IMPROVING THE ADVERSE IMPACT AND VALIDITY TRADE-OFF IN PARAETO OPTIMAL COMPOSITES: A COMPARISON OF WEIGHTS DEVELOPED ON CONTEXTUAL VS TASK PERFORMANCE

by

HOWIN TSANG
B.S. University of Florida, 2008

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2010

ABSTRACT

Recent research in reducing adverse impact in personnel selection has focused on the use of various weighting schemes to balance levels of adverse impact and the validity of selection processes. De Corte Lievens & Sackett (2007) suggested the use of the normal boundary intersection method to create a number of weights that optimize adverse impact and criterion validity. This study seeks to improve the efficacy of this solution by looking at specific types of performance, namely task and contextual performance. It will investigate whether a focus on contextual performance will improve the trade-off by requiring smaller losses in validity for greater gains in adverse impact.

This study utilized data from 272 applicants for exempt positions at a multinational financial institution. The two sets of Paraeto optimal composite were developed, one based on contextual performance and the other based on task performance. Results were analyzed based on levels of adverse impact and validity of weights generated using each method. Results indicate that reducing adverse impact required a greater validity trade-off for task performance than contextual performance. Application of this method would allow for greater reductions to adverse impact than the original method while retaining a validity coefficient of 95% of the maximum achieved with regression weighting. Though this method would limit practitioners to selecting based on contextual performance, the use of minimal cut-off scores on task predictors or job experience could allow employers to incorporate task measures while further reducing adverse impact.

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of advisors, friends, and family. First and foremost, I owe my deepest gratitude to Dr. William Wooten for his guidance and support throughout this entire process. His shared enthusiasm for the subject inspired me to gain a deeper understanding of all aspects this study and the field of I/O psychology. For all the countless hours reviewing drafts and providing invaluable technical and professional commentary on my writing, thank you.

It has been a great honor and pleasure working with Dr. Nancy Reed, whose insightful and detailed analysis of every point being made provoked deeper research, thought, and understanding. Her support and encouragement throughout the program and this study has truly fostered my learning and love for I/O psychology.

I am extremely grateful to Dr. Fritzsche, whose critical but constructive review of both my proposal and thesis facilitated the development of new ideas. This study may never have begun without a few ideas spawned in her lectures and words of encouragement that followed.

My sincerest gratitude goes to Kristi Pippin for her unwavering support in every other aspect of life throughout this process, and for her editorial support from an outside perspective. Finally, a special thanks goes to my sister Sylvia for her help in editing, reviewing, and driving me three hours to Orlando and back shortly after my wisdom tooth extraction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Ever since the Civil Rights Act of 1964 and the creation of the EEOC, adverse impact and various methods of mitigating adverse impact have been at the core of numerous studies in Industrial Organizational Psychology. Many in personnel selection have a goal of reducing adverse impact, whether for legal or social reasons, but a dilemma arises when we must choose between measures that have the highest validity and those that result in the least amount of adverse impact. Ignoring adverse impact puts the organization at risk, opening up potential for legal action from aggrieved parties. Ignoring validity, also puts the organization at risk, potentially hiring employees who will be unable to perform effectively.

The Uniform Guidelines on Employee Selection Procedures (1978), henceforth referred to as the Uniform Guidelines, defines adverse impact as "A substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group."

The number that really matters to businesses is the adverse impact ratio. That is, the ratio of minority group members hired out of total applicants, compared to the ratio of majority group members hired out of total applicants. $\frac{Minority\ Selection\ Ratio}{Majority\ Selection\ Ratio} = Adverse\ Impact\ Ratio$ For example, let's say a given organization hired 5 out of 25 minority applicants (5/25 or 20% selection ratio) and hired 24 out of 80 majority applicants (24/80 or 30% selection ratio). The adverse impact ratio for this organization would be .667. $\frac{20\%\ (Minority)}{30\%\ (Majority)} = 66.7\%$. In this event, or any other where the adverse impact ratio falls below .80, the organization may be open to a claim of racial discrimination, and therefore subject to lengthy court action and hefty monetary judgments.

Historical Approaches

A number of solutions have been presented over the years including within group norming and banding. Banding had some potential to allow for selection based on gender/race when presented with a number of "equally qualified candidates". Within group norming showed promise with the potential of creating selection measures that were both more valid for subgroups while reducing adverse impact. Unfortunately, despite strong evidence of group differences on scores, there is no evidence of differential predictive validity between different racial/ethnic groups on major personality and cognitive ability tests (Sackett & Wilk, 1994). Basically this means norming different racial groups separately does not help predict job performance any better than a single norm without separating groups. Group norming would actually lead to less predictive validity since the norms created for each group would be based on a smaller sample size.

After being challenged in the courts in the 1980's the Department of Justice labeled race norming as a form of reverse discrimination and called for a review of the practice. In 1989, the National Academy of Sciences reviewed the General Aptitude Test Battery (GATB), a Department of Labor selection tool that was at the center of the controversy for using race norming (Ewoh & Guseh, 2001; Baydoun & Neuman, 1992). The practice was contested because percentiles were formed for each racial group with the top score for whites being treated as equal to the top score for blacks and other minorities. The GATB was found to have an overall validity of .30 based on 750 criterion validity studies (Baydoun & Neuman, 1992). Although it resulted in different scores for different racial groups, the NAS review found the GATB was not a less valid predictor of job performance for minorities than it was for whites, thus it was not racially biased (Ewoh & Guseh, 2001). The review recommended that the test

should not be used as the sole basis of selection since minorities tended to score lower than majority candidates. The case with the GATB set a precedent for the idea that criterion validity is not enough to win a discrimination challenge and that race norming shouldn't be used if a measure is equally predictive for both minority and majority groups. The practices of within-group norming and selecting solely based on race within bands in selection were ultimately prohibited in the Civil Rights Act of 1991 (Sackett & Wilk, 1994).

