2008

Using A Contingency-based Method For Combining Individual Assessment Center Dimension Ratings Into Overall Assessment Ratings

Keisha Wicks
University of Central Florida

USING A CONTINGENCY-BASED METHOD FOR COMBINING INDIVIDUAL ASSESSMENT CENTER DIMENSION RATINGS INTO OVERALL ASSESSMENT RATINGS

by

KEISHA KENTRELL WICKS
M.S. University of Tennessee at Chattanooga, 2004

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Science
at the University of Central Florida
Orlando, Florida

Summer Term
2008

Major Professor: Robert Pritchard

ABSTRACT

The current study applies a newly proposed mechanical combination method along with four traditional mechanical combination methods to assessment center scoring. These comparisons were made for two job levels (Fire Lieutenant and Fire Captain). The study further assesses the level of adverse impact for the various methods at three cut-off scores. Results indicated that the new contingency-based scoring method was successfully implemented in the assessment center. Results were mixed regarding whether the contingencies developed for the two job levels were different. Further, results indicated that although the various combination methods were highly correlated as expected, there were clear distinctions in the decisions made based on the different combination methods. Specifically, the various combination methods resulted in different candidates comprising the qualifying cut-off ranks. Finally, results showed that the contingency-based method had less adverse impact overall when compared to the other four methods. Future research is proposed in addition to a discussion of the limitations of the study. The main limitation was a lack of criterion data.

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES viii

CHAPTER ONE: INTRODUCTION 1

CHAPTER TWO: LITERATURE REVIEW 4

What is an Assessment Center? 4

Criterion-related Validity of ACs 6

Assessment Center Adverse Impact 7

Construct Validity of ACs 9

Scoring Methods 11

Empirical Evidence for the Combination Methods 13

Job Analysis Weighting 17

Method 1: Applying Job Analysis Weights to Components 18

Method 2: Applying Job Analysis Weights to KSAs 19

Unit Weighting 19

Method 3: Unit Weighting of the Exam Components 20

Method 4: Unit Weighting of the Exam KSAs 20

Current Study 22

Proposed Contingency Combination Method 22

Advantages of the Contingency Method 25

Comparing the Scoring Methods 29

CHAPTER THREE: METHODOLOGY 32

Participants 32
Fire Lieutenant Test Administration Procedure 34
Fire Captain Test Administration Procedure 35
Assessment 36
Contingency Development 37
Calculating OARs. 48
Determining Cut-off Ranks 48
CHAPTER FOUR: FINDINGS 52
Hypothesis 1 Results 56
Hypothesis 2 Results 64
Hypothesis 3 Results 70
Hypothesis 4 Results 73
CHAPTER FIVE: CONCLUSION 96
Discussion 96
Practical Implications 101
Limitations and Future Research 102
Conclusion 103
APPENDIX A: FIRE LIEUTENANT ASSESSOR DEMOGRAPHICS 105
APPENDIX B: FIRE CAPTAIN ASSESSOR DEMOGRAPHICS 107
APPENDIX C: CONFIDENTIALITY AGREEMENT 109
APPENDIX D: DETAILS REGARDING AC SCENARIOS AND ASSESSMENT 111

APPENDIX E: FINAL CONTINGENCIES FOR FIRE LIEUTENANT 129

APPENDIX F: FINAL CONTINGENCIES FOR FIRE CAPTAIN 132

APPENDIX G: FIRE CAPTAIN SME RATING FORM 136

APPENDIX H: IRB LETTER 138

LIST OF FIGURES

Figure 1 Example Contingency Graph 24

Figure 2 Comparison of Four Contingencies 26

Figure 3 Contingency Template with Maximum, Minimum, and Zero Point 43

Figure 4 Example of a Linear Contingency 44

Figure 5 Example of a Critical Mass 45

Figure 6 Example of a Diminishing Return 46

Figure 7 Example Contingencies with Distinct Non-linearities 63

Figure 8 Similar Contingencies for the Two Job Levels 67

Figure 9 Dissimilar Contingencies for the Two Job Levels 68

Figure 10 Fire Lieutenant Cognitively Loaded Dimensions 94

Figure 11 Fire Captain Cognitively Loaded Dimensions 95

LIST OF TABLES

Table 1 Example Summary Table of the Weights Applied in the Four Weighting Methods 21

Table 2: Fire Lieutenant Candidate Demographics 33

Table 3: Fire Captain Candidate Demographics 33

Table 4: SMEs for Contingency Development 34

Table 5: Fire Lieutenant Test Plan 35

Table 6: Fire Captain Test Plan 36

Table 7: Fire Lieutenant Reliability of Assessor Ratings for all Exam Components 53

Table 8: Fire Captain Reliability of Assessor Ratings for all Exam Components 55

Table 9 Changes in Group Composition Between the Contingency Approach and the Reversed Contingency Approach 70

Table 10 Correlation Coefficients for the Five Methods 71

Table 11 Changes in Group Composition Between Contingencies and Other Methods 72

Table 12 Descriptive Statistics and Effect Sizes 74

Table 13 Fire Lieutenant Adverse Impact Ratios at the Critical Cut-off Ranks 77

Table 14 Fire Captain Adverse Impact Ratios at the Critical Cut-off Ranks 77

Table 15 Summary of Adverse Impact Statistic Comparisons for All Ranks 79

Table 16 Fire Lieutenant Statistical Significance Tests for the Contingency Method at Rank 10 82

Table 17 Fire Lieutenant Statistical Significance Tests for the Contingency Method at Rank 39 82

Table 18 Fire Lieutenant Statistical Significance Tests for the Contingency Method at Rank 48 82

Table 19 Fire Lieutenant Statistical Significance Tests for the Job Analysis by Component Method at Rank 10 83

Table 20 Fire Lieutenant Statistical Significance Tests for the Job Analysis by Component Method at Rank 39 83

Table 21 Fire Lieutenant Statistical Significance Tests for the Job Analysis by Component Method at Rank 48 83

Table 22 Fire Lieutenant Statistical Significance Tests for the Job Analysis by KSA Method at Rank 10 84

Table 23 Fire Lieutenant Statistical Significance Tests for the Job Analysis by KSA Method at Rank 39 84

Table 24 Fire Lieutenant Statistical Significance Tests for the Job Analysis by KSA Method at Rank 48 84

Table 25 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 10 85

Table 26 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 39 85

Table 27 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 48 85

Table 28 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by Component Method at Rank 10 86

Table 29 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by Component Method at Rank 39 86

Table 30 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by Component Method at Rank 48 86

Table 31 Fire Captain Statistical Significance Tests for the Contingency Method at Rank 10 87

Table 32 Fire Captain Statistical Significance Tests for the Contingency Method at Rank 22 87

Table 33 Fire Captain Statistical Significance Tests for the Contingency Method at Rank 31 87

Table 34 Fire Captain Statistical Significance Tests for the Job Analysis by Component Method at Rank 10 88

Table 35 Fire Captain Statistical Significance Tests for the Job Analysis by Component Method at Rank 22 88

Table 36 Fire Captain Statistical Significance Tests for the Job Analysis by Component Method at Rank 31 88

Table 37 Fire Captain Statistical Significance Tests for the Job Analysis by KSA Method at Rank 10 ... 89

Table 38 Fire Captain Statistical Significance Tests for the Job Analysis by KSA Method at Rank 22 ... 89

Table 39 Fire Captain Statistical Significance Tests for the Job Analysis by KSA Method at Rank 31 ... 89

Table 40 Fire Captain Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 10 90

Table 41 Fire Captain Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 22 90

Table 42 Fire Captain Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 31 90

Table 43 Fire Captain Statistical Significance Tests for the Unit Weighting by Component Method at Rank 10 91

Table 44 Fire Captain Statistical Significance Tests for the Unit Weighting by Component Method at Rank 22 91

Table 45 Fire Captain Statistical Significance Tests for the Unit Weighting by Component Method at Rank 31 91

Table 46: Fire Captain Test Plan..... 116

CHAPTER ONE: INTRODUCTION

Assessment centers (ACs) are popular job selection tools (Connelly, Ones, Ramesh, & Goff, 2008; Lowry, 1997; Schleicher, Day, Mayes, & Riggio, 2002; Stillman & Kirkley, 2007; Woehr & Arthur, 2003) due to their perceived fairness and usefulness (e.g., Spychalski, Quinones, Gaugler, & Pohley, 1997), clearly divulged content validity (e.g., Neidig, & Neidig, 2008; Thornton & Mueller-Hanson, 2004), and their well-established predictive validity (e.g., Arthur, Day, McNelly, & Edens, 2003; Hardison & Sackett, 2004). The use of ACs as selection tools to predict performance in managerial jobs and selection for promotion has become increasingly prevalent in many organizations (Thornton, 1992).

From the time they were first introduced by the American Telegraph and Telephone Company (AT & T) years ago (Bray & Grant, 1966), ACs have managed to reach far beyond organizations in the United States spreading into Europe, South America, and Indonesia. In fact, police departments in London implemented ACs as final screening tools for the selection of police officers well before the U.S. began to implement such selection techniques (Tielsch & Whisenand, 1977). Survey reports indicated that the use of ACs in Britain is rising more quickly than alternative selection devices reaching a high of 65% of organizations using ACs (Industrial Relations Services, 1997). A study by Shackleton (1991) noted a reported increase in the use of ACs in the United Kingdom from 21% in 1986 to 59% in 1991 with another study by Boyle, Fullerton, and Yapp (1993) reporting their use in medium and large United Kingdom organizations at 45%. ACs have been increasingly used in the United States for over 40 years (Bray & Grant, 1966; Eurich, Krause, Cigularov, & Thornton, 2006). Due to the perception that selection processes can be subjective, inexact, and sometimes inequitable, ACs are desirable because they are perceived to be objective and reliable (Hinrichs & Haanpera, 1976). Despite

their popularity and impressive predictive validity coefficients, ACs have been criticized for poor construct validity and high cost (Donahue, Truxillo, Cornwell, & Gerrity, 1997; Schleicher, Day, Mayes, & Riggio, 2002; Woehr & Arthur, 2003). Dean, Roth, and Bobko (2008) conducted a recent meta-analysis showing that the group differences between Blacks and Whites in ACs may be larger than we believe them to be. This is one of the focal issues that I will address.

In the present study, I applied the concept of contingencies from the organizational productivity literature to assessment center scoring and made comparisons regarding whether or not different decisions are made when different scoring methods are applied. Contingencies are graphical representations of a type of utility function highlighting the relationship between dimensions and their contribution to the specified criterion of interest (Pritchard, 1990). They were first described by Naylor, Pritchard, and Ilgen (1980) and later applied to measuring performance in organizations (Pritchard, Jones, Roth, Stuebing, & Ekeberg, 1988). The basic study used these contingencies, to be described in more detail below, to form overall assessment ratings (OARs). The OARs formed by the contingency approach are then compared to more traditional methods of obtaining overall scores. The contingency approach is expected to improve the job relatedness of the scoring. Thus, predictive validity is expected to be higher than that obtained with more traditional assessment center scoring methods. The gold standard would be to go beyond expectations of predictive validity by actually conducting a criterion-related validity study using measures of performance on the job to provide empirical evidence regarding the predictive validity of the proposed approach. Unfortunately, this type of study was not feasible. Specifically, out of the total number of job candidates involved in the AC for this study, the number of individuals actually hired (i.e., candidates for which performance data were available) was insufficient to conduct a proper criterion-related validity study. An added benefit

of the contingency approach is that the AC may be easily customizable within a given job family by simply modifying the contingencies reflecting the identified weights, rather than by redesigning the entire AC. Ultimately, the contingency approach to scoring may provide a way to improve the job relatedness of AC scores with anticipated lower levels of adverse impact.

The next chapter will elaborate on the research evidence supporting ACs, the criticism of traditional OARs, various strategies for obtaining the OAR, and the origins of contingencies. Finally, the current study will be introduced and research questions and hypotheses will be presented.

CHAPTER TWO: LITERATURE REVIEW

What is an Assessment Center?

Assessment Centers (ACs) were originally designed to predict managerial success through the standardized measurement of various traits in multiple exercises using multiple assessors (Byham, 1980). The traits, also referred to as dimensions, are job-related individual differences constructs (Gatewood & Feild, 2001). Examples of common dimensions assessed in ACs include oral communication, leadership ability, and analytical ability. Relevant dimensions are identified from job analytic data (Gatewood & Feild, 2001; Thornton, 1992). The number of dimensions assessed during an AC range from as few as three to as many as twenty-five dimensions (Sackett & Hakel, 1979) with many ACs typically using approximately six dimensions; however, most researchers agree that the fewer dimensions you have, the better (Bycio, Alvares, & Hahn, 1987; Lance et al., 2000; Schneider & Schmitt, 1992). Gaugler and Thornton (1989) found greater observational accuracy when three dimensions rather than six or nine dimensions were assessed.

