validity coefficient in the cross-validation sample compared to the calibration sample. The mean difference in the cross-validation validity between the two methods as reported in Table 2 was $Diff_r = 0.153$, larger than what was found in the low job complexity condition. Consequently, the ACO approach resulted in a cross-
validation validity coefficient 31% lower than the unit weighted approach in the moderate job complexity condition.

**High job complexity.** The criterion-related validity and Cohen’s $d$ effect size estimates for each method deployed within the high job complexity condition are presented in Table 4. Consistent with the other job complexity conditions, the results in Table 4 show that the ACO produced a lower cross-validation validity coefficient ($r = 0.430$) compared to the unit weighted approach ($r = 0.576$). Similar to the moderate job complexity condition, the ACO produced a lower calibration validity coefficient ($r = 0.495$) compared to the unit weighted approach ($r = 0.581$). The amount of shrinkage observed in the ACO approach (0.065) was again greater than what was observed with the unit weighted approach (0.005). The mean difference in the cross-validation validity between the two methods as reported in Table 2 was $\text{Diff}_r = 0.146$. Hence, the ACO approach resulted in a cross-validation validity coefficient 25% lower than the unit weighted approach in the high job complexity condition.

**Racioethnic Group Differences of ACO and Unit Weighting**

Research Question 2 sought to compare the degree of racioethnic group differences in GCA as calculated by the ACO algorithm versus a traditional unit weighted scoring strategy across three levels of job complexity. Results specific to each complexity level are described below.

**Low job complexity.** The results in Table 1 show that within the cross-validation sample, the ACO produced a lower Cohen’s $d$ value of 0.349 compared to the unit weighted approach ($d = 0.845$). This indicates that there was a smaller majority-minority group difference in GCA scores produced via the ACO algorithm than a traditional unit weighted scoring strategy. Within the calibration sample the ACO produced a much lower $d$ value of 0.081 compared to the
unit weighted approach \( (d = 0.842) \). Table 2 shows that the mean difference in the cross-validation \( d \) value between the two methods was \( \text{Diff}_d = 0.496 \). Thus, the ACO approach resulted in a \( d \) value 59% lower than the unit weighted approach in the low job complexity cross-validation sample.

**Moderate job complexity.** The results in Table 3 show that within the cross-validation sample the ACO once again produced a lower \( d \) value of 0.475 compared to the unit weighted approach \( (d = 0.730) \). Within the calibration sample the ACO again produced a much lower \( d \) value of 0.090 compared to the unit weighted approach \( (d = 0.710) \). Table 2 shows that the mean difference in the cross-validation \( d \) value between the two methods was \( \text{Diff}_d = 0.253 \). As a result, the ACO approach produced a \( d \) value 35% lower than the unit weighted approach in the moderate job complexity cross-validation sample.

**High job complexity.** The results in Table 4 show that within the cross-validation sample the ACO was able to achieve a lower \( d \) value of 0.442 compared to the unit weighted approach \( (d = 0.602) \). As with the previous two job complexity conditions, the ACO performed admirably in the calibration sample resulting in a lower value 0.090 compared to the unit weighted approach \( (d = 0.602) \). Table 2 shows that the mean difference in the cross-validation \( d \) value between the two methods was \( \text{Diff}_d = 0.160 \). Thus, the ACO approach resulted in a \( d \) value 27% lower than the unit weighted approach in the high job complexity cross-validation sample.

**Criterion-Related Validity of ACO and Point-Biserial Correlation**

Research Question 3 sought to compare the criterion-related validity of ACO generated GCA predictor scores with those derived from the point-biserial correlation method across three levels of job complexity. Results specific to each complexity level are described below.
Low job complexity. The results in Table 1 show that the ACO produced a lower cross-validation validity coefficient ($r = 0.107$) compared to the point-biserial method ($r = 0.223$). In contrast, the ACO produced a higher calibration validity coefficient ($r = 0.276$) compared to the point-biserial method ($r = 0.245$). The amount of shrinkage observed in the ACO approach (0.169) was greater than the point-biserial method (0.022). Table 5 presents the mean of the differences between the ACO and point-biserial method. The mean difference in the cross-validation validity between the two methods was Diff$_v$ = 0.116. Consequently, the ACO approach resulted in a validity coefficient 52% lower than the point-biserial method in the low job complexity cross-validation sample.