As various methods of score adjustment have been shot down over the years through the court of public opinion, Supreme Court decisions, and legislation, we are left with the original dilemma of choosing between validity and adverse impact. While it would be great to use a selection measure that's both highly predictive and results in low adverse impact, in most cases they either don't exist or are too costly to be feasible. Assessment Centers would be a good example of such a selection tool, with high validity and less adverse impact but they come with fairly high costs (Hunter & Hunter, 1984). Cognitive ability testing results in some of the highest predictive validity across jobs for job performance, but produces high levels of adverse impact (Hunter & Hunter, 1984; Salgadt, et al. 2003; Schmidt & Hunter, 1981; Schmidt & Hunter, 1998). Other measures like the personality measure of conscientiousness have been shown to have lower levels of adverse impact but also tend to be less predictive of job performance (Barrick & Mount, 1991; Bobko, Roth & Potosky, 1999; Schmidt & Hunter, 1998).

Rather than individually selecting measures with the highest validity, it would be most efficient to pick those with the greatest combined validity. Tests with high validity individually might appear to be useful, but if they all measure the same construct and don't have high incremental validity, then the additional test adds little to nothing to the overall validity of the selection process. Incremental validity refers to the additional variance explained that wasn't

accounted for by the original measure or construct (Brackett & Mayer, 2003). While conscientiousness alone isn't a highly predictive measure of job performance, it has greater incremental validity than most other measures when combined with cognitive ability (Avis, Kudisch, & Fortunato, 2002; Bobko, Roth , Potosky, 1999; Schmidt and Hunter, 1998). This is because personality measures different constructs in predicting job performance and has little overlap with what's measured by cognitive ability tests. Schmidt and Hunter (1998) found that the combined validity of conscientiousness and mental ability tests resulted in a .14 gain in predictive validity over a measure of General Mental Ability (GMA) alone. Comparatively, composites of GMA with Bio Data or GMA with Assessment Center data only resulted in a .01 and .02 gain in validity respectively.

Just as it is important for new measures to have incremental validity, it is important to combine measures to try to maximize incremental validity. The regression weighting method is favored because it maximizes the incremental validity of combining two different measures (De Corte, Lievens & Sackett, 2008). A method proposed by De Corte, Lievens and Sackett (2007) is based on a concept similar to maximizing incremental validity, however instead of just maximizing validity, it adds the goal of reducing adverse impact.

Current Methods

Current literature presents a number of different ways for building composite scores with various goals, including maximizing validity, minimizing costs and reducing adverse impact. Some of the more common methods for combining predictors in recent research include regression weighting, unit weighting, and different ad hoc weighting methods. Regression weighting is a method of weighting scores based on the criterion validity exhibited by each selection measure. This method would "maximize the linear relationship between the predictors

and criterion," meaning it would theoretically maximize the predictive validity of the final composite score compared to other composite weights, assuming the sample is representative of the population (Bobko, Roth & Buster, 2007; Schmitt, et al., 1997). However, one of the main problems with this method is that it does not attempt to reduce adverse impact. Since cognitive ability has high predictive validity across many positions, it's generally likely to be weighted heavily and contribute to greater adverse impact.

Another more simple method of combining measures is to simply use unit weights. Composites formed using unit weighting gives each measure a weight of 1. For smaller sample sizes where a proper regression weights can't be calculated with a high degree of confidence, it may be more appropriate to use unit weighting (Bobko, Roth & Buster, 2007; Dana & Dawes, 2004). A major argument for the use of unit weights over regression weighting is that the maximized validity is only perfectly applicable to the sample. The applicability of the regression weights to the population as a whole depend on how representative the sample is of the population. In employment (and even some experimental) settings, there are concerns with range restriction and various other factors that may reduce generalizability to the population.

Unfortunately, unit weights still don't address the problem of adverse impact. The final results of a unit weighting based selection system would directly reflect the amount of adverse impact found on average among all the tests used. For example, let's say a company uses a job knowledge test, a role playing exercise, a cognitive ability measure, and an in-basket test in their selection battery. If their respective adverse impact ratios are 0.34, 0.77, 0.25, and 0.52, the final AI ratio would be 0.47. Now let's assume the job knowledge test was a very poor predictor of job performance. Although the role playing and in-basket exercises have significantly less

adverse impact, the resultant AI levels are dragged down by the other two tests regardless of the predictive validity of any of the other tests.

Doverspike, Winter, Healy and Barrett (1996) created an ad hoc weighting scheme using a simulation of a hiring situation. Using simulated testing and hiring data, they tried using unit weighting (where all tests are weighted equally), weighting individual tests at 1.0 and all others at 0, and various other combination of weights. There was no systematic method of determining these various weights; a number of different weights, including .15, .25, .35, .50, were applied to each of the three tests with the total adding up to 1.0. The benefit of such a system is that allows for the weighting of a test battery when it isn't feasible or preferable to use a regression based system. This simple solution also allows practitioners to take both adverse impact and predictive validity into consideration. Unfortunately, the lack of a systematic method to determine which weights should be used could leave the researcher or practitioner with a less than optimal weighting system for both adverse impact and predictive validity.