In ACs, dimensions are measured in multiple exercises. Exercises are the techniques used to elicit the behaviors to be observed during the AC. Example exercises include role-plays, in-baskets, and the structured interviews. It is important to ensure that there are multiple exercises, which are standardized, content- valid, and realistic (Gatewood & Feild, 2001; Woehr & Arthur, 2003). The purpose of having multiple exercises is to assess whether the candidate's performance on the dimensions of interest is consistent across different exercises measuring those same dimensions. For example, the role-play involves having the job candidate take on the role of the actual position he/she is applying for (e.g., Police Chief). While pretending to be in

this role, the job candidate interacts with at least one other person (e.g., a subordinate) to address a specific issue in a given scenario (e.g., citizen complaint against the employee). The candidate is observed and rated on how well he/she demonstrated the targeted dimensions for that exercise. Dimensions such as conflict management and judgment and decision making are commonly assessed during the role play exercise.

An in-basket exercise usually involves providing the job candidate with a multitude of tasks to be performed. Specifically, candidates are typically given a basket or an envelope filled with paperwork specifying several tasks that need to be addressed (e.g., responding to a citizen complaint of harassment, preparing for a neighborhood meeting to address the public on an incident that has occurred, preparing a written statement to the mayor explaining a widely publicized accusation of discrimination against the department and how that is being handled, and highlighting major issues to be addressed in a press conference to be given in one hour.) Examples of specific dimensions that might be rated in an in-basket exercise include written communication, analytical ability, and judgment and decision making.

During the administration of an AC, job candidates are observed engaging in the various exercises and rated on the dimensions being assessed in a given exercise based on the work behaviors they exhibit. It is critical to have multiple trained assessors to rate the performance of the job candidates completing the AC (Gatewood & Feild, 2001). Idiosyncrasies of the assessor can possibly lead to low correlations across exercises when only one assessor is used to rate an exercise (Sackett & Dreher, 1982; Turner & Muchinsky, 1982). Biases of any given assessor are not as influential when there are multiple assessors who have gone through extensive training (Schleicher, Day, Mayes, & Riggio, 2002). Research shows that there is a higher degree of interrater reliability when the assessors are trained (Bray & Grant, 1966; Schleicher, Day,

Mayes, & Riggio, 2002; Spychalski, Quinones, Gaugler, & Pohley, 1997). The assessors are typically subject matter experts (SMEs) who are intimately familiar with the job and requirements to be successful on the job.

When scoring the ACs it is common practice to first have assessors make independent preliminary ratings when formulating individual dimension ratings in the ACs. Subsequently, raters will commonly engage in discussion when there are discrepant preliminary ratings to reach a consensus on the final individual dimension ratings. The overall dimension ratings are then combined either statistically or judgmentally to form an overall assessment rating (OAR) reflecting the job candidate's overall AC performance across all exercises and dimensions. This OAR is used to either predict the candidate's standing on the relevant job criterion (e.g., overall job performance, promotability, training potential, etc.) or provide feedback to the candidate for developmental purposes.

Criterion-related Validity of ACs

The overall assessment ratings (OARs) that result from assessment centers have been consistently shown to have moderate to high correlations with various job criteria including military recruiter performance (Borman, 1982), promotion, overall training performance (Feltham, 1988), overall job performance (Hermelin, Lievens, & Robertson, 2007; Jansen & Stoop, 2001; Klimoski & Strickland, 1977; Ross, 1980), potential ratings (Gaugler, Rosenthal, Thornton, & Bentson, 1987), achievement (Schmitt, Gooding, Noe, & Kirsch, 1984), status change (Schmitt, Gooding, Noe, & Kirsch, 1984), wages (Schmitt, Gooding, Noe, & Kirsch, 1984), military officer training performance (Tziner & Dolan, 1982), and salary growth (Mitchel, 1975). Research suggests that ACs best predict advancement criteria for managerial and

promotional jobs (Klimoski & Strickland, 1981; Turnage & Muchinsky, 1984). The average predictive validity coefficients for ACs tend to be around .40. A large number of police ACs have been conducted in Britain (Linnane, 1985) with validity coefficients comparable to those shown in the U.S. (Feltham, 1988). In meta-analyses by Hunter and Hunter (1984) and Gaugler, Rosenthal, Thornton, and Bentson (1985) reported AC predictive validities ranged from .37 to .43, which supports the idea that ACs tend to demonstrate decent predictive validity. Another meta-analysis by Arthur, Day, McNelly, and Edens (2003) examining 34 AC articles supported the criterion-related validity of ACs at the dimension level by demonstrating validity coefficients ranging from .25 to .39 with dimensions such as consideration, awareness of others, communication, drive, influencing others, organizing and planning, and problem solving. Hermelin, Lievens, and Robertson (2007) conducted a more recent AC meta-analysis including 26 studies, which yielded a corrected correlation coefficient of .28 between the OAR and supervisory ratings of job performance.

Assessment Center Adverse Impact

Adverse (disparate) impact refers to group differences in the outcome of an employment decision, with one or more groups being negatively affected. The $4/5^{th}$ rule outlined in the Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines; U.S. Equal Employment Opportunity Commission, 1978) is commonly used to determine adverse impact. The $4/5^{th}$ rule indicates there is adverse impact when the selection ratio for the minority group is less than $4/5^{th}$ (i.e., 80%) of the selection ratio for the majority group. A limitation that has been noted regarding the $4/5^{th}$ rule is that it fails to take into account the potential impact of sampling error (Morris & Lobsenz, 2000). When dealing with a small sample size, the $4/5^{th}$ rule is likely

to yield adverse impact even when selection rates are equal in the population (Roth, Bobko, & Switzer, 2006).

ACs have been consistently shown to result in less adverse impact than do aptitude tests (e.g., Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Hoffman & Thornton, 1997; Thornton, Murphy, Everest, & Hoffman, 2000; Tyler & Bernardin, 2003). However, this is not to suggest that ACs are completely without adverse impact. Adverse impact ratios play a pivotal role in many employment discrimination lawsuits and have become a standard part of the evaluation of employee selection procedures to determine if discrimination exists (Collins & Morris, 2008). There is evidence that ACs have adverse impact, which has resulted in lawsuits. For example, a group of Alcohol, Tobacco, and Firearms (ATF) special agents filed a lawsuit based on their claim that the AC upon which promotion decisions were being made resulted in adverse impact against African Americans (*Stewart v. Rubin*, 1996). A settlement agreement was drafted, which recognized the discrimination and awarded financial compensation to the special agents able to show validity of their claims in addition to the agreement that a new selection instrument would be developed to minimize adverse impact against protected groups. Also, a class action lawsuit was filed against the Alabama Department of Transportation claiming that the AC used to select employees adversely impacted African Americans (*Reynolds v. Alabama DOT*, 1994). Based on the finding that there was a pattern of discrimination against African Americans, a consent decree was put into place in an effort to reduce and monitor adverse impact against African Americans with this department.

Similar lawsuits have been filed despite the fact that the adverse impact ratios in ACs are more impressive than those shown with aptitude tests; consequently, there clearly still remains adverse impact against protected groups when using ACs that can be further reduced (Hoffman

& Thornton, 1997; Tyler & Bernardin, 2003). Therefore, any efforts to reduce adverse impact in selection are noteworthy. An important question for the current research is whether the newly proposed contingency method of developing overall scores shows less adverse impact than traditional methods of producing overall scores.

Construct Validity of ACs

Researchers and practitioners continue to seek explanations to better understand why evidence for the construct validity of ACs is less impressive than the criterion-related validity results (Arthur, Day, & Woehr, 2008; Howard, 2008; Lance, 2008; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens, 2008; Moses, 2008; Sackett & Dreher, 1982; Schuler, 2008; Stillman & Kirkley, 2007). Specifically, research has shown over and over again that despite the intent of ACs to measure specified trait-based dimensions, the obtained dimension ratings continue to consistently reflect exercise effects (e.g., how the candidates performed on the various exercises) rather than reflecting performance on the dimensions of interest (Bowler & Woehr, 2006; Lance et al., 2000; Lievens & Klimoski, 2001; Sackett & Tuzinski, 2001). Several different reasons for the lack of construct validity related to the design of ACs have been proposed with a few common suggestions including the use of abstract rather than concrete dimensions for assessors to rate (e.g., Donahue, Truxillo, Cornwell, & Gerrity, 1997; Hennessy, Mabey & Warr, 1998), a lack of extensive training for assessors (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002), and too many dimensions for assessors to rate (e.g., Gaugler & Thornton, 1989; Schneider & Schmitt, 1992).

Several researchers have made strong arguments regarding how AC ratings are influenced by idiosyncrasies and cognitive processes of the raters such as how they select,

organize, store, and retrieve information when making judgments (Feldman, 1981; Landy & Farr, 1980). A major implication of poor assessor training is inaccuracy of ratings (Schleicher, Day, Mayes, & Riggio, 2002). Frame-of-reference (FOR) training, which was originally proposed by Bernardin and Buckley (1981), has been suggested as a technique to increase accuracy of ratings by replacing individual standards with common frames of reference for making ratings. FOR training involves highlighting the dimensions related to the job and the work behaviors associated with those dimensions, examining behaviors indicative of poor, moderate, and outstanding levels of performance on the dimension, practicing rating candidates using the new frame of reference, and finally providing feedback regarding the accuracy of the practice ratings (Pulakos, 1984). There is empirical evidence supporting the use of FOR training for increasing the accuracy of ratings by reducing halo, central tendency, and leniency (Cardy & Keefe, 1994; Schleicher & Day, 1998; Woehr & Huffcutt, 1994).

A major implication of having too many dimensions for assessors to rate is the cognitive load placed on assessors, which leads to error (Bycio, Alvares, & Hahn, 1987; Sackett & Dreher, 1982). Several researchers have suggested reducing the number of dimensions to reduce the cognitive overload (Gaugler & Thornton, 1989; Lance et al., 2000; Sackett & Hackel, 1979; Schneider & Schmitt, 1992). Despite the careful planning and extreme care taken in the development phase of ACs in addition to their unquestionable predictive validity, there still remains a grave concern regarding the ability of assessors to accurately rate candidates when attempting to integrate so many different pieces of information (Lance et al., 2000; Schneider & Schmitt, 1992).

Scoring Methods

The actual scoring method used to combine the final ratings made by the assessors is an integral part of the construct validity concern surrounding ACs (Brannick, 2008) and serves as the focal issue of this paper. Specifically, the manner in which the data are combined will have a direct impact on whether you are actually measuring the dimensions you intend to measure. For example, suppose you are measuring candidates on the dimensions of leadership, written communication, oral communication and analytical ability to predict expected promotability to the next level of the job. It has been determined that these dimensions are not equally important to the criterion of interest; however, you score the ratings collected on each dimension in a manner that does not accurately reflect the differential importance of each dimension and how that relates to the criterion of interest. If the scoring method utilized is not taking these differential weights into account when determining final ratings, the OAR will not accurately reflect what it is you are trying to measure. However, it may be true for another job that the dimensions are in fact equally important to the criterion of interest; thus, the scoring method needs to reflect the equal weighting of each dimension. Therefore, it is important to look at different scoring techniques such as weighted and non-weighted composites, including the advantages and disadvantages of each.

The most common method for combining individual dimension ratings to determine OARs continues to be the use of clinical judgments (Feltham, 1988; Thornton, 1992). The clinical judgment approach relies upon discussion of individual pieces of information by assessors to make final overall judgments. Specifically, assessors review the individual dimension ratings (e.g., leadership ability, oral communication, analytical ability) for a given assessee in each exercise (e.g., role-play, In-basket, structured interview, etc.). Assessors engage

in discussions of their interpretation of all individual dimension ratings combined for a given assessee. Finally, assessors reach a consensus regarding the OAR for the assessee based on their judgments regarding how the assessee performed on the various dimensions in the various exercises. The idea underlying this approach is that assessors can take things into account (e.g., special circumstances/situations, nonverbal cues, etc.) that an equation cannot.

An increasingly popular combination method is the mechanical approach, which is also commonly referred to as the statistical or actuarial approach (Feltham, 1988; Thornton, 1992) and was borrowed from clinical psychology. Meehl (1954a) defined this method as making predictions solely on straightforward application of an equation or table without relying on the judgments of clinicians; however, it is possible for the data to be collected judgmentally based on consensus. According to Sarbin, Taft, and Bailey (1960), the manner in which the data are collected as well as combined should be mechanical. Specifically, mathematical formulas reflecting relative weights of each dimension based on relative importance to the job criteria of interest would be used to calculate OARs. The idea behind this approach is that it would apply objectivity to an otherwise subjective process.