Moderate job complexity. The results in Table 3 show that the ACO cross-validation validity coefficient once again produced a lower validity coefficient ($r = 0.352$) compared to the point-biserial method ($r = 0.505$). Unlike the low job complexity condition, the ACO produced a lower calibration validity coefficient ($r = 0.425$) compared to the point-biserial method ($r = 0.517$). The amount of shrinkage observed in the ACO approach (0.073) was again greater than what was observed with the point-biserial method (0.012) which produced a slightly lower validity coefficient in the cross-validation sample compared to the calibration sample. The mean difference in the cross-validation validity between the two methods as reported in Table 5 was Diff$_v$ = 0.152. Thus, the ACO approach resulted in a cross-validation validity coefficient 30% lower than the point-biserial correlation approach in the moderate job complexity condition.

High job complexity. Consistent with the other job complexity conditions, the results in Table 4 show that the ACO produced a lower cross-validation validity coefficient ($r = 0.430$) compared to the point-biserial method ($r = 0.576$). As with the moderate job complexity condition, the ACO produced a lower calibration validity coefficient ($r = 0.495$) compared to the
point-biserial method \((r = 0.585)\). The amount of shrinkage observed in the ACO approach (0.065) was again greater than what was observed with the point-biserial approach (0.009). The mean difference in the cross-validation validity between the two methods as reported in Table 5 was \(\text{Diff}_r = 0.150\). As a result, the ACO approach produced a cross-validation validity coefficient 25% lower than the point-biserial method in the high job complexity condition.

**Racioethnic Group Differences of ACO and Point-Biserial Correlation**

Research Question 4 sought to compare the degree of racioethnic group differences in GCA as calculated by the ACO algorithm versus the point-biserial correlation method across three levels of job complexity. Results specific to each complexity level are described below.

**Low job complexity.** The results in Table 1 show that within the cross-validation sample, the ACO produced a lower \(d\) value of 0.349 compared to the point-biserial method \((d = 0.840)\). This indicates that there were smaller group differences in GCA scores produced via the ACO algorithm than the point-biserial method. Within the calibration sample the ACO produced a much lower \(d\) value of 0.081 compared to the point-biserial method \((d = 0.840)\). Table 5 shows that the mean difference in the cross-validation \(d\) value between the two methods was \(\text{Diff}_d = 0.491\). As such, the ACO approach resulted in a Cohen’s \(d\) value 58% lower than the point-biserial method in the low job complexity cross-validation sample.

**Moderate job complexity.** The results in Table 3 show that within the cross-validation sample the ACO once again produced a lower \(d\) value of 0.475 compared to the point-biserial method \((d = 0.726)\). Within the calibration sample the ACO again produced a much lower \(d\) value \((d = 0.090)\) compared to the point-biserial method \((d = 0.705)\). Table 5 shows that the mean difference in the cross-validation \(d\) value between the two methods was \(\text{Diff}_d = 0.251\).
Hence, the ACO approach resulted in a Cohen’s $d$ value 35% lower than the point-biserial method in the moderate job complexity cross-validation sample.

**High job complexity.** The results in Table 4 show that within the cross-validation sample the ACO was able to achieve a lower $d$ value of 0.442 compared to the point-biserial method ($d = 0.602$). As with the previous two job complexity conditions, the ACO performed admirably in the calibration sample resulting in a lower $d$ value of 0.090 compared to the point-biserial method ($d = 0.601$). Table 5 shows that the mean difference in the cross-validation $d$ value between the two methods was $\text{Diff}_d = 0.161$. As a result, the ACO approach produced a Cohen’s $d$ value 27% lower than the point-biserial method in the high job complexity cross-validation sample.