A new approach to weighting proposed by De Corte, Lievens and Sackett (2007) addresses many of the shortcomings of the current methods available to practitioners. The proposed solution involves using the normal boundary intersect (NBI) method to create Paraeto optimal composites which maximize AI Ratio for any given level of predictive validity. This method is similar to the ad hoc weighting method but improves upon it by creating a systematic method for determining the optimal weights for a given situation.

The results from this procedure are illustrated in Figure 1. At one end of the weighting scheme criterion validity is optimized in the same way typical regression weights normally is, and a set of weights that optimizes the reduction of adverse impact on the other end. Then the normal boundary intersect method would be used to determine the a number of points in between

finding weights that maximize adverse impact reduction at each validity level beginning with maximum validity levels at one end approaching the maximum adverse impact ratio at the AI optimized end. Additionally the normal boundary intersection method creates a uniformly distributed set of points allowing for an estimation of the Pareto surface (or the line that represents the highest AI Ratio level for any given validity level).

While this method provides a very useful tool for maximizing both validity and adverse impact, it doesn't make it easy to select along the line. The decrease in predicted performance for gains in AI ratio may be too large for some employers to be comfortable with. The slope of the line is dependent on a number of factors including the types of instruments including both predictors and criteria used as inputs. Both researchers and practitioners have devoted much attention to various predictors. However, far less attention has been placed on criterion measures and their effects on developing selection procedures, so this is where we will begin.

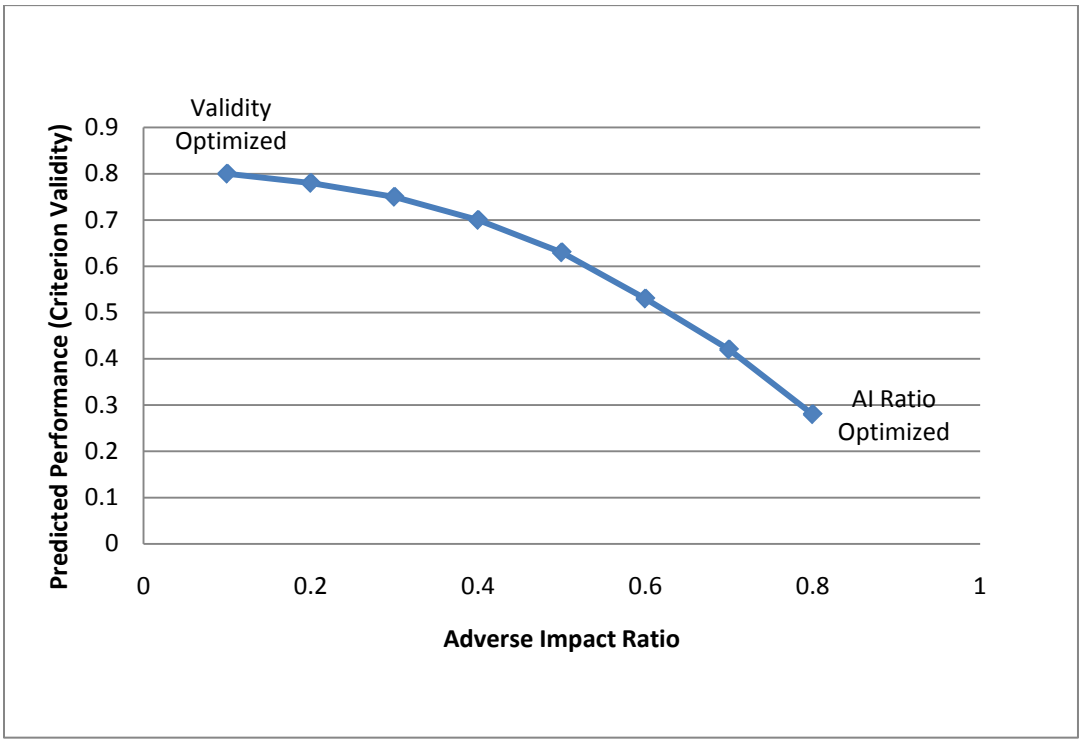

Figure 1: NBI Weighted Composites

Contextual vs Task Performance

Motowidlo and Van Scotter (1994) found that contextual and task performance should be distinguished from each other and considered different constructs. Their study showed experience to be highly correlated with task performance while personality measures were highly correlated with contextual performance. This supports the findings of McHenry, et al. (1990) which found that ability tests tend to be better at predicting general and job-related tasks, while personality measures were better at predicting "giving extra effort, supporting peers, and exhibiting personal discipline".

Task performance may be defined as "the effectiveness with which job incumbents perform activities to the organization's technical core" or how well a job incumbent performs job related tasks (Borman & Motowidlo, 1997). Tasks themselves tend to vary across different jobs and positions while contextual behaviors are typically similar across jobs. Borman and Motowidlo (1997) described contextual performance as consisting of the following:

1. Persisting with enthusiasm and extra effort as necessary to complete own task activities successfully
2. Volunteering to carry out task activities that are not formally part of own job
3. Helping and cooperating with others
4. Following organizational rules and procedures
5. Endorsing, supporting, and defending organizational objectives

Research has found focusing on contextual performance potentially is another method of reducing adverse impact. Hattrup, Rock and Scalia (1997) examined the effects of varying weights of different performance dimensions. The study found weighted composite scores that placed a greater weight on contextual performance resulted in substantially reduced levels of adverse impact.

This study will expand on the method proposed be De Corte, Lievens and Sackett (2007) by creating two sets of Paraeto optimal composites, one using contextual performance and the other using task performance data.

Hypothesis

Hypothesis: The weights developed using task performance will demonstrate a set of composite scores that illustrates visually different slopes representing different trade-offs between validity and adverse impact.