According to Tziner and Dolan (1982), the mechanical approach has the advantage of combining individual dimension ratings without the biases inherent in a clinical combination of ratings where scores on one dimension may unjustly influence overall ratings. Jones (1981) provided evidence showing that inter-assessor influence is a concern when relying on the clinical method of combining data and may affect final ratings. Researchers have provided consistent support for the superiority of mechanical combinations of information over the more common clinical combinations in predicting job performance (Cascio, 1987; Sackett & Wilson, 1992; Zedeck & Cascio, 1984).

In addition to its superiority in predicting job performance, a mechanical combination of ratings is more time efficient and less expensive than a consensus meeting (Feltham, 1988). Specifically, the cost of ACs can be reduced drastically by replacing the involved discussion process with a statistical approach to integrating ratings (Coulton & Field, 1995). Furthermore, the biases associated with subjective ratings (e.g., halo) and the strain placed on assessors when forced to acknowledge, analyze, and ultimately combine various pieces of individual data from various exercises can be reduced.

Empirical Evidence for the Combination Methods

Across a variety of studies (e.g., Borman, 1982; Feltham, 1988; Mitchel, 1975; Sawyer, 1966; Tziner & Dolan; 1982; Wollowick & McNamara, 1969) validity coefficients based on statistical combinations of data were significantly higher than those based on clinical combination methods. For example, in a comprehensive review of 45 studies involving a general review of clinical versus statistical prediction, Sawyer (1966) found that the mechanical approach has been consistently shown to be superior or at least equal to the clinical method in terms of predictive validity regardless of whether data were collected mechanically or clinically. Support for this conclusion was provided by Wollowick and McNamara (1969) in their finding that the common clinical approach to obtaining OARs for a group of lower and middle managers resulted in a correlation of .37 with increase in management responsibility while the mechanical approach of using multiple correlation of the AC components yielded a correlation of .62 with the same criterion. Further support for this conclusion was shown by Mitchel (1975) in a longitudinal study of three subsamples of AC participants using dimension ratings and scores on

paper-and-pencil tests. The study found that OARs determined by the clinical approach had an average predictive validity of .22 with salary growth compared to the average predictive validity of .42 for the mechanical approach; however, the mechanical approach tends to exhibit inflated validity coefficients without adjusting for shrinkage. It is worth noting that the average correlations with salary growth dropped to .28 when the regression equations were cross-validated across subsamples. However, cross-validation typically results in a conservative estimate of 'true' correlations (Humphreys, 1985). Although this may demonstrate the need for cross-validation when making comparisons of clinical versus mechanical prediction approaches, the superiority of the mechanical approach over the clinical approach still held after cross-validation. Feltham (1988) conducted a study comparing the overall predictive validity of clinically determined OARs with a unit-weighted composite over five different criteria in an AC developed to select police constables to be placed on an accelerated promotion scheme. Results indicated that the predictive validity of the unit-weighted composite was superior to the corresponding OAR coefficients for all five criteria examined thereby lending further support to the conclusion that mechanical combinations of information have greater predictive validity than clinical combinations of information. Even after correcting for differential effects of range restriction, the mechanical approach still remained superior to the clinical approach for all five criteria examined (Feltham, 1988).

Determining the best combination method may not always be a simple task due to the fact that there are disadvantages with different methods. Arthur, Doverspike, and Barrett (1996) noted that the most important concern is determining how to best weight and combine individual scores that yield overall scores with the strongest relationship with the criteria of interest. According to Arthur, Doverspike, and Barrett (1996) two basic techniques for determining test

weights include unit weighting and a job-analysis based weighting procedure. Unit weighting refers to weighting each component equally to determine the overall rating (Dawes, 1979, Feltham, 1988, Schmidt, 1971), while the second method relies on subjective judgments based on job-analysis importance weights to apply differential weights to the various components and ultimately formulate overall scores. The obvious problem with the first method is that Subject Matter Experts (SMEs) may have strong preferences regarding differential importance of certain dimensions, which may even be supported by the job analysis weights (Arthur, Doverspike, & Barrett, 1996); however, this information is disregarded in the unit weighting method. The problem with the second method is that the job analysis weights are based on subjective judgments in addition to no clear link between these judgments and final test weights (Arthur, Doverspike, & Barrett, 1996). Despite noted concerns with each method, unit weighting has been suggested as a viable alternative to standard regression methods for prediction purposes for several reasons (e.g., Einhorn & Hogarth, 1975). Unit weights do not use degrees of freedom because they are not estimated from the data. Another advantage of unit weights is that they have no standard error. The unit weighting method has other advantages over some of the other suggested weighting strategies such that they are less affected by sampling error, outliers do not impact the scores, and the weights do not change over time (Buster et al., 2003).

In light of the noted concerns regarding both basic weighting methods, Arthur, Doverspike and Barrett (1996) conducted a study examining two promotional firefighter tests and one entry-level police test using a relative content contribution (RCC) weighting method in which more weight was assigned to dimensions that were linked to work behaviors that were more important to overall job performance based on job analysis information than those dimensions linked to work behaviors deemed less important to overall job performance relative

to the other dimensions. Further, the researchers compared the RCC weighting method with two kinds of unit weighted calculations including the sum of the raw scores as well as the weighted average of the z-scores for each test. Additionally, the RCC weighting method was compared with a multiple regression method using data from a criterion-related study conducted prior to test administration. Despite the fact that a multiple regression method is not commonly seen in a content-validity study, the researchers included this method for the purpose of comparison (Arthur, Doverspike, & Barrett, 1996).

Results supported the use of the RCC weighting system by revealing they were fairly consistent across job ranks in addition to having high levels of interrater reliability. Results indicated that the various methods were highly correlated although a considerable number of individuals shifted in and out of the top 100 (the number of individuals to be selected) with the different scoring methods.

Consistent with the literature on weighting methods, there may not be significant differences in the overall ranking of candidates using the different weighting strategies; however, sizeable displacement may occur from a selection standpoint based on the shifting of individuals in and out of the eligible top ranks to be selected based on weighting method. Specifically, the final scores obtained for any given method are likely to be highly correlated with the final scores obtained using the other methods. Thus, the overall ranks that candidates receive based on their final scores will also be highly correlated when applying the different weighting strategies. However, the important issue is not the level of correlation among the final ranking of candidates. The practical concern is whether this change in final score results in the displacement of candidates from qualifying for the eligible top ranks to be selected for the position. For example, suppose the cut-off for candidates to be selected for a given position is at

rank 10. Based on one weighting strategy, the candidate has a rank of 10, but based on another weighting strategy, the candidate has a rank of 11. Even though the two ranks are highly correlated, this candidate has now been displaced from qualifying for the eligible top rank to be selected based on the weighting strategy applied. This is precisely the type of comparison that will be made in the current study.

Ultimately, the process by which various pieces of information are integrated to formulate one overall score is one of the most important steps in ACs and can have an immense impact on selection outcomes. Due to the overwhelming evidence supporting the use of a mechanical combination of data, this study will focus on such methods for combining data and make comparisons of a variety of ways in which data are currently being weighted and ultimately combined mechanically with a newly proposed mechanical combination method. Specifically, four methods will be discussed in detail before examining the proposed combination method that will be used to make comparisons with the four methods. These methods involve combining data based on two general weighting strategies (e.g., job analysis weighting and unit weighting) with two distinct approaches to applying the weights and combining the data within each general category.

Job Analysis Weighting

With this weighting strategy, the weights assigned are solely based on information provided during the job analysis process regarding the job tasks and KSAs. The job tasks are rated on criticality, which is a combination of ratings of how important each task is to successfully performing the job as well as how frequently each task is performed on the job. Specifically, SMEs (e.g., incumbents and supervisors) are asked to indicate how often they

perform a given task regardless of the importance of the task. For the importance rating, SMEs are asked how important each task is for successfully performing their job regardless of the frequency or amount of time spent on the task. Ultimately, the component rated most critical will have the highest job analysis weight (JAW). To facilitate one's understanding of the usefulness of the JAWs, they were converted to relative selection weights by dividing the JAW for each KSA by the sum of all JAWs for KSAs being assessed.

Method 1: Applying Job Analysis Weights to Components

With this weighting strategy, selection weights were first determined by dividing the JAW for each KSA by the sum of all JAWs for the KSAs being assessed. The selection weights for all KSAs associated with a specific component of the AC are summed to obtain the weight for that component in the OAR. For example, let's say a component (e.g., work sample) has five KSAs linked to it. The selection weights for those five KSAs are .069, .034, .076, .057, and .041, respectively. The selection weight applied to the work sample component would be .277 (i.e., sum of the relevant KSA selection weights) when calculating the OAR. To avoid overweighting (or underweighting) a component due to each score being calculated based on differing scales, it is necessary to transform the raw scores into standard scores to put all scores on a common metric with equally spaced intervals. This is accomplished by standardizing, or calculating *z* scores, for each dimension score.

Transforming raw scores into standard scores allows for the comparison between candidates' scores originally obtained on different rating scales. For example, it would be irrational to combine raw scores obtained using a 1-5 Likert scale and make comparisons with a combination of raw scores obtained using a 0-10 checklist because of the different scales and

units of the two scales. Therefore, it would be necessary to transform the raw scores to a common metric to allow for comparisons. Standardizing also has the effect of controlling for variability in the different scales. Scales with larger variability will have a larger effect on the composite score. Standardizing equalizes the variability and eliminates this problem. After converting the overall dimension scores to *z* scores, the next step in the scoring process is to multiply the standardized dimension scores by their corresponding selection weights and sum those scores.

Method 2: Applying Job Analysis Weights to KSAs

Before applying the job analysis weights to each dimension, the first step taken is to calculate an overall score for each of the KSAs measured by the examination components. This is accomplished by averaging the final individual KSA ratings for both assessors. For example, if a candidate received a rating of 3 on the leadership KSA from the first assessor and a rating of 4 on that same KSA from the second assessor, this candidate would receive a final leadership KSA of 3.5. Each score is standardized as described previously and summed with the other ratings for that KSA to calculate an overall score for the KSA.

Unit Weighting

This weighting strategy involves assigning equal weights to all components (i.e., implying equal importance of all KSAs measured by the various components) despite differential importance of various KSAs as determined by the job analysis.

Method 3: Unit Weighting of the Exam Components

With this weighting strategy each component is considered a unique score where each rating for a given component is combined into a single score for that component. For example, suppose an AC had 3 components (e.g., structured interview, work sample, and role-play). This would result in each candidate having 3 final scores (one overall score for each component), which are later summed to produce an OAR. All ratings received on each individual component are combined to produce the overall score for that component. Each score is standardized as described previously and summed with the other ratings for that component to calculate an overall score for the component. If different rating scales are used, the scores are first standardized prior to summing the item ratings and standardized again after summation.

Method 4: Unit Weighting of the Exam KSAs

With this weighting strategy each KSA is considered a unique score where each rating assessing a given KSA is summed into a single score that is transformed into a standard score and ultimately summed with the other KSA scores to obtain an overall score for that KSA. Each KSA will most likely not be assessed in all of the exercises; and there are likely to be differences in the number of times each KSA is assessed in the entire AC. All ratings received for a given KSA on the various exercises are combined to produce the overall score for that KSA. For example, suppose an AC had a total of 10 KSAs that were being assessed using 3 different exercises (e.g., structured interview, work sample, and role-play). This would result in each candidate having 10 final KSA scores (one overall score for each KSA), which are later summed to produce an OAR.

Table 1 provides a summary table of example weights applied using the four different weighting methods. "KSA Number" provides the actual number of the KSA/dimension being assessed. "JA by KSA" provides example weights applied to the KSAs using the job analysis weighting method. "JA by Component" provides example weights applied to each component/exercise in the AC using the job analysis weighting method. "Unit by Component" provides example weights applied to each component/exercise using the equal weighting method. "Unit by KSA" provides example weights applied to the KSAs using the unit weighting method.

Table 1 Example Summary Table of the Weights Applied in the Four Weighting Methods

KSA Number	JA by KSA	JA by Component	Unit by Component	Unit by KSA
KSA 1	0.268	0.268	0.333	0.200
KSA 3	0.459	0.458	0.333	0.600
KSA 3	0.274	0.274	0.333	0.200

The four previously discussed approaches all represent mechanical combinations of data that have been weighted using different general strategies (e.g., job analysis weights and unit weights) to obtain the final OAR. The focus of this paper is to calculate final OARs for two job levels (e.g., Fire Lieutenant and Fire Captain) using those four strategies in addition to the proposed new approach to mechanical combination and make a variety of comparisons of those results.