**Discussion**

Organizations looking to utilize pre-employment assessments of GCA to enhance productivity levels while promoting racioethnic diversity within the workplace can be confronted with a challenging dilemma. The current study sought to address this diversity-validity dilemma by exploring a strategy not typically utilized by psychologists and selection practitioners. By investigating the performance of the ACO algorithm in maximizing both validity and diversity under a variety of job complexity conditions, this study contributes to the personnel selection literature. Upon implementation, the ACO method performed as designed in constraining group differences to less than $d$ values of 0.10 within the calibration samples. Much more importantly, across all three levels of job complexity, the ACO approach was able to produce lower group differences within cross-validation GCA predictor scores compared to the traditional unit weighting and point-biserial correlation approach.
Although the $d$ values in ACO produced GCA scores were consistently lower than those produced via comparison methods, the magnitude of those ACO derived cross-validation group differences within moderate and high job complexity conditions were higher compared to the low complexity level. With that, all ACO cross-validation $d$ values fell below the threshold of 0.50 for medium effects. An important feat when considering that cross-validation $d$ values across complexity levels ranged from 0.60, a medium effect size, to 0.85, a large effect size, for the comparison methods.

Decreases in racioethnic group differences in GCA scores did not occur without a Pareto-optimal like tradeoff, however. As the ACO was executed under a constraint of $d = .10$, there were observed decreases in cross-validation criterion-related validity for the ACO produced predictor scores. Across all job complexity conditions, ACO criterion-related validities were lower than those produced by either traditional unit weighting or the point-biserial correlation method. Degree of differences in validity coefficients were most pronounced at the low job complexity level where cross-validation ACO criterion-related validity was 70% lower than the comparison methods versus 29% lower in the high job complexity level. By and large, the degree of reduction in racioethnic group differences was proportional to the reduction in criterion-related validity. When group membership is more highly correlated with both predictor scores and performance, ACO will have a more difficult time identifying a solution that minimizes between group differences and maximizes criterion-related validity. As such, it appears that the ACO-generated solution is more localized and generalizes less well to cross-validation samples when this is the case.

While not the focus of the current investigation, it should be noted that the point-biserial correlation method did little to improve upon the traditional unit weighted approach in terms of
criterion-related validity. Both methods consistently produced similar validities across complexity levels and samples. Although the degree of similarity between the two methods was unexpected, the findings are consistent with Wilks’ (1938) theorem. Wilks provided a mathematical proof explaining that under certain conditions, a composite variable derived from a unit weighting procedure will exhibit a correlation approaching 1.0 with composites weighted by other methods. The factors that determine the correlation between composites are the mean correlation among the predictors, the number of predictors, and the variability of the weights (Ree, Carretta, & Earles, 1998). When there is a high number of variables that are positively correlated with one another and relatively little variability among the weights, there will only be a modest effect on the order of scores (Aiken, 1966, Ree & Earles, 1991). Thus, the correlation between a unit-weighted linear composite score and composite score obtained by summing many items with differential item weights will be quite high under such circumstances. Although it is reasonable to expect that differentially weighting GCA items based on predictor validities should enhance prediction, for the data generated in the present study point-biserial empirical keying did little to improve upon unit weighting. Across all three job complexity conditions, the point-biserial correlation method resulted in positively intercorrelated item scores and item weights with low variability. Given these conditions, the point-biserial correlation method had little effect on the validity of the produced GCA predictor explaining why the performance of the unit weighted and point-biserial procedures were so similar.

**Practical Implications**

There are a number of practical applications that may be gained from the current investigation. For organizations seeking to minimize racioethnic group differences on cognitively loaded pre-employment assessments while maintaining the highest predictive
capabilities possible, this study provides estimates of what may be expected when applying ACO in terms of both validity and racioethnic group differences for typical conditions across multiple levels of job complexity. Additionally, group difference constraint parameters within the ACO algorithm can be altered such that practitioners can easily see the tradeoffs in diversity and validity depending upon their own value judgements and what is deemed most desirable.

Rarely do organizations make use of a single predictor in hiring decisions. Many organizations address the diversity-validity dilemma by combining GCA with valid noncognitive predictors that lack discernable racioethnic group differences. Schmitt et al. (1997) found that by creating composite predictors comprised of personality, structured interview data, biodata, and GCA, it was possible under certain circumstances to reduce racioethnic group differences. Unfortunately, it was also found that in many situations the observed group differences in composite predictor scores surpassed that of GCA alone. Additionally, under a range of reasonable selection ratios the reduction in effect size was not enough to remove the threat of violating the four-fifths rule. Along those lines, Ryan et al. (1998) have cautioned that the use of personality will not guarantee a reduction in GCA-related adverse impact. Their study found that group differences in GCA-personality composite scores resulted in adverse impact across a wide range of selection ratios even when personality was heavily weighted. Though appealing, the addition of alternative predictors with small group differences to cognitive assessments will not necessarily result in reduced effect sizes. The complex reality is that there are many variables to consider when attempting to reduce mean differences on pre-employment assessments. Based on the findings in this study, the application of ACO should be considered an efficient option when attempting to reduce group differences on cognitive predictors. It may also act as a point of
comparison to evaluate alternative methods and predictor composites with similar goals of effect size reduction, many of which require far more testing time on the part of the applicant.