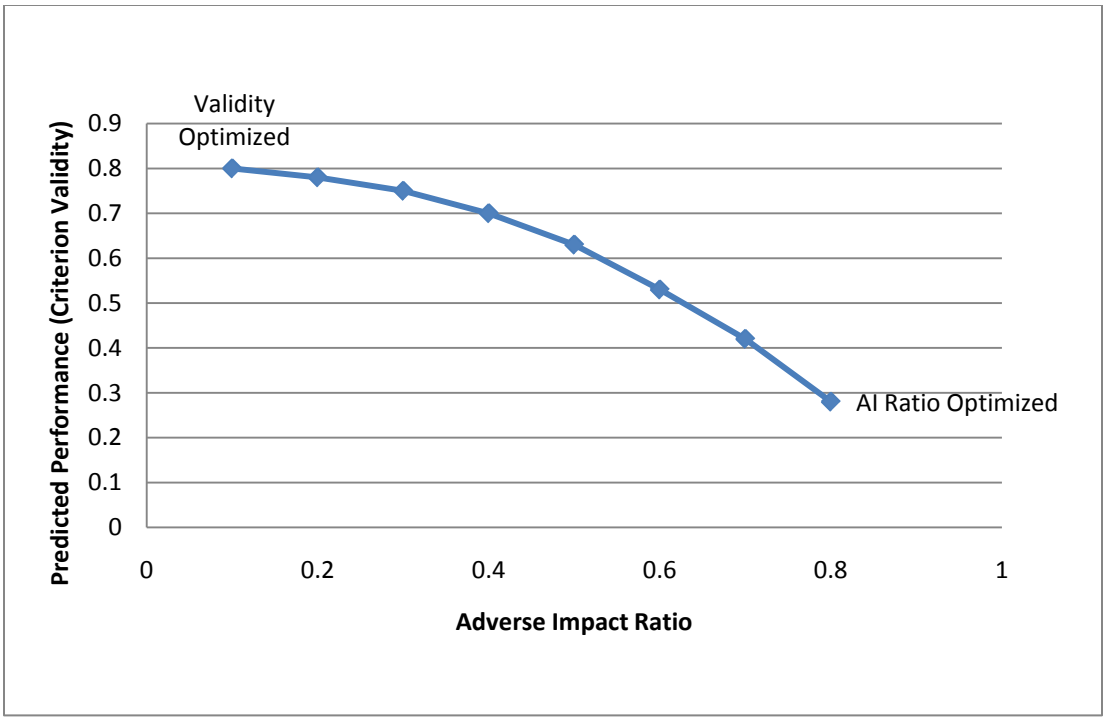

Figure 2: NBI with Contextual Performance

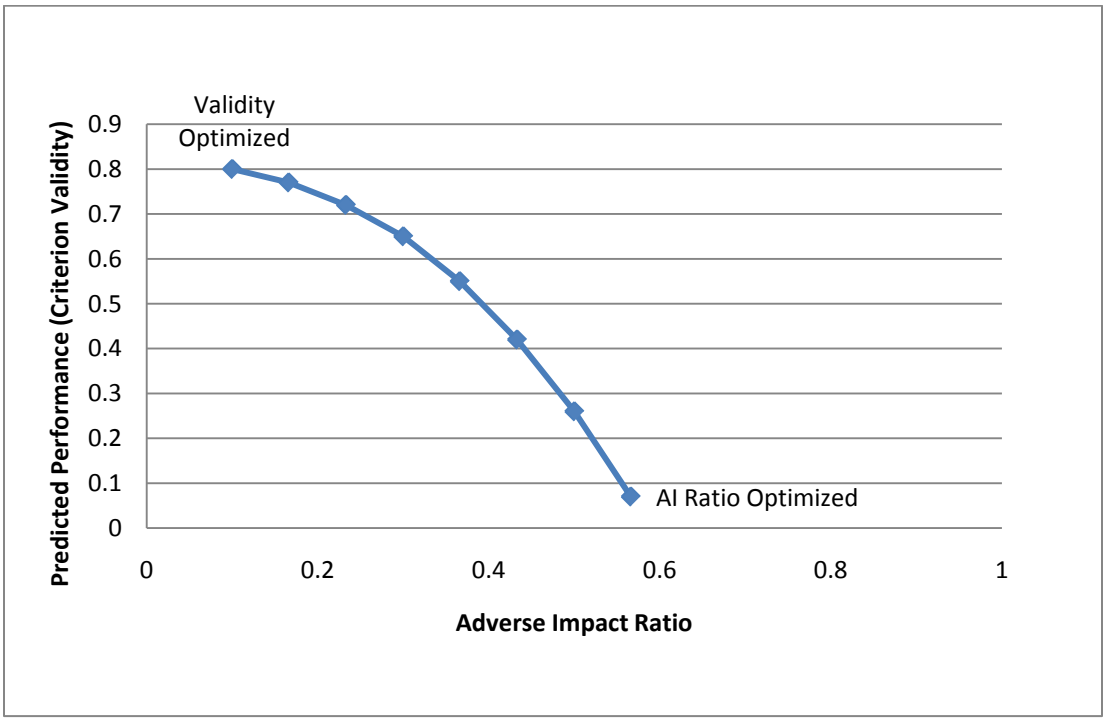

Figure 3: NBI with Task Performance

METHODOLOGY

Participants

The participants for this study represent applicants for a range of exempt positions at a large multinational financial services organization. Study participants include 272 applicants, culled from a much larger sample, all of whom took a series of employment selection tests. Participants consisted of 50% non-minority and 50% minority applicants, and 63.1% females and 35.9% males.