Current Study

The current study examined the contingency method for combining individual ratings into one overall rating with the goal a yielding a better combination method than traditionally used. Specifically, a mechanical/statistical combination method based on contingencies was compared with four other mechanical combination methods based on job analysis weights and unit weights. Because prior research indicates that a contingency approach often results in different decisions in comparison to more traditional methods of obtaining overall scores (Pritchard, Watson, Kelly, & Paquin, 1998; Pritchard & Roth, 1991), the main issue addressed in this study was to examine whether different job candidates were likely to be promoted when OARs were calculated based on contingencies compared to the other four methods. The reason for considering the new contingency approach is to provide a scoring method that is more job-related than the other four methods. Specifically, based on the ability of the contingency method to identify non-linearities and reflect those findings accordingly in the scoring process, this technique is expected to yield final scores that are more closely related to the targeted job.

Proposed Contingency Combination Method

The newly proposed mechanical combination method is based on a contingency approach. Contingencies are graphical representations of the relationship between each dimension and its contribution to the specified criterion of interest (Pritchard, 1990). Contingencies first were proposed by Naylor, Pritchard, and Ilgen (1980) and operationalized as part of the Productivity Measurement and Enhancement System (ProMES) developed by Pritchard (1990) as a motivational tool designed to measure and ultimately enhance productivity

using feedback. Contingencies are a type of graphic utility function. The horizontal axis shows the range of possible dimension ratings starting with its minimum up to its maximum level possible. For example (see Figure 1), if you are examining the dimension of effectively developing employees on a 7-point scale in a particular AC exercise, the horizontal axis for that dimension would reflect the possible range of scores a person could receive on the leadership dimension (i.e., 1 = unacceptable, 4 = acceptable, 7 = outstanding). The y-axis represents the amount of contribution made to the criterion of interest (e.g., effectiveness) ranging from -100 (highly negative effectiveness) through 0 (meeting minimum expectations) to +100 (highly positive effectiveness). The shape of the graph defines how each level of the dimension relates to the criterion of interest. For example, a score of 1 on this dimension reflects an effectiveness score of -60 whereas a score of 4 reflects an effectiveness score of 70. There is a contingency for each dimension of performance. The overall criterion score is obtained by converting the performance score on each dimension to its corresponding effectiveness score and then summing the effectiveness scores. For example, if a candidate is rated on five dimensions and these ratings correspond with contingency effectiveness scores of 60, 40, -20, -40, and 80, the overall effectiveness score for this candidate would be the sum of these numbers, which is 120.

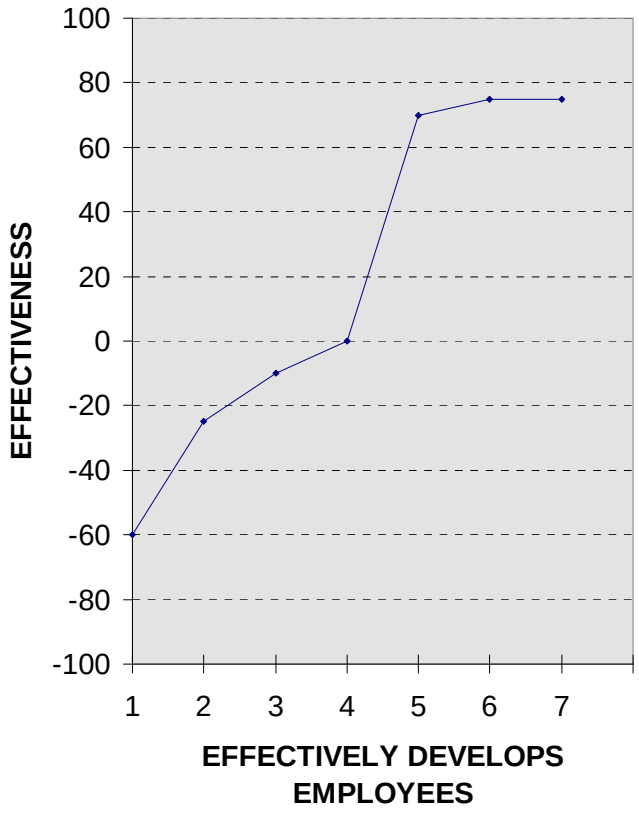


Figure 1 Example Contingency Graph

Several studies support the use of contingencies for formulating overall scores for various jobs (Pritchard, 1992; Pritchard, 1995; Pritchard, Holling, Lammers, & Clark, 2002; Pritchard, Harrell, DiazGranados, & Sargent, 2007; Pritchard, Paquin, Decuir, McCormick, & Bly, 2002); however, this method has not been used for producing OARs based on combinations of various dimension ratings. Based on previous findings that the use of contingencies has been shown to be successful in a variety of other settings (Pritchard, 1995; Pritchard, Harrell, DiazGranados, & Sargent, 2007; Pritchard & Roth, 1991), it was expected that this approach could be successfully implemented in an AC arena using SMEs. Success in the AC context first means that the SMEs are able to use the approach and actually develop contingencies.

Thus,

Hypothesis 1: SMEs are expected to be able to successfully develop contingencies.

Advantages of the Contingency Method

There are several contributions of the contingency-based method (Pritchard, 1992) beyond existing combination methods. First, one can easily identify the importance of each dimension relative to the other AC dimensions by evaluating the range of scores on the y-axis. Specifically, dimensions with larger ranges add to or take away from the criterion of interest in greater amounts than those with smaller ranges. This is shown in Figure 2, which shows the contingencies for four dimensions. Effective communication skills is the most important dimension with its range expanding from -90 to +100. The other dimensions are important, but not as important as indicated by their lower ranges. Technical knowledge and proficiency has the lowest range (e.g., -100 to +5), thus is the least important of the four measures.

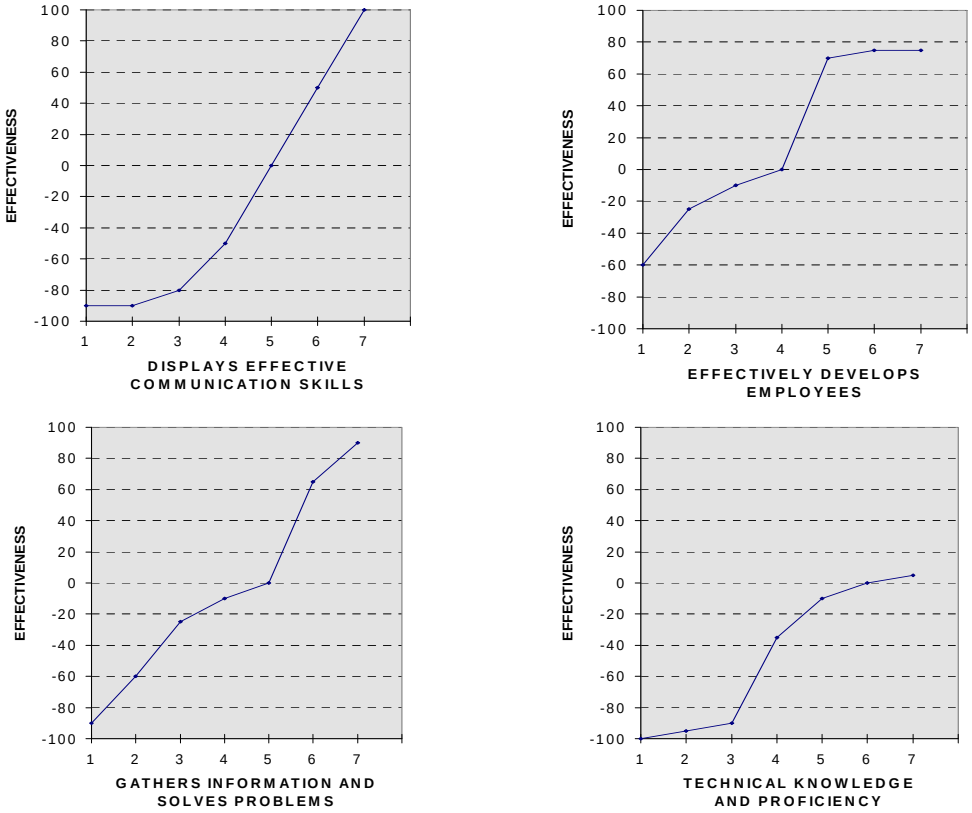


Figure 2 Comparison of Four Contingencies

Also, contingencies identify the minimum expected level of performance on each dimension assessed. This is the dimension score associated with an effectiveness score of zero. For example, Figure 2 shows that the minimum expected level for the dimensions of displays effective communication skills, effectively develops employees, gathers information and solves problems, and technical knowledge and proficiency are 5, 4, 5, and 6, respectively. Contingencies also identify what levels of performance on the dimension are good or bad. For example, Figure 2 shows that any level of performance on the measure of "gathers information and solves problems" above a 5 indicates performance above the minimally acceptable with values of 6 and 7 being very positive. Levels of performance below a 5 indicate poor

performance with values of 1 and 2 being very low. Additionally, contingencies place each dimension rating on a common scale of amount of contribution made to the criterion of interest. In the case of assessment centers for promotion, the criterion is expected performance if promoted. Because all the dimensions are on the same scale, the effectiveness scores from each dimension can be summed to obtain a single overall score representing contribution made to the criterion of interest (Pritchard, 1992). More specifically, if an AC had a total of 10 dimensions assessed, the y-axis score (i.e., amount of contribution made to the criterion of interest) for each dimension would be summed to calculate the single OAR.

Another advantage of contingencies is their ability to identify non-linearities where there are changes in effectiveness where certain levels of change in the dimension level do not yield equal amounts of change in the amount of contribution made to the criterion of interest. For example, Figure 2 shows that a rating of "6" on the measure of "effectively develops employees" has the same effectiveness score as a rating of "7" indicating that no further contribution will be made to the criterion of interest once you obtain at least a score of "6" on the measure. In this case, a point of diminishing returns is reached at a score of 6 and a rating higher than 6 does not indicate higher expected performance. The more traditional weighting systems do not capture this non-linearity. Unit or any other linear weighting system applies an equal weight no matter what the value of the dimension is. This assumes that any given change in the rating produces an equal change in expected performance. If all contingencies were linear, this new method would not provide any unique information beyond that of the traditional methods. Having this added bonus of capturing non-linearity, which is only provided with the contingency method, is the main reason for the expectation of higher validity with this new scoring method.

Contingencies also provide for the ability to identify which dimensions one should focus on if looking to make the greatest increase in expected gain in contribution to the criterion of interest. This can be accomplished by examining the function of the graph and calculating expected gain in contribution if you were to improve a certain amount on a given dimension. For example, Figure 2 shows that on the one hand, increasing a rating from a score of 5 to a score of 6 on the measure of "gathers information and solves problems" provides a gain of 65 points for the overall effectiveness score. On the other hand, increasing the rating from a score of 3 to a score of 4 on the measure of "effectively develops employees" provides a gain of 10 points for the overall effectiveness score. Therefore, if an individual has a score of 5 on "gathers information and solves problems" and a score of 3 on "effectively develops employees," the suggestion would be to focus on improving the rating on "gathers information and solves problems" rather than the improving the rating on "effectively develops employees" due to the gain of 65 points in overall effectiveness versus a gain of 10 points. These characteristics of contingencies have the potential to provide unique information about the AC performance dimensions specific to the job being assessed.

For these characteristics to actually provide unique information in an AC, a number of features should be expected. The first is that contingencies should be sensitive to different jobs. If jobs that are truly different result in the same contingencies, the contingencies are not sensitive to job differences. Anecdotal evidence suggests that for the jobs in this study, certain dimensions may be more important at the level of Fire Lieutenant, while others may be more important at the level of Fire Captain. For example, judgment and decision making, analytical ability, incident command, and conflict management tend to be used more often in the field; thus, they were expected to have different features exhibited in their contingencies (e.g., differences in

the ranges and differences in linearity) for the job of Fire Lieutenant compared to Fire Captain. Thus,

> Hypothesis 2: Contingencies developed for dimensions used more often in the field (e.g., judgment and decision making, analytical ability, incident command, and conflict management) will differ based on the level of the job.