As noted earlier, we should have more confidence in the generalizability of the cross-validation criterion-related validities and group differences produced by the ACO scoring algorithm compared to like values derived from the calibration samples. Nonetheless, the ACO was able to consistently produce both higher criterion-related validity and much lower group differences in the calibration samples compared to the cross-validation samples. If organizations had sufficiently large and generalizable samples to apply the ACO scoring algorithm to, it would be reasonable to expect that they would receive even better results than those reported here in the cross-validation samples. Large organizations employing a high number of people could conceivably produce enough predictor and criterion data to train the ACO algorithm. This would result in stable differential weights and less shrinkage between calibration and cross-validation samples as those reported in the current study.

**Limitations and Future Directions**

One primary limitation of this study was that it relied exclusively on simulated data and the parameters modeled. Namely, the racioethnic group differences on the GCA predictor, the criterion-related validity coefficients, and the moderating impact of job complexity on these values. Parameter values were informed by meta-analytic estimates. While the results of the present study are generalizable to situations with similar properties, it is likely that within real world organizations the actual criterion-related validities and group differences are not identical to those reported in the meta-analytic studies used to inform the data simulated for this research. Future research should explore the performance of the ACO when applied to predictor and criterion data where there is increased variability between items in both group differences and
validity. Under such conditions, the ACO should be capable of identifying items that exhibit comparatively higher validity with lower group differences and weight such items more heavily. Such a situation would likely result in increased performance of an ACO derived predictor composite.

An additional limitation was that only two groups were the target of the study: Whites and Blacks within the United States. This decision was based on a lack of reliable parameter values for other racioethnic groups. As the United States becomes increasingly diverse, more work will be needed so that other racioethnic groups are more fully incorporated into personnel selection research.

Along with applying the ACO algorithm to real world predictor data, future studies could also explore the impact of introducing ACO derived GCA predictors into larger selection batteries. It may be possible to provide incremental validity above and beyond the other predictors with minimal negative impact on racioethnic group differences. It would be interesting to examine how much of the variance explained by an ACO generated GCA predictor overlapped with noncognitive predictors compared to that of a traditionally scored test of GCA.

**Conclusion**

This study contributes to the body of literature focused on solving the diversity-validity dilemma. Specifically, this research informs practitioners of a novel metaheuristic strategy capable of minimizing racioethnic predictor score differences while maximizing criterion-related validity. Researchers too can benefit from this discussion as metaheuristic algorithms, along with other forms of machine learning, may likely become increasingly utilized within the study of personnel selection. Researchers who possess an understanding of both personnel psychology as
well multi-objective optimization methods have an opportunity to make material advancements in how future organizations select individuals for employment.
REFERENCES


doi:10.1177/1094428112440328


Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample</th>
<th>Complete</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r    d</td>
<td>r  d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit weighting</td>
<td>0.227</td>
<td>0.842</td>
<td>0.229</td>
<td>0.842</td>
<td>0.223</td>
</tr>
<tr>
<td>Ant Colony Optimization</td>
<td>-</td>
<td>-</td>
<td>0.276</td>
<td>0.081</td>
<td>0.107</td>
</tr>
<tr>
<td>Point-Biserial Correlation</td>
<td>-</td>
<td>-</td>
<td>0.245</td>
<td>0.840</td>
<td>0.223</td>
</tr>
</tbody>
</table>

*Note.* Reported $r$ and $d$ values are means of the validity coefficients and Cohen's $d$ effect size estimates calculated from 100 simulated samples. SD values are in parentheses.
### Mean of the Differences Between Methods: Unit Weighting and Ant Colony Optimization