Predictor Measures

Each participant took a number of selection instruments but only two were selected for use in this study. The first measure is the *Teamwork and Organization Test*. This test is a measure of knowledge of cooperation and interdependence of activities in organizational settings, utilizing behavioral scenarios and representing a wide range of organizational activities in a multiple choice format. The second instrument selected for this study was the *Planning, Organizing and Scheduling Test* or POST. This is an individual assessment instrument designed to assess knowledge of planning, organizing and scheduling activities, utilizing behavioral scenarios representing areas ranging from general life experiences to work-related activities, also presented in a multiple choice format. These measures were chosen because they were administered to all participants in the sample and were consistent with the objectives of this study. The Teamwork test is expected to focus more on measuring contextual performance while the POST focuses on task performance.

Criterion Measures

Criterion performance was measured using a set of behaviorally anchored rating scales (BARS). Again, only two of the many BARS utilized by the organization were used for the study to accommodate both consistency and study objectives. The first criterion measure used was the Problem Solving/Troubleshooting rating. This rating scale focused on the ability to analyze job related problems, troubleshoot problem situations, and apply job knowledge to resolve problem issues. The second criterion was the Public & Customer Relations performance rating. This rating focused on the ability to communicate with customers, the ability to assess customer needs and satisfaction, and a willingness to address customer's problems. The first criterion represents an application of cognitive abilities, and, conceptually, is strongly related directly to task performance. The second criterion represents a human relations orientation, and, conceptually, is strongly related to contextual performance. In this sample 235 participants were measured on the Problem Solving/Troubleshooting BARS and 77 participants measured on the Public & Consumer Relations BARS.

In table 1, we can see the Public & Customer Relations BARS had a 0.041 correlation with the POST measure and 0.132 with the TEAM measure. As expected the contextual predictor is more strongly correlated with the contextual performance measure (Team) than the task performance measure (POST). The Problem Solving & Troubleshooting BARS correlated 0.145 with the POST measure and 0.103 with the Team measure. Again, we see the task performance measure correlates with the task predictor than the contextual predictor.

Table 1

Instrument Correlations

		POST	Team	Public	Problem
POST	Pearson Correlation	1			
	Sig. (2-tailed)				
	N	272			
Team	Pearson Correlation	0.080	1		
	Sig. (2-tailed)	0.188			
	N	272	272		
Public	Pearson Correlation	0.041	0.132	1	
	Sig. (2-tailed)	0.721	0.251		
	N	77	77	77	
Problem	Pearson Correlation	0.145*	0.103	0.535*	1
	Sig. (2-tailed)	0.022	0.102	0.000	
	N	251	251	56	252

* Significant at 0.05

Procedure

The criterion validity of the POST and the Teamwork measures will be calculated for both the Teamwork and Problem Solving BARS. The method outlined by De Corte, Lievens and Sackett (2007) will be calculated using the normal boundary intersect method. The normal boundary intersect will be calculated using TROFSS, a program developed by De Corte for De Corte, Lievens and Sackett (2007) and other research uses. The program uses criterion validity (of the POST and TEAM measures), number of candidates, percentage of population each group represents, effect size of the difference between groups for each measure, and selection ratio as inputs. These calculations will be performed twice, once using the validity data generated with the Public and Customer Relations BARS and the second time with the Problem Solving BARS.

The weights generated by this program will be applied to the data set to develop a set of simulated scores for each participant. A participant's score for each measure will be converted into a standard score. Group averages will be calculated separately for majority and minority

groups. Then, the number of candidates that would have been hired based on the hiring criteria will be determined. Levels of adverse impact will be calculated based on who would theoretically be hired with this weighting scheme and the validity of this process will be determined by comparing the BARS ratings of selected and rejected candidates for both the Teamwork and Problem Solving BARS.

RESULTS

A selection ratio of 0.30 was utilized for these calculations since it balances discrimination between candidates of different quality and number of candidates selected. The results are broken into two tables, with Table 2 representing outcomes using Public & Customer Relation BARS and Table 3 displaying the outcomes using the Problem Solving & Troubleshooting BARS. The tables are sorted such that AI Ratio decreases and validity increases from top to bottom. The AI Ratio column represents the percentage of minorities selected divided by the number of non-minorities selected, represented by the following equation: $\frac{Minority\ Selection\ Ratio}{Majority\ Selection\ Ratio} = Adverse\ Impact\ Ratio$. The validity column correspond to the correlation between participant's scores given a set of weights (developed by the TROFFS program) and their scores on the respective BARS.

At the AI optimized end of the Public & Customer Relations BARS in Table 2, using a weight of 0 for POST and 1 for Team results in an AI ratio of 0.700 and a validity coefficient of 0.297. At the validity optimized end with Team weighted at 0.331 and POST at 0.669, we find an AI ratio of 0.429 and a validity coefficient 0.318, as represented by option 21 on Table 2.

The line showing in Figure 4 represents the set of Paraeto optimal trade-offs between the two optimal points. The point furthest to the left represents a set of weights that would result in validity being maximized. On the far right, is the point where AI ratio is maximized. Each point in between moving from left to right shows the amount of validity sacrificed for gains in AI ratio using an optimal set of weights to maximize both at set intervals. On average the AI ratio increases while validity decreases at a rate of 0.077 for the Public & Customer Relations BARS. This means validity decreases very slowly as AI ratio increases as visualized in Figure 4.