Comparing the Scoring Methods

One way to compare the results of the contingency-based method with the other four methods is to correlate the final scores obtained for each method. The logic is if the correlations are high, the different methods provide similar information. However, this comparison is likely to provide no valuable information for assessing the value of the contingency-based method (Pritchard, Watson, Kelly, & Paquin, 1998). From a mathematical standpoint, it is expected that composites formed using the same variables will correlate very highly with one another irrespective of the weighting strategy utilized when there are more than 10 variables that are moderately correlated with one another (Arthur, Doverspike, & Barrett, 1996; Ree, Carretta, & Earles, 1998; Wilks, 1938). For example, Pritchard, Watson, Kelly, and Paquin (1998) found that overall scores based on contingencies correlated very highly (r = .87-.97) with three other methods of combining evaluations of teacher effectiveness.

Although an interpretation of the contingency approach yielding high correlations with the other four combination methods could be that the contingency method fails to add any unique information, evidence has been provided to suggest otherwise. For example, prior research has shown that highly correlated combination methods can result in different decisions from a selection standpoint based on the shifting of individuals in and out of the eligible top ranks to be

selected based on weighting method (Arthur, Doverspike, & Barrett, 1996; Pritchard & Roth, 1991; Pritchard, Watson, Kelly, & Paquin, 1998). Specifically, different people are represented at the extremes where cut-offs for selection are likely to be made.

To appropriately assess the value of the contingency-based combination method in comparison to the other four methods, it would be more fitting to examine the composition of the individuals comprising various cut-off ranks to determine if different job candidates comprised the eligible ranks when different scoring methods are used. Specifically, different cut-offs, which are discussed in more detail in the next chapter, can be used in actual settings. It is important to examine which individuals make those cut-offs because these are the individuals that will qualify for the eligible ranks. This means that looking at whether there are different folks in the top groups (e.g., eligible ranks) using the different methods is one way to see if the methods are different. The combination approach suggested by the current study was to rescale each level of performance on each AC dimension to a common scale (i.e., expected promotability). This was accomplished by using SMEs to establish contingencies for all dimensions and converting all possible dimension ratings expected promotability scores to ultimately sum all of those dimension scores to obtain an overall effectiveness score. Thus, although the various scoring methods are expected to be highly correlated:

> Hypothesis 3: The final results obtained for the contingency approach are expected to be different than the final results obtained for the traditional approaches in that the people remaining in the top ranks will change depending on the combination method used.

Also, the contingency approach directly links to important aspects of the job without overweighting cognitively loaded dimensions beyond the level of performance actually needed

on that dimension to reap the maximum gain in expected promotability. Minorities scoring lower on those particular dimensions, which are known to produce more adverse impact, would not be penalized on their overall expected promotability score if they perform at least at the level providing the maximum gain in expected promotability. Therefore, overall scores for minorities would be comparable to their majority member counterparts even if the majority group members perform better on the cognitively loaded dimensions. Consequently, there should be less adverse impact for the contingency method. Thus,

> Hypothesis 4: Scoring the AC with the contingency-based method is expected to result in less adverse impact than scoring the AC using the traditional combination methods.

CHAPTER THREE: METHODOLOGY

Participants

The sample for this study consisted of 444 AC participants including 326 Fire Lieutenant candidates and 118 Fire Captain candidates during an assessment center for at large testing facility. Tables 2 and 3 list the demographics of the candidates for Fire Lieutenant and Fire Captain, respectively. For both tables the "Total Applications" refer to the total number of individuals submitting a complete application for the job. "Failed to Appear at Test" indicates the number of individuals invited to the AC, but failed to attend. "Completed Test" refers to the total number of individuals completing the AC. Ten current Fire Lieutenants, Fire Captains, and Fire Battalion Chiefs voluntarily served as SMEs for the contingency development processes of the proposed study (see Table 4 for demographic information). Assessors included thirty-six volunteers for Fire Lieutenant and thirty-four volunteers for Fire Captain with rank of Fire Lieutenant or higher in their home department representing fire and rescue departments from across the United States (see Appendices A and B for demographic information).

Table 2: Fire Lieutenant Candidate Demographics

Demographic Category	Total Applications	Failed to Appear at Test	Completed Test
Black	153	36	116
White	310	98	209
Other	4	2	1
Total	467	136	326
Male	444	130	310
Female	23	6	16
Other	0	0	0
Total	467	136	326

Table 3: Fire Captain Candidate Demographics

Demographic Category	Total Applications	Failed to Appear at Test	Completed Test
Black	54	6	46
White	72	10	55
Other	29	11	17
Total	155	27	118
Male	132	20	103
Female	5	0	5
Other	18	7	10
Total	155	27	118

Table 4: SMEs for Contingency Development

Fire Lieutenant			
SME #	Rank	Race	Sex
1	Lieutenant	Black	Female
2	Captain	Black	Female
3	Lieutenant	Black	Male
4	Captain	Black	Male
5	Captain	White	Male
Fire Captain			
6	Captain	Black	Male
7	Battalion Chief	Black	Male
8	Captain	White	Male
9	Battalion Chief	White	Male
10	Captain	White	Male

Fire Lieutenant Test Administration Procedure

The Fire Lieutenant test was administered at a large testing facility over the course of two days where candidates were required to attend both days of the test to be considered for the position. Candidates checked-in for the test by providing photo identification, turning in a signed Test Agreement Form, signing the candidate roster, and signing a Confidentiality Agreement for the exam (see Appendix C). The AC for Fire Lieutenant consisted of two components including a video-based supervisory exam and a video-based technical exam. The two exams (see Table 5 for the specific dimensions assessed by each component of the exam) contained three independent scenarios to which candidates had to respond (see Appendix D for more details of the scenarios).

Table 5: Fire Lieutenant Test Plan

KSAO Dimension	Assessment Method					
	Supervisory Exam			Technical Exam		
	1	2	3	1	2	3
Dimension 1: Policies and Procedures		X				
Dimension 2: Safety and Life Preservation				X	X	
Dimension 3: Firefighting Tactical Knowledge				X	X	
Dimension 4: Supervisory Ability	X	X	X			
Dimension 5: Leadership Ability			X			
Dimension 6: Conflict Management			X			
Dimension 7: Fire Behavior Knowledge				X		
Dimension 8: Analytical Ability				X	X	
Dimension 9: Judgment and Decision Making					X	
Dimension 10: Oral Communication			X			
Dimension 11: Written Communication			X			
Dimension 12: Incident Command/IMS				X	X	X

Note. Appendix D provides a more detailed description of each of the three scenarios for both exams.

Fire Captain Test Administration Procedure

The Fire Captain exam was administered at a large testing facility over the course of one day. Candidates checked-in for the test by providing photo identification, signing the candidate roster, and signing a Confidentiality Agreement for the exam. The Fire Captain Exam consisted of three unique phases (see Appendix D for a more detailed description), which simulated a single shift for a Fire Captain and required candidates to complete several tasks that may be performed during a shift by a Fire Captain (see Table 6 for the specific dimensions assessed by each task).

Table 6: Fire Captain Test Plan

KSAO Dimension	Assessment Method								
	Phase I		Phase I					Phase III	
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9
Dimension 1: Judgment and Decision Making			X						
Dimension 2: Management Ability	X	X							
Dimension 3: Oral Communication					X				
Dimension 4: Written Communication								X	X
Dimension 5: Professionalism				X	X				X
Dimension 6: Incident Command						X	X		
Dimension 7: Supervisory Ability				X					
Dimension 8: Leadership							X		
Dimension 9: Conflict Management				X					
Dimension 10: Analytical Skills									X
Dimension 11: Departmental/Jurisdictional Knowledge			X		X				
Dimension 12: Technical Knowledge – Emergency Response						X	X		
Dimension 13: Technical Knowledge – Firefighting						X	X		

Note. Appendix D provides a more detailed description of each of the tasks for all three phases of the exam.

Assessment

The assessments of the Fire Lieutenant and Fire Captain exams were conducted over the course of one week each. Assessors worked in appropriately diverse pairs to rate candidate performance on each component of the examination to ensure that each candidate was scored by at least one individual who was demographically similar (e.g., match on race, sometimes gender) to him/herself, thus avoiding potential rater biases based on race or gender. All assessors

underwent specialized training sessions conducted by experienced job analysts which included information and practice opportunities for observing and recording behavior, categorizing behavior, evaluating behavior, and making ratings of behavior, as well as how to appropriately take notes and categorize notes in terms of performance dimensions. Assessors also received training regarding the administrative aspects of the assessment, such as completing rating forms and entering ratings into the computer system.

As a critical part of training, assessors were given multiple opportunities to practice making ratings based on observations of candidate responses to each component of the exam. Each assessor made independent ratings of performance using the benchmarks developed for the exercise. Benchmarks for many of the exercises were anchored to three points of a five-point rating scale: (1) Unacceptable, (3) Acceptable and (5) Outstanding. For other exercises, a checklist rating scale was used in which candidates were given one point for each benchmark hit. All assessors made independent, preliminary ratings of performance using benchmark rating forms provided. Once preliminary ratings were obtained, assessors discussed their ratings and made independent, final ratings with the requirement that final ratings had to fall within one scale point of each other for Likert rating scale, and checklist ratings had to match perfectly. In special cases where discrepancies could not be reconciled, the candidates were rated by a second panel to ensure appropriate consensus ratings were reached.

Contingency Development

Ten subject matter experts (SMEs, five for the Fire Lieutenant job and five for the Fire Captain job) with rank of Fire Lieutenant or higher in their home department volunteered to

participate in the contingency development process. These SMEs represented fire and rescue departments from across the United States. There were several steps followed in the development of the contingencies for each dimension for both the Fire Lieutenant and Fire Captain jobs, which closely followed the basic steps of contingency development provided by Pritchard (1990). Before starting the contingency development process, the SMEs were given a detailed presentation explaining the contingency process, its advantages, and how it was related to the Fire Lieutenant and Fire Captain ACs. Additionally, all SMEs were given a detailed description of the specific job they were developing contingencies for (e.g., Fire Lieutenant or Fire Captain) including all exercises, all dimensions, job analysis weights for each dimension, and the rating scales used for each dimension. Finally, SMEs were provided with a contingency development worksheet, which was used to record information for the first three steps of contingency development. This worksheet recorded information such as the dimensions assessed, the ratings and rankings for each dimension, the minimum expected level of performance for each dimension in addition to the effectiveness scores associated with each minimum and maximum level. A more detailed explanation of these ratings is provided below during the description of the steps of contingency development.

The first step of the contingency development process would typically be for SMEs to identify the minimum and maximum rating levels for each dimension, which are reflected on the x-axis on the contingency graph. However, in the AC context these levels were easily identifiable based on the rating scales used during assessment. For example, if the leadership dimension was rated using a 5-point scale, as many of the dimensions were, ranging from 1 = unacceptable, 3 = acceptable, to 5 = outstanding, clearly the minimum dimension level would be "1" and the maximum dimension level would be "5."

However, there were five dimensions (e.g., incident command, safety and life preservation, firefighting tactical knowledge, and analytical ability) for Fire Lieutenant and four dimensions (e.g., incident command, management ability, departmental/jurisdictional knowledge, emergency response technical knowledge, and firefighting technical knowledge) for Fire Captain that utilized two different rating scales (e.g., 5-point rating scale and a behavioral checklist). Therefore, scores had to be standardized (e.g., *z* scores) to obtain an overall dimension rating on a common metric so as not to overweight or underweight any score. This made it a little bit more complex to explain the minimum and maximum dimension levels to SMEs due to their unfamiliarity with *z* scores. To facilitate the SMEs ability to successfully complete contingency development, they were first given a brief overview of what are *z* scores. I explained the notion of the standard normal distribution with a mean = 0 and standard deviation = 1 and provided an example of the standard normal distribution curve for the SMEs to refer back to throughout the contingency development process. The SMEs were then informed that the minimum and maximum levels for the *z* scores were -3 and +3, respectively.

After discussing this information with the SMEs in addition to explaining how *z* scores transformed to percentiles, the group agreed that it would be easier and more intuitive to use percentiles on the x-axis during contingency development rather than *z* scores. Ultimately, the percentiles corresponding to *z* scores ranging from -3 to +3 (e.g., 1^{st} percentile and 99^{th} percentile, respectively) were used as the minimum and maximum dimension level ratings for both jobs. All minimum and maximum values were recorded on the contingency worksheet.

The second step was to agree upon the minimum acceptable level of performance for each dimension. This level on the x-axis would correspond to an effectiveness score (point on the y-axis of the graph) of zero. For example, if the leadership dimension was rated using a 5-

point scale ranging from 1 = unacceptable, 3 = acceptable, to 5 = outstanding, SMEs agreed that a score of "3" represented the minimum expected level of performance for leadership. If the leadership dimension was rated using a behavioral checklist ranging from zero to seven, SMEs discussed what number represented the minimum acceptable level for the dimension until they reached a consensus. I explained to the SMEs that this point on the graph would reflect a level of performance that is nether good nor bad, but acceptable. I further explained that this is a level of performance for which an individual would not be complimented nor criticized. I provided the SMEs with examples of previously developed contingencies and highlighted the minimum acceptable level of performance for each to provide them with a visual representation of how this minimum acceptable level of performance is reflected on the contingency graph.