<table>
<thead>
<tr>
<th>Complexity</th>
<th></th>
<th>Calibration</th>
<th></th>
<th></th>
<th></th>
<th>Cross-Validation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Diff(_r)</td>
<td></td>
<td>Diff(_d)</td>
<td></td>
<td>Diff(_r)</td>
<td></td>
<td>Diff(_d)</td>
</tr>
<tr>
<td>Low</td>
<td>-0.046 (0.03)</td>
<td>0.761 (0.12)</td>
<td>0.116 (0.05)</td>
<td>0.496 (0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>0.090 (0.04)</td>
<td>0.616 (0.13)</td>
<td>0.153 (0.06)</td>
<td>0.253 (0.16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.086 (0.04)</td>
<td>0.512 (0.11)</td>
<td>0.146 (0.06)</td>
<td>0.160 (0.13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Differences were calculated by subtracting values produced via ant colony optimization (ACO) from like values produced via unit weighting. Negative mean differences indicate that ACO means are greater than their unit weighted counterparts.
### Table 3

**Comparison of Criterion-Related Validity and Group Differences Across Methods: Moderate Job Complexity Condition**

<table>
<thead>
<tr>
<th>Method</th>
<th>Complete</th>
<th></th>
<th>Calibration</th>
<th></th>
<th>Cross-Validation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>$d$</td>
<td>$r$</td>
<td>$d$</td>
<td>$r$</td>
<td>$d$</td>
</tr>
<tr>
<td>Unit weighting</td>
<td>0.510 (0.02)</td>
<td>0.712 (0.10)</td>
<td>0.512 (0.03)</td>
<td>0.710 (0.12)</td>
<td>0.510 (0.04)</td>
<td>0.730 (0.16)</td>
</tr>
<tr>
<td>Ant Colony Optimization</td>
<td>-</td>
<td>-</td>
<td>0.425 (0.04)</td>
<td>0.090 (0.01)</td>
<td>0.352 (0.06)</td>
<td>0.475 (0.16)</td>
</tr>
<tr>
<td>Point-Biserial Correlation</td>
<td>-</td>
<td>-</td>
<td>0.517 (0.03)</td>
<td>0.705 (0.12)</td>
<td>0.505 (0.04)</td>
<td>0.726 (0.16)</td>
</tr>
</tbody>
</table>

*Note. Reported $r$ and $d$ values are means of the validity coefficients and Cohen's $d$ effect size estimates calculated from 100 simulated samples. SD values are in parentheses.*
<table>
<thead>
<tr>
<th>Method</th>
<th>Complete</th>
<th>Calibration</th>
<th>Cross-Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$d$</td>
<td>$r$</td>
</tr>
<tr>
<td>Unit weighting</td>
<td>0.580 (0.02)</td>
<td>0.601 (0.09)</td>
<td>0.581 (0.03)</td>
</tr>
<tr>
<td>Ant Colony Optimization</td>
<td>-</td>
<td>-</td>
<td>0.495 (0.04)</td>
</tr>
<tr>
<td>Point-Biserial Correlation</td>
<td>-</td>
<td>-</td>
<td>0.585 (0.02)</td>
</tr>
</tbody>
</table>

*Note.* Reported $r$ and $d$ values are means of the validity coefficients and Cohen’s $d$ effect size estimates calculated from 100 simulated samples. SD values are in parentheses.
Table 5

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Sample Complexity</th>
<th>Diff(_r)</th>
<th>Diff(_d)</th>
<th>Diff(_r)</th>
<th>Diff(_d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td></td>
<td>-0.030 (0.03)</td>
<td>0.759 (0.12)</td>
<td>0.116 (0.05)</td>
<td>0.491 (0.19)</td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td>0.093 (0.04)</td>
<td>0.615 (0.13)</td>
<td>0.152 (0.06)</td>
<td>0.251 (0.16)</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>0.090 (0.04)</td>
<td>0.511 (0.11)</td>
<td>0.150 (0.06)</td>
<td>0.161 (0.13)</td>
</tr>
</tbody>
</table>

Note. Differences were calculated by subtracting values produced via ant colony optimization (ACO) from like values produced via point-biserial correlation. Negative mean differences indicate that ACO means are greater than their point-biserial counterparts.