For the Problem Solving & Troubleshooting BARS at the AI optimized end using weights of 0 for POST and 1 for Team results in an AI ratio of 0.700 and a validity coefficient of 0.320, which we can see in Table 3. At the validity optimized end using weights of .637 for POST and .363 for Team we find an AI ratio of 0.289 and a validity coefficient 0.441. The AI ratio increases while validity decreases at a rate of 0.294 for the Problem Solving & Troubleshooting BARS. Compared to the 0.077 slope for the Public & Customer Relations BARS, this results in a much steeper downward curve compared to the Public & Customer Relations BARS, due to the greater losses in validity per AI ratio gains as seen in Figure 5. This is in line with the hypothesis which stated the Problem Solving BARS would see a difference in validity for gains in AI ratio compared to the Public & Customer Relations BARS.

To compare the results of this study to typical alternatives, we look to final set of weights in both Table 2 and Table 3. The final set of weights in both tables represents the AI Ratio and Validity obtained if unit weights were used to make selection decisions on this sample, therefore both tests were weighted equally at 0.500. For this sample both AI Ratio and validity are lower using unit weights than any other available option.

A final consideration is the statistical significance of the validity in each option. For both BARS the validity of the Unit Weight option, was not found to be statistically significant at 0.05. Every other set of weights yielded validity coefficients that were significant at 0.05.

Table 2

Outcomes Using Public & Customer Relations BARS

POST	Team	AI Ratio	Validity
0.000	1.000	0.700	0.297
0.013	0.987	0.689	0.298
0.026	0.974	0.678	0.300
0.039	0.961	0.667	0.301
0.052	0.948	0.656	0.302
0.066	0.934	0.645	0.303
0.079	0.921	0.633	0.304
0.093	0.907	0.622	0.306
0.107	0.893	0.610	0.307
0.121	0.879	0.598	0.308
0.135	0.865	0.586	0.309
0.150	0.850	0.573	0.310
0.165	0.835	0.561	0.311
0.180	0.820	0.547	0.312
0.197	0.803	0.534	0.313
0.214	0.786	0.520	0.314
0.232	0.768	0.505	0.315
0.252	0.748	0.489	0.316
0.274	0.726	0.472	0.317
0.299	0.701	0.452	0.317
0.331	0.669	0.429	0.318
†0.500	0.500	0.364	*0.122

Table 3

Outcomes Using Problem Solving & Troubleshooting BARS

POST	Team	AI Ratio	Validity
0.000	1.000	0.700	0.320
0.024	0.976	0.680	0.326
0.049	0.951	0.659	0.332
0.072	0.928	0.639	0.338
0.096	0.904	0.619	0.344
0.119	0.881	0.599	0.351
0.143	0.857	0.579	0.357
0.166	0.834	0.559	0.363
0.190	0.810	0.540	0.370
0.213	0.787	0.520	0.376
0.237	0.763	0.500	0.382
0.262	0.738	0.481	0.389
0.287	0.713	0.461	0.395
0.314	0.686	0.442	0.401
0.341	0.659	0.422	0.408
0.371	0.629	0.402	0.414
0.403	0.597	0.382	0.420
0.439	0.561	0.361	0.426
0.481	0.519	0.340	0.432
0.536	0.464	0.317	0.438
0.637	0.363	0.289	0.441
† 0.500	0.500	0.240	*0.168

* Not significant at 0.05

† AI Ratio and validity calculated using unit weights.