Although the SMEs did engage in several minutes of discussion regarding what should be the minimum expected level of performance for the various dimensions, there was no disagreement when it was time to make a final decision. Similarly, for the dimensions showing percentiles on the x-axis, SMEs discussed what percentile represented the minimum acceptable level for the dimension until they reached a consensus. All minimum acceptable levels of performance were recorded on the contingency worksheet for each dimension.

The third step is for the SMEs to identify the effectiveness values (y-axis scores) for the minimum and maximum dimension levels identified in the first step. To accomplish this task the SMEs first ranked and rated each minimum and maximum dimension value regarding importance to the overall criterion (e.g., expected promotability). I explained to the SMEs that the maximum value deemed most important received a rank of 1; the second most important received a rank of 2, and so on until all maximums were ranked. To facilitate the process of ranking each dimension, I asked the SMEs to imagine a person is at the minimum expected level

on all 12 dimensions and they had the ability to move this person up on only one dimension. The dimension that would improve the expected promotability for the person the most would receive a rank of 1. The dimension that would improve the expected promotability the second most would receive a rank of 2. This process continued all maximum dimension values were ranked. Similarly, the minimum value deemed worst for the overall criterion (e.g., expected promotability) if all dimensions were at the zero point leaving only one dimension to be at its minimum would receive a rank of "1." The process for ranking all other minimums was analogous to the process described for ranking the maximums (i.e., the second worst had a rank of 2, the third worst had a rank of 3, etc.). All rankings were recorded on the contingency worksheet.

Following the guidelines set by Pritchard (1990) the first ranked maximum automatically received an effectiveness rating of +100. The other maximums received ratings corresponding to their importance relative to the first ranked maximum using percentages of 100. For example, if the maximum value for leadership was ranked 1, the maximum value for supervisory skills was ranked 2, and the maximum value for analytical skills was ranked 3, leadership would automatically receive an effectiveness rating of +100. Supervisory and analytical skills would receive effectiveness scores reflecting their importance relative to leadership which would be lower than +100 (e.g., + 95, + 90, respectively). Specifically, SMES were asked to indicate the level of importance for the second dimension compared to the first. If they believed the maximum score on supervisory skills was 95% as valuable for performance as was scoring the maximum on leadership, the supervisory skills maximum would receive an effectiveness rating of +95. If they believed the maximum on supervisory skills was only half as valuable as getting the maximum on leadership, they would give it an effectiveness score of +50. If they felt getting

the maximum on analytical skills was 90% as important as getting the maximum on leadership, this maximum would receive an effectiveness score of +90, and so on. The SMEs were also instructed that the effectiveness scores should correspond to the ranks such that a maximum with the rank of "3" should not have an effectiveness score larger than a maximum with the rank of "1."

The main distinction when rating the minimum dimension values was that the negative with a rank of "1" did not automatically receive an effectiveness score of -100. This number could have been more or less negative (e.g., -150 or -75, respectively) depending on how the SMEs viewed the amount of negative contribution to the overall criterion made by the most negative dimension compared to the most positive contribution made by the most positive dimension. SMEs discussed this issue to determine the effectiveness value for the minimum with a rank of one and preceded to rate the other minimums relative to the most negative minimum using a process parallel to what they had done for the ratings of the maximums. All ratings were recorded on the contingency worksheet. At this point in the process, three points on each contingency had been determined, the effectiveness score for the minimum level on the dimensions, the effectiveness score for the maximum on the dimension, and the minimally acceptable score which corresponded to an effectiveness score of 0.

Once these first three steps were completed, I provided the SMEs with a 15 minute break to allow me the opportunity to utilize the information on the contingency worksheet to plot the effectiveness scores for the minimum, maximum, and minimum expected level for each dimension on a transparency with a blank contingency template (see Figure 3). When the SMEs returned from their break, they were presented with the contingency graph using an overhead projector where transparencies were presented for each dimension, one at a time.

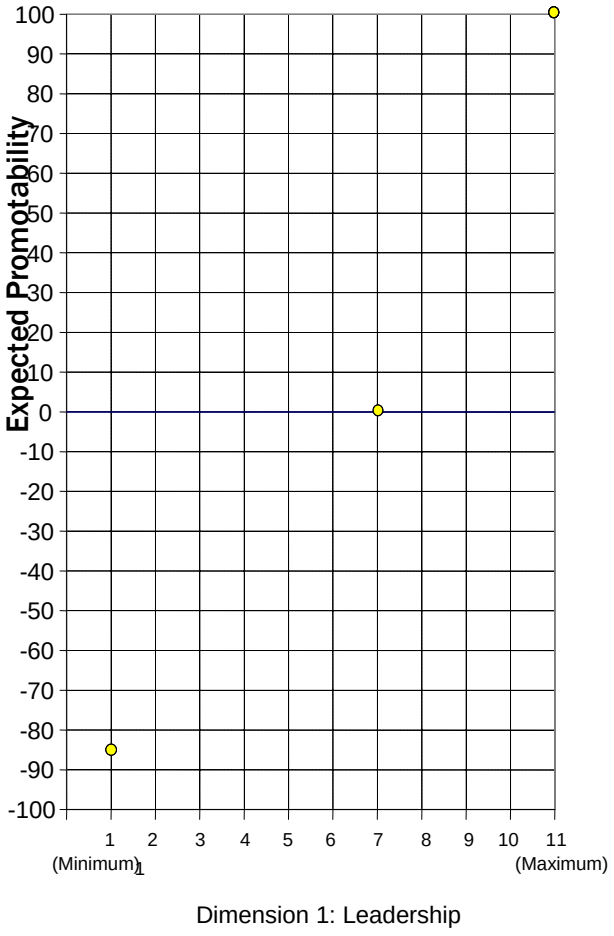


Figure 3 Contingency Template with Maximum, Minimum, and Zero Point

The ultimate goal was for the SMEs to engage in group discussion to determine the shape of each graph by identifying the remainder of the points on the graphs. At this point during the contingency process I explained three distinct shapes that the graphs could have (e.g., linear, critical mass, and diminishing return) and provided examples to facilitate their understanding of the concepts. Specifically, a linear graph would mean that for each increase on the x-axis there would be an equal increase in expected promotability on the y-axis. For example, Figure 4 shows that moving from a 3.00 to a 3.50 on supervisory ability results in an increase of 25 for the

expected promotability, and moving from a 3.5 to a 4 also yields an increase of 25 for the expected promotability.

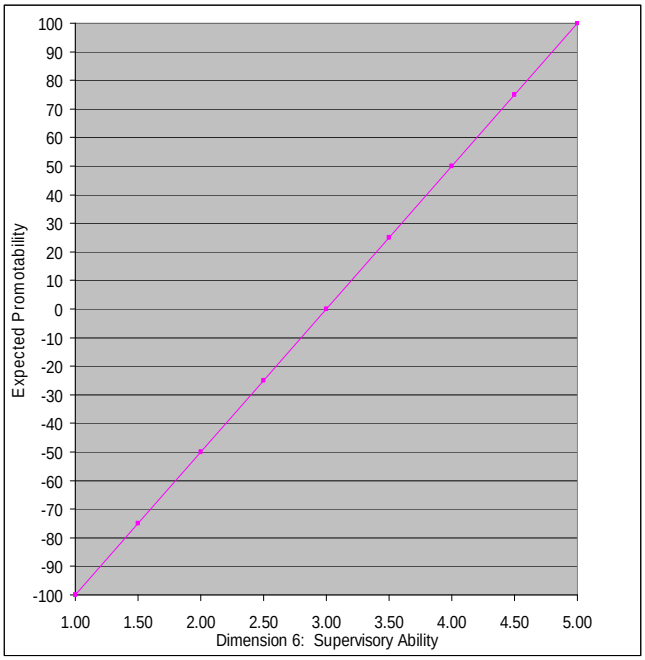


Figure 4 Example of a Linear Contingency

A critical mass would be a graph that starts out very low in terms of expected promotability and sharply increases. For example, Figure 5 shows that the expected promotability is very low at the levels of 1.00 to 2.00 on oral communication; however, there is a steep increase in expected promotability after 2.00.

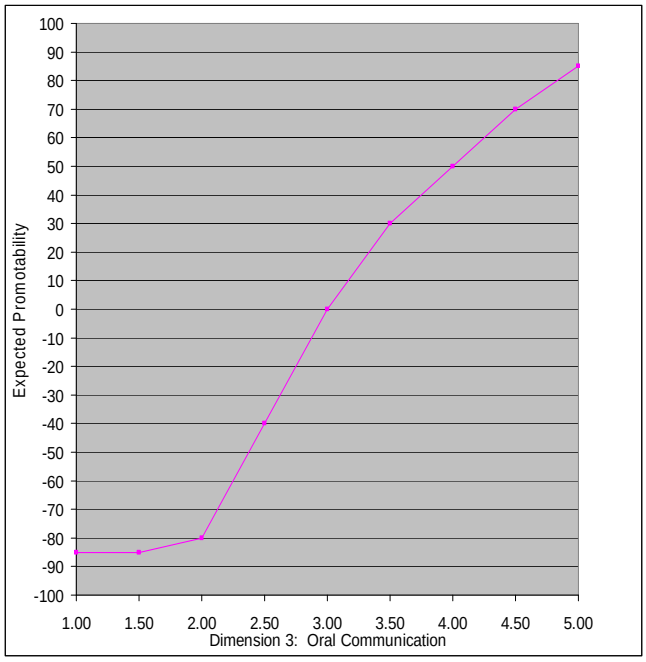


Figure 5 Example of a Critical Mass

A diminishing return would be a graph where a certain point would be reached on the x-axis that would provide little or no further increases in expected promotability. For example, Figure 6 shows that once a person has reached a level of 4.00 on written communication, they have already obtained the highest score on expected promotability that is possible for this dimension. Thus, any further increases on this dimension would not provide additional increases in their expected promotability score.

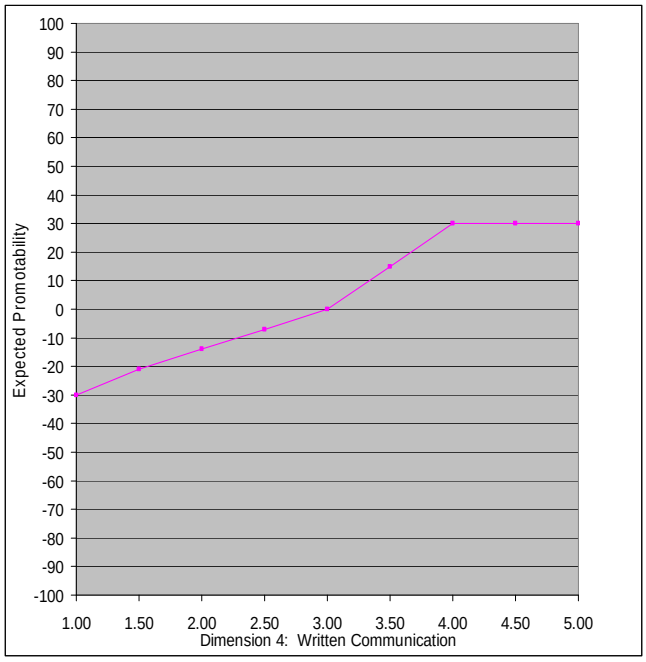


Figure 6 Example of a Diminishing Return

After explaining these different shapes to the SMEs, we focused on the shape of the top half of the graph (i.e., the shape connecting the zero point with the maximum) followed by the shape of the bottom half of the graph (i.e., the shape connecting the zero point with the minimum). The use of an overhead projector with transparencies highlighting the contingency template was used to facilitate this process. Once this process was complete, the shape of the graph was fine tuned to ensure it accurately reflected the relationship between the dimension values and the overall promotability scores relative the other dimensions. For example, if the SMEs developed a contingency reflecting a diminishing return starting at the 4.00 point on a 5-point rating scale, I made sure they understand and agreed that this meant that there would be no further gains in expected promotability for this dimension beyond a rating of 4.00. This type of overview and discussion took place after completing all steps of contingency development for all

dimensions. Final contingency graphs for all dimensions for both Fire Lieutenant and Fire Captain can be found in Appendices D and E, respectively.