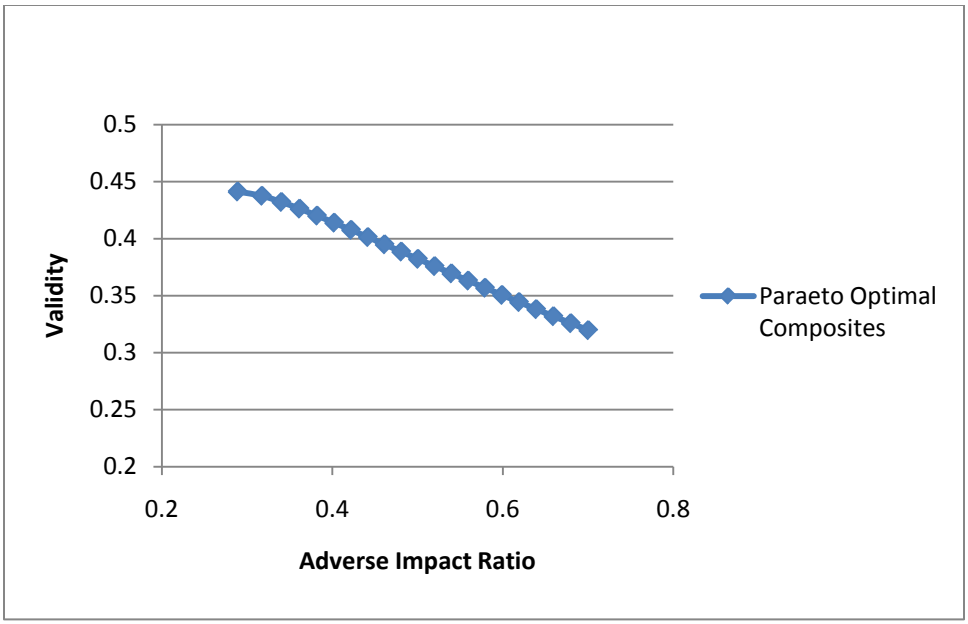

Figure 4: Problem Solving & Troubleshooting

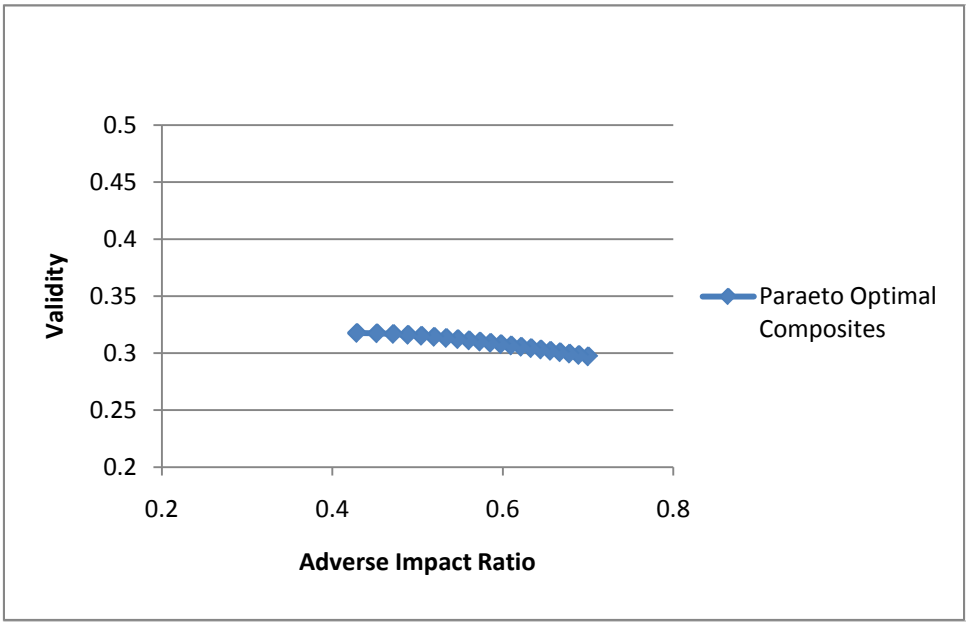

Figure 5: Public & Customer Relations

DISCUSSION

As we can see, moving towards optimizing AI Ratio has a much smaller impact on contextual performance than on task performance. For the purposes of reducing adverse impact, this method is clearly more effective than unit weighting or regression based methods. With gains in AI Ratio of up to 0.34 and 0.28 respectively over the alternatives, it has great potential to be considered part of the solution. The max AI Ratio of .70 using weighting developed based on contextual performance comes quite close, but alone is not enough to reach the 0.80 minimum selection ratio suggested by the EEOC for this sample.

For this sample, we also see all available options generated were better than their unit weighting alternative. In situations where the AI Ratio a of unit weighting scheme falls in the middle of other weighting options, practitioners could use this as a cut off for their set of options, assuming reduction of adverse impact is an objective of their selection process.

One of the primary issues for practitioners using this model is deciding how to select among the available options. If the goal is to increase the AI Ratio while maintaining something close to the maximum predictive validity of the selection process then, as De Corte, Lievens and Sackett (2007) suggests, it depends on how we define "close". In their case, it was 95%, a number which also seems appropriate for this study. For the Public & Customer Relations BARS if close is defined as being within 95% of the maximum validity level, only first four options are eliminated since 95% of 0.318 is 0.302, the fifth option from the top in Figure 2. This would result in a maximum AI Ratio of 0.66, a gain of 0.30 over unit weights. Alternatively, for the Problem Solving & Troubleshooting BARS using the 95% standard would eliminate the first sixteen options. In this case the max AI ratio while maintaining 95% of max AI Ratio would be 0.38, a gain of .24 over the unit weight alternative. By focusing on contextual

performance, this means an almost negligible loss of validity (5%) could result in substantial gains in AI Ratio (30%).

Practitioners interested in reducing adverse impact could use this method to develop alternative selection procedures. Research shows length of tenure is a solid predictor long term of task performance (McDaniel, Schmidt & Hunter, 1988). Given this, employers could consider weighting contextual performance more heavily for positions that expect longer tenures. Alternatively, if there are smaller group differences with task performance cut offs (instead of top down selection), employers could use task performance minimum cut offs while making final decisions based on contextual performance using this model.

The results of this study have important implications for human resources professionals and industrial and organizational psychologist practitioners. Unit weighting and regression weighting are both utilized as almost default options for weighting a set of selection tests. In this study, unit weighting greatly underperformed all other weighting options in both validity and adverse impact. This means that even a validity optimized set of weights would have less adverse impact than unit weighting.

This study also highlights another important point for practitioners, the real differences in results when one focuses on contextual versus task performance. A focus on contextual performance in this study provides employers with the option of greatly increasing the adverse impact ratio while maintaining close to maximum validity. On the other hand, focusing on task performance results in more of a mutually exclusive dichotomy where one must choose validity at the expense of adverse impact or vice versa.

The importance of considering contextual performance criteria can be seen in the recent Supreme Court case Ricci v. DeStefano (2009) where the court upheld a reverse discrimination

ruling by a lower court. New Haven Fire Department used a selection process composed of a written and an oral component. Without any known basis, the New Haven Fire Department decided the written component would be weighted at 60% of the total score and the oral component would be weighted at 40%. After discovering that their selection process would result in no blacks and only two Hispanics eligible for promotion the NHFD decided to throw out the test. Court ruled the city violated Title VII of the Civil Rights Act of 1964 by throwing out tests based on race.

The heavy emphasis on job knowledge (and thus task performance) and the written portion of this test left the city in a situation where almost no minority was eligible for promotion despite half their candidates being minorities. By taking contextual and other types of performance into consideration they could have gained a more complete view of each candidate's overall performance while selecting more qualified minorities for the position. In the majority opinion, Justice Kennedy rejected the idea that employers "must be in violation of the disparate-impact provision before it can use compliance as a defense in a disparate-treatment suit. " This means the court expects employers to take good faith actions to reduce adverse impact as part of the selection process rather than waiting until a situation that could potentially turn into a law suit pops up to make a race based decisions.