Throughout the contingency development process I made it a point to take detailed notes regarding the ability of the SMEs to develop the contingencies. Specifically, there were five questions I addressed in my notes regarding whether the SMEs were able to develop the contingencies including: 1) Were the members able to perform each step of contingency development? 2) Was there disagreement at each step, how much, how was it resolved? 3) Was total consensus obtained at the end of the process? 4) Did SMEs appear to be involved during the process? 5) Did SMEs agree resulting contingencies were accurate reflections of their positions? In addition to my notes, I also sent these same questions and 5- point rating scales to the five SMEs for Fire Captain contingency development (see Appendix G for the questionnaire sent to SMEs). They were asked to indicate whether their views of the process were similar mine. I would have liked to collect this information from the five SMEs for Fire Lieutenant contingency development as well, but I did not collect their contact information during my meeting with them.

Support for the first hypothesis would be demonstrated if there was general consensus by the SMEs and my notes regarding the success of the development process and if the resulting contingencies provide more information about the AC performance dimensions than would be apparent by examination of the dimension ratings themselves (e.g., evidence that there are differential minimum performance levels across dimensions, differential nonlinearity across dimensions). It was necessary to address these various concerns, which each contributed unique evidence, to adequately determine if contingencies were successfully developed.

It was expected for the second hypothesis, based on the knowledge of the jobs, that only four of the common dimensions (e.g., leadership ability, supervisory ability, written communication, and oral communication) would have contingencies with similar shapes and similar ranges. Further support for this hypothesis would be shown if there were differences in candidates at the various ranks based on the reverse contingency method. If the contingencies were the same, results would show that the same candidates are selected. Although there was no statistical test conducted on the differences in candidates at the various ranks based on this reverse contingency method, there was a subjective determination made of the practical difference in the candidates.

Calculating OARs.

For the four traditional combination methods for both Fire Lieutenant and Fire Captain, the weighted dimension scores were summed to obtain OARs. Specifically, each of the weighted dimension scores was summed to form an OAR, which was then used to rank the candidates. For consensus ratings, the standardized individual ratings for each assessor were first averaged before summing scores to form an OAR. For the contingency-based method, the overall score was obtained by converting the candidates' overall dimension scores to their corresponding effectiveness scores and then summing all effectiveness scores to rank candidates.

Determining Cut-off Ranks

Cut-off ranks to identify which candidates are hired in the operational ACs were determined and used for research purposes to test some of the hypotheses in the study. The cut-

off ranks for Fire Lieutenant were top 10, top 39, and top 48. The cut-off ranks for Fire Captain were top 10, top 22, and top 31. The "rule of 10" which the organization used requires that when there is only one vacancy for a given position, 9 ranks are added to that one vacancy to produce a cut-off at rank 10. Specifically, 9 ranks are added to the number of vacancies for the targeted position to determine where to set the cut-off. As a result, the minimum cut-off possible is always at rank 10 given there is at least one vacancy. If there are two vacancies, the cut-off moves to rank 11, and so on. This was the process used to determine the first cut-off rank 10 with the expectancy of at least one vacancy.

Determining the other two cut-off ranks was more involved such that it was based on first calculating the projected number of hires over the life of the register (i.e., list including all candidates passing the selection test and subsequently deemed eligible for the job, which typically expires after 18 months). The projected number of hires was based on the sum of the number of hires from the past register, divided by the number of months this register was in place, times the projected life of the current register in months. Once the projected number of hires has been calculated, the number of ranks is determined by assuming top down selection of the number projected to be hired over the life of the register. The rank of the individual at this cut-off would determine the starting rank cut-off used for the list of eligible job candidates. Adding 9 additional ranks (i.e., the rule of ten) would provide the next cut-off rank. For example, if the projected number of hires for a position has been determined to be 20, the next step in determining the cut-off is to first identify what rank the 20^{th} individual on the list holds. This individual's rank would be used as the next cut-off rank. To determine the next cut-off rank, 9 ranks would be added to this individual's rank.

A step-by-step example of this process may make this clearer. First, the projected number of hires over the life of the register must be determined. This number was estimated by using the number of hires from the past register (i.e., sum = 38), divided by the number of months (31) this register was in place (December 2004 to June 2007), times 18 (the projected life of the current register in months). This calculation results in the number 22. Next, the number of ranks was determined by assuming top down selection of 22 individuals (i.e., the number projected to be hired over the 1.5 year life). The rank of the 22nd individual corresponded to Rank 22, which results in a list of eligible candidates starting with Rank 22 and going to Rank 1. This is how the cut-off rank 22 was determined. Moreover, adding 9 additional ranks (i.e., the rule of ten) to this cut-off produced the final cut-off rank 31 for Fire Captain. The exact process was used to determine the cut-off ranks for Fire Lieutenant.

Although this was the chosen strategy for determining cut-off ranks for this study, another strategy would have been to arbitrarily choose cut-off ranks that may seem more intuitive (e.g., top 10, top 20, and top 30) and consistent at both job levels. However, the current approach was taken to justify the cut-offs from an applied standpoint. Specifically, the organization from which the data were collected had a procedure already in place for determining final ranks based on OARs, which provides a rationale for how the cut-offs chosen are linked to possible selection decisions.

The issue of which individuals are included in the top cut-off group is one way of determining the differences between the various ways to get the overall assessment center score. Support for the third hypothesis would be shown if the composition of the top cut-off group is different when the contingency method is used rather than the other four methods. This would indicate that the contingency method yields different selection decisions based on the differences

in candidates comprising the various cut-off ranks when compared to the other combination approaches. However, if the rank order of job candidates is the same when using the new contingency method compared as it is with the other four methods, this would mean that the contingency approach fails to provide any unique information. The same decisions for eligibility of promotion would be made as with the other four methods, thus, the third hypothesis would not be supported. The fourth hypothesis will be tested by looking at differences in adverse impact at the various cut-off ranks for all combination methods. This hypothesis will be supported if the contingency method results in less adverse impact overall than the other methods.

CHAPTER FOUR: FINDINGS

To gain an understanding of how well the items on each AC component performed, reliabilities were estimated. Inter-rater reliability estimates of the various AC components were first examined for the 5-point ratings using the two-way mixed model intraclass correlation coefficient (ICC). This estimate reflects the extent to which ratings provided by different assessors are proportional when they are expressed as deviations from their means (Tinsley & Weiss, 1975). The reliability of each AC component was calculated by computing the intraclass correlation (ICC) for the preliminary ratings made by the two assessors for each dimension assessed for the Fire Lieutenant and Fire Captain examinations. Cronbach's Alpha, a measure of internal consistency, was used to estimate the reliabilities of the checklists that were used to measure dimensions in the Fire Captain exam. Table 7 presents the intraclass correlations for each component of the Fire Lieutenant examination by performance dimension. Results indicated that the reliability for the assessor ratings ranged from .71 to .98 across the test components.

Table 7: Fire Lieutenant Reliability of Assessor Ratings for all Exam Components

Exam Components	Dimension	Reliability
Supervisory Video Scenario 1	Supervisory Ability Rating 1	.76
	Supervisory Ability Rating 2	.82
Supervisory Video Scenario 2	Policies and Procedures	.82
	Supervisory Ability	.80
Supervisory Video Scenario 3	Supervisory Ability	.74
	Leadership	.72
	Conflict Management	.71
	Written Communications	.82
	Verbal Communications	.81
Technical Video Scenario 1	Safety and Life Preservation	.96
	Firefighting Tactics	.96
	Fire Behavior	.96
	Incident Command	.96
	Analytical Ability	.94
Technical Video Scenario 2	Incident Command Rating 1	.85
	Incident Command Rating 2	.81
	Incident Command Rating 3	.85
	Firefighting Tactics Rating 1	.76
	Firefighting Tactics Rating 2	.85
	Firefighter Tactics Rating 3	.98
	Safety and Life Preservation Rating 1	.85
	Safety and Life Preservation Rating 2	.92
	Analytical Ability	.84
	Judgment and Decision Making	.87
Technical Video Scenario 3	Incident Command Rating 1	.96
	Incident Command Rating 2	.87
	Incident Command Rating 3	.89

Table 8 presents the intraclass correlations and internal consistencies for each component of the Fire Captain examination by performance dimension. Results indicated that the reliability of ratings made by assessors for all except two exercises ranged from .62 to .84. The checklists for Task 7, which measured Technical Knowledge of Emergency Response (Dimension 12), had an internal consistency of .50. In addition, the checklist for Task 7, which measured Leadership (Dimension 8), had an internal consistency of .03. All inter-rater reliabilities for the benchmarks making up the checklists measuring Technical Knowledge of Emergency Response for Task 7 were greater or equal to .65. All inter-rater reliabilities for the benchmarks measuring Leadership in Task 7 were greater than or equal to .68. These reliability estimates suggest that pairs of raters consistently utilized these benchmarks when assessing candidates' Technical Knowledge of Emergency Response and Leadership. Furthermore, the SMEs who developed Task 7, and its benchmarks, indicated through ratings that each of these benchmarks were relevant to the task and related to the dimensions being measured. For these two reasons, the checklists measuring Technical Knowledge of Emergency Response and Leadership in Task 7 were retained.

Table 8: Fire Captain Reliability of Assessor Ratings for all Exam Components

Exam Components	KSA	Reliability
In-Basket - Task 1	Management Ability	.79
In-Basket - Task 2	Management Ability	.65
Video Work Sample Task 3	Departmental/Jurisdictional Knowledge	.74
	Judgment and Decision Making	.76
Video Work Sample Task 4	Professionalism	.82
	Supervisory Ability	.79
	Conflict Management	.72
Video Work Sample Task 5	Oral Communication	.75
	Professionalism	.78
	Departmental/Jurisdictional Knowledge	.66
Video Work Sample Task 6	Incident Command Rating 1	.67
	Technical Knowledge-Emergency Response	.79
	Incident Command Rating 2	.82
	Technical Knowledge-Firefighting	.77
Video Work Sample Task 7	Incident Command	.66
	Technical Knowledge-Emergency Response	.50
	Technical Knowledge-Firefighting	.62
	Leadership Ability	.03
Write-up Task 8	Written Communication	.80
Write-up Task 9	Written Communication (Memo to Chief)	.79
	Analytical Skills	.84
	Written Communication (Letter to Citizen)	.72
	Professionalism	.75

Hypothesis 1 Results

The first hypothesis stated that SMEs are expected to be able to successfully develop contingencies. In one sense, it is clear that this hypothesis was supported because both groups did, in fact, develop sets of contingencies for their respective jobs. However, we can investigate this issue in more detail by considering the following questions: 1) Were the members able to perform each step of contingency development? 2) Was there disagreement at each step, how much disagreement, and how was the disagreement resolved? 3) Was total consensus obtained at the end of the contingency process? 4) Did SMEs appear to be involved during the process? 5) Did SMEs agree resulting contingencies were accurate reflections of their positions? To answer these questions, I will use my notes taken during the contingency development process and supplement these with the reactions data from the five SMEs who responded to the questionnaire about contingency development

I first visually inspected the actual contingencies developed to assess whether these resulting contingencies appeared to be well-developed. There were five questions qualitatively analyzed to shed light on this issue including: 1) Was there differential importance among the various dimensions for both job levels? 2) Was the differential importance from the contingencies related to the job analysis weights given to the dimensions prior to the contingency development? 3) Were differential minimum acceptable performance levels identified for the various dimensions? 4) Did the minimum acceptable performance identified by SMEs differ from a rating of "Acceptable" on the initial scale use to rate candidates on the given dimension? 5) Was there distinct non-linearity in the functions of the final contingency graphs?

The first question, which was supplemented with a 1-5 point rating scale (e.g., 1 = strongly disagree, 3 = agree, and 5 = strongly agree), assessed if the SMEs were able to perform each step of contingency development. The mean rating on this item was 5.0, the maximum possible value. My notes for both jobs indicated that SMEs were able to understand and perform the steps of contingency development. Qualitatively comparisons of the SME reactions data with my notes for Fire Captain also indicated that SMEs had little to no difficulty carrying out the steps of contingency development. This process was admittedly a little bit more complex for the contingencies that were developed using percentiles rather than the raw scores from the rating scales. Even though it was expected that there would be somewhat of a challenge for SMEs to develop contingencies for such dimensions due to the SMEs' unfamiliarity with *z* scores, results indicated that SMEs quickly comprehended the need for converting the raw scores to *z* scores and using corresponding percentiles as values on the contingency graphs. All five SMEs strongly agreed that members of the group were able to perform each step of contingency development by providing a rating of 5 on the scale. This conclusion was consistent with my notes.