A possible limitation to this study is that using participants that were not all measured on both criterion measures resulted in some sample size differences. The differences in sample sizes may have been the source of some error variance. Different samples means the estimated predictive validity may have been different depending on the sample. This would translate into different validity coefficients and AI ratios that outline the Paraeto curve.

While it's possible that this may have changed the height of the curve (its overall validity) and the length of the curve (the range of AI ratios), it's unlikely to be the source of major changes to the shape of the curve. The steeper drop off in validity for the problem solving BARS can be attributed to larger group differences on the predictor with higher validity. Since this means weights with higher validity results into bigger group differences. The small drop off in validity for Public & Customer relations can be attributed to the fact that the opposite was true. The predictor with smaller group differences had a higher predictive validity for Public & Customer relations.

Not every real life situation works out as it has in this sample with these measures. However, the adverse impact and validity levels of the tests used run parallel to existing research. Tests like personality testing tend to have lower adverse impact and higher predictive validity for contextual performance and ability tests tend to have higher levels of adverse impact and higher predictive validity for task based performance (Barrick & Mount, 1991; Borman & Motowidlo, 1997; Hunter & Hunter, 1984; Hunter, 1986; McHenry, et al. 1990).

CONCLUSION

While this study highlights the benefits of using contextual performance, it is not necessarily a perfect fit for every organization across the board. Rather than considering either task or contextual performance by default employers should evaluate their employees on both to see whether the shortcomings of their employees are more a result of one or another. The problem highlighted with the New Haven Fire Department is a common one, namely that contextual performance is often wholly ignored in the selection process. By bringing contextual into consideration and applying the method developed by De Corte, Lievens & Sackett (2007), HR professional and I/O practitioners could potentially increase the overall validity of the process while reducing adverse impact.

REFERENCES

Avis, J. M., Kudisch, J. D., & Fortunato, V. J. (2002). Examining the Incremental Validity and Adverse Impact of Cognitive Ability and Conscientiousness on Job Performance. *Journal of Business and Psychology*, *17*(1), 87-105.

Barrick, M. R., & Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology*, *44*(1), 1 - 26.

Baydoun, R. B., & Neuman, G. A. (1992). The future of the General Aptitude Test Battery (GATB) for use in public and private testing . *Journal of Business and Psychology*, *7*(1), 81-91.

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.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The Usefulness of Unit Weights in Creating Composite Scores: A Literature Review, Application to Content Validity, and Meta-Analysis. *Organizational Research Methods*, *10*(4), 689-709.

Borman, W. C., & Motowidlo, S. J. (1997). Task Performance and Contextual Performance: The Meaning for Personnel Selection Research . *Human Performance*, *10*(2), 99 - 109.

Brackett, M. A., & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*(9), 1147-1158.

Dana, J., & Dawes, R. M. (2004). The Superiority of Simple Alternatives to Regression for Social Science Predictions. *Journal of Educational and Behavioral Statistics*, *29*(3), 317-331.

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact.. *Journal of Applied Psychology*, *84*(5), 695-702.

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.

Doverspike, D., Winter, J. L., Healy, M. C., & Barrett, G. V. (1996). Simulations as a Method of Illustrating the Impact of Differential Weights on Personnel Selection Outcomes . *Human Performance*, *9*(3), 259-273.

Ewoh, A. I., & Guseh, J. S. (2001). The Status of the Uniform Guidelines on Employee Selection Procedures. *Review of Public Personnel Administration*, *21*(3), 185-199.

Hattrup, K., Rock, J., & Scalia, C. (1997). The Effects of Varying Conceptualizations of Job Performance on Adverse Impact, Minority Hiring, and Predicted Performance. *Journal of Applied Psychology*, *82*(5), 656-664.

Hattrup, K., & Rock, J. (2002). A Comparison of Predictor-Based and Criterion-Based Methods for Weighing Predictors to Reduce Adverse Impact. *Applied H.R.M. Research*, *7*(1), 22-28.

Hunter, J. E., & Hunter, R. F. (1984). Validity and Utility of Alternative Predictors of Job Performance. *Psychological Bulletin*, *96*(1), 72-98.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*(3), 340-362.

Kehoe, J. F. (2008). Commentary on Pareto-Optimality as a Rationale for Adverse Impact Reduction: What would organizations do?. *International Journal of Selection and Assessment*, *16*(3), 195-200.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job Experience Correlates of Job Performance. *Journal of Applied Psychology*, *73*(2), 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.

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.

Potosky, D., Bobko, P., & Roth, P. L. (2008). Some Comments on Pareto Thinking, Test Validity, and Adverse Impact: When 'and' is optimal and 'or' is a trade-off. *International Journal of Selection and Assessment*, *16*(3), 201-205.

Ricci v. DeStefano, 129 S.Ct. 2658 (2009).

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.

Sackett, P. R., & Ellingson, J. E. (1997). The Effects of Forming Multi-Predictor Composites On Group Differences and Adverse Impact. *Personnel Psychology*, *50*(3), 707-722.

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., Corte, W. D., & 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.

Salgado, J. R., Anderson, N., Moscoso, S., Bertua, C., & Fruyt, F. D. (2003). International Validity Generalization of GMA and Cognitive Abilities: A European Community Meta-Analysis. *Personnel Psychology*, *56*(3), 573-606.

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. (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.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, *37*(3), 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.