Another aspect of successfully completing each step is the time it took. Including the presentation introducing contingencies (15 minutes) and doing the steps for completing the contingencies, it took the SMEs approximately 2.5 hours to complete the entire process. Approximately half way through developing the contingencies, I provided the SMES with a 15 minute break. Another 30 minutes were dedicated to revisiting the resulting contingencies to be sure the SMEs agreed that the final graphs accurately depicted their views regarding their relationship to expected promotability in comparison to the other dimensions. The SMEs never indicated that time was a problem and seemed completely alert and willing to dedicate the

required time to developing the contingencies. Therefore, my conclusion is that SMEs successfully completed the process in a reasonable amount of time.

The second question is whether there was disagreement at each step, how much, and how was it resolved. Qualitatively comparisons of the SME reactions data with my notes indicated that there was no disagreement at any step of contingency development for Fire Captain. My notes for the position of Fire Lieutenant indicated there was very little disagreement during contingency development. Specifically, there was one instance when developing the contingency for "Fire Behavior Knowledge" at the Fire Lieutenant job level where some disagreement surfaced regarding what values should reflect the minimum acceptable level of performance. This dimension was rated using a 0 to 7 checklist, and SMEs were somewhat torn between assigning the value of 4 or 5 as the minimum acceptable level of performance. To resolve this issue, I simply stood back and allowed the SMEs to engage in discussion supporting their cases for what they felt the value should be. Ultimately, without any intervention, the SMEs listened to each other's reasoning behind their suggestions and agreed on a final value of 5. All five SMEs completing the questionnaire strongly agreed that there was no disagreement during the steps of contingency development. Their mean rating on this item was 5.0, the maximum possible value. This rating was consistent with my notes. It should be noted that the responding SMEs participated in contingency development for Fire Captain, not Fire Lieutenant; therefore, they did not witness the instance of disagreement noted above.

The third question, also supplemented with the 1-5 point rating scale, asked if total consensus was obtained during the process or did SMEs appear to agree just to move along with the process. My notes for both jobs indicated there was consensus among the SMEs. Qualitatively comparisons of the SME reactions data with my notes also indicated there was

consensus obtained throughout the contingency development process (M = 4.8, SD = .45) for Fire Captain. Specifically, four of the five SMEs gave a rating of 5 on the 1-5 point scale the remaining one gave it a 4.

The fourth question, which was also supplemented with the 1-5 point rating scale, assessed whether or not SMEs appeared to be involved during the contingency development process. The mean rating on this item was 5.0, the maximum possible value. My notes for both jobs indicated that SMEs engaged in a high level of involvement throughout the process. Qualitatively comparisons of the SME reactions data with my notes indicated consensus that SMEs appeared to be involved during the process. The SMEs clearly understood the logic of contingencies and their potential advantages. They were very inquisitive about the process itself and how the final results of the process would be used. Several of the SMEs indicated an interest in implementing such a process at their respective fire departments based on the fact that they felt the contingencies were very informative and job-related. Specifically, SMEs were impressed with the feedback that could be provided regarding identifying priorities for improvement. SMEs indicated that this information has a high level of value and would be beneficial to their subordinates who were originally unsuccessful in their performance during a similar AC process. Also, the SMEs inquired about whether I would be willing to send them copies of the final contingencies in addition to a summary of the findings of how this approach compared to the other combination methods. All five SMEs strongly agreed that SMEs appeared involved during contingency development by providing a mean rating of 5.0, which was consistent with my notes.

The fifth question, which was also supplemented with the 1-5 point rating scale, assessed whether the contingencies accurately reflected the targeted position. After completing all steps

of contingency development for both job levels, I revisited all contingencies with the two groups of SMEs in a second session to ensure that the final graphs accurately reflected the SMEs' views of how that particular dimension related to expected promotability in comparison to the other dimensions. My notes for both jobs indicated that the resulting contingencies accurately represented the targeted job. Qualitatively comparisons of the SME reactions data with my notes also indicated that the final contingencies accurately represented the position of Fire Captain. Specifically, three of the five SMEs provided ratings of 5 on the 1-5 point rating scale and two SMEs provided ratings of 4 (M = 4.6, SD = .55).

I made visual inspections of the contingencies to answer the five specific questions about the resulting contingencies to assess whether they appeared to be well-developed. The first question was about differential importance among the contingencies for both job levels. Results indicated that there was clearly differential importance for the various dimensions for both Fire Lieutenant and Fire Captain. As seen in Appendices E and F, which show the final contingencies for both jobs, there are clear differences in the ranges of the contingencies. As noted earlier in the paper, the larger the range of the contingency, the more important that dimension is to expected promotability. For example, the Fire Lieutenant and Fire Captain contingencies show that analytical ability is the most important dimension with the largest range of -100 to +100 at both job levels. Fire behavior knowledge is clearly the least important dimension at the Fire Lieutenant level with the smallest range of -25 to + 45. The ratio of most important to least important for Fire Lieutenant is 2.9 to one. Specifically, most important is 100 + 100 = 200. Least important is 25 +45=70; therefore, the ratio is 200/70, or 2.9 to one. Thus there was considerable differential importance in the contingencies. Conflict management is clearly the least important dimension at the Fire Captain level with the smallest range of -25 to +

25. The ratio of most important to least important for Fire Captain is four to one. Specifically, most important is 100 + 100 = 200. Least important is 25 +25=50; therefore, the ratio is 200/50, or four to one. Thus there was also considerable differential importance in these contingencies as well.

The second question was whether the differential importance from the contingencies related to the job analysis weights given to the dimensions prior to the contingency development by a different set of SMEs. I further investigated instances where the design team realized there were differences in differential importance based on contingencies versus prior job analysis weights, which did they ultimately choose and why. Results indicated that the differential importance from the contingencies was positively related to the job analysis weights given to the dimensions prior to contingency development for Fire Lieutenant (r = .79, p = .00) and for Fire Captain (r = .82, p = .00).

The SMEs doing the contingencies for both Fire Captain and Fire Lieutenant did note that there was some disagreement concerning the weights previously assigned to the various dimensions. For example, as compared to the SMEs doing the earlier weights, SMEs doing the contingencies often had different opinions regarding the importance of a given dimension for successful performance on the job relative to the other dimensions. However, the contingency SMEs for both job levels decided to stick with the importance weights previously assigned to the given dimensions. This decision was made because the SMEs for contingency development decided that the SMEs used to develop weights for the dimensions were recruited for the purpose of developing these weights and were intimately familiar with the ACs and all of its components (e.g., exercises and tasks contained in those ACs). Therefore, they felt this added expertise may have provided the prior SMEs with some additional information that they were unaware of that

influenced the weights they applied to the various dimensions. Therefore, the SMEs for contingency development agreed to defer to the expert opinions of the job analysis weights determined previously due to the more intimate exposure and knowledge the previous SMEs had regarding the AC and its components.

The third and fourth questions assessed if there was differential minimum acceptable performance levels identified for the various dimensions and whether the minimum acceptable performance differed from a rating of "Acceptable". Results indicated that for all of the dimensions assessed using the 1-5 rating scale (e.g., 1 = unacceptable, 3 = acceptable, 5 = outstanding) for both job levels, the minimum acceptable levels of performance were all 3; therefore, there were no differential minimum acceptable levels of performance for these dimensions with no differences from the rating of "Acceptable" on the 5-point scale. However, there were differential minimum acceptable levels of performance for the other dimensions. Specifically, two dimensions for Fire Lieutenant (e.g., incident command and safety and life preservation) both had minimum acceptable levels at the 75^{th} percentile. However, firefighting tactical knowledge and analytical ability both had minimum acceptable levels at the 70^{th} percentile (see Appendix E for the Fire Lieutenant contingencies). For Fire Captain, all contingencies using percentiles had minimum acceptable levels at the 75^{th} percentile except for management ability, which had a minimum acceptable level of performance at the 80^{th} percentile (see Appendix F for the Fire Captain contingencies).

The fifth question examined if there distinct non-linearity in the functions of the final contingency graphs. Results indicated that there was distinct non-linearity for several of the contingencies for both job levels (see Appendices E and F for all final contingencies). A specific example of such non-linearity is observed in Figure 7, which shows that both contingencies have

a critical mass (graph starts out very low in terms of expected promotability and sharply increases) at the bottom of the graph and a diminishing return (graph where a certain point would be reached on the x-axis that would provide no further increases in expected promotability) at the top of the graph. The shapes of these graphs are distinctly non-linear.

Overall, results of all qualitative analyses and visual inspections of the contingencies support hypothesis 1. Specifically, there was general consensus by the SMEs and the researcher concluding a successful development of the contingencies. In addition, the resulting contingencies provided unique information about the AC performance dimensions than would be apparent by examination of the dimension ratings alone. For example, the contingencies reflect evidence that there are differential minimum performance levels across dimensions and differential nonlinearity across dimensions, which is unique information that cannot be determined from the raw dimension ratings.

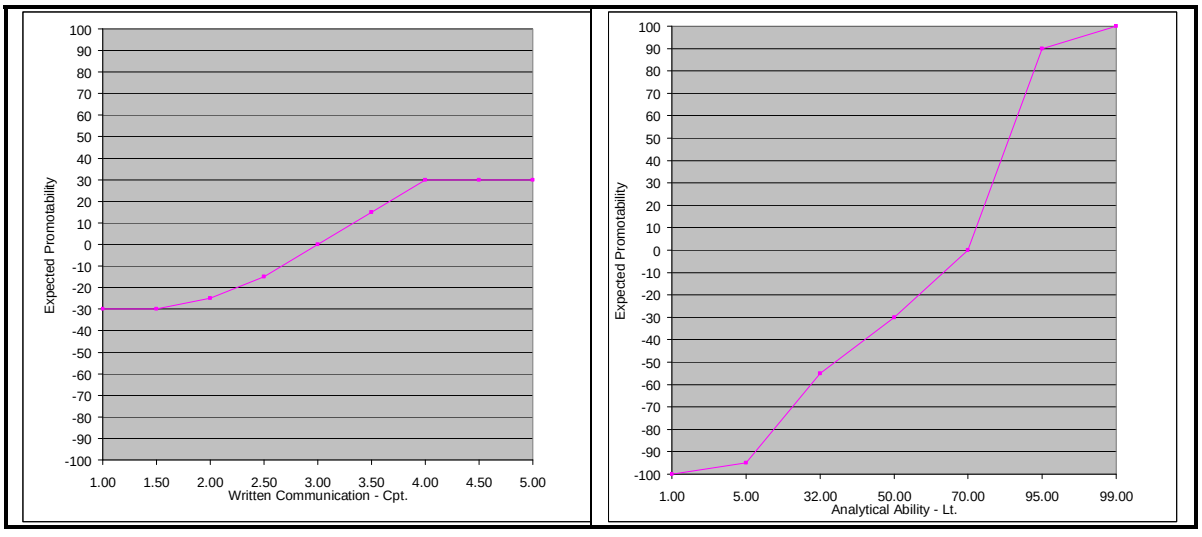


Figure 7 Example Contingencies with Distinct Non-linearities

Hypothesis 2 Results

The second hypothesis stated that contingencies developed for dimensions used more often in the field (e.g., judgment and decision making, analytical ability, incident command, and conflict management) will differ based on the level of the job. No statistical significance tests were used to test this hypothesis. Instead, I conducted visual inspections of the individual contingencies for the eight AC dimensions that were common across both job levels. Job analysts earlier indicated that the Fire Lieutenant job requires more field work than the Fire Captain job, so differences were expected across the dimensions that are used more often in the field (i.e., judgment and decision making, analytical ability, incident command, and conflict management). First, linearity was examined (i.e., linear contingencies for one job level, but non-linear for the other; critical mass for one, but diminishing return for the other). Results indicated that six of the eight common dimensions had similar shapes for their resulting contingencies (see Figure 8). However, it was expected, based on the knowledge of the jobs, that only four of the common dimensions (e.g., leadership ability, supervisory ability, written communication, and oral communication) would have contingencies with similar shapes. As seen in Figure 9 only two of the eight common dimensions had dissimilar contingencies even though it was expected that four dimensions would have contingencies with different shapes. Specifically, there are clear non-linearities in the Fire Lieutenant contingencies for the dimensions of analytical ability and conflict management; however, these contingencies have more of a linear function at the Fire Captain level.

Second, the ranges of the contingencies for each job level were examined to see if they differed, thereby indicating differences in importance. Results indicated that the contingencies for the dimension of conflict management resulted in very dissimilar ranges for the two job

levels further indicating a difference in importance of this dimension at the two job levels. Results further indicated that seven of the eight dimensions that were common across the two job levels had similar ranges (see Figures 8 and 9). However, it was expected, based on the knowledge of the jobs, that only four of the common dimensions (e.g., leadership ability, supervisory ability, written communication, and oral communication) would have similar ranges. This result partially supported hypothesis 2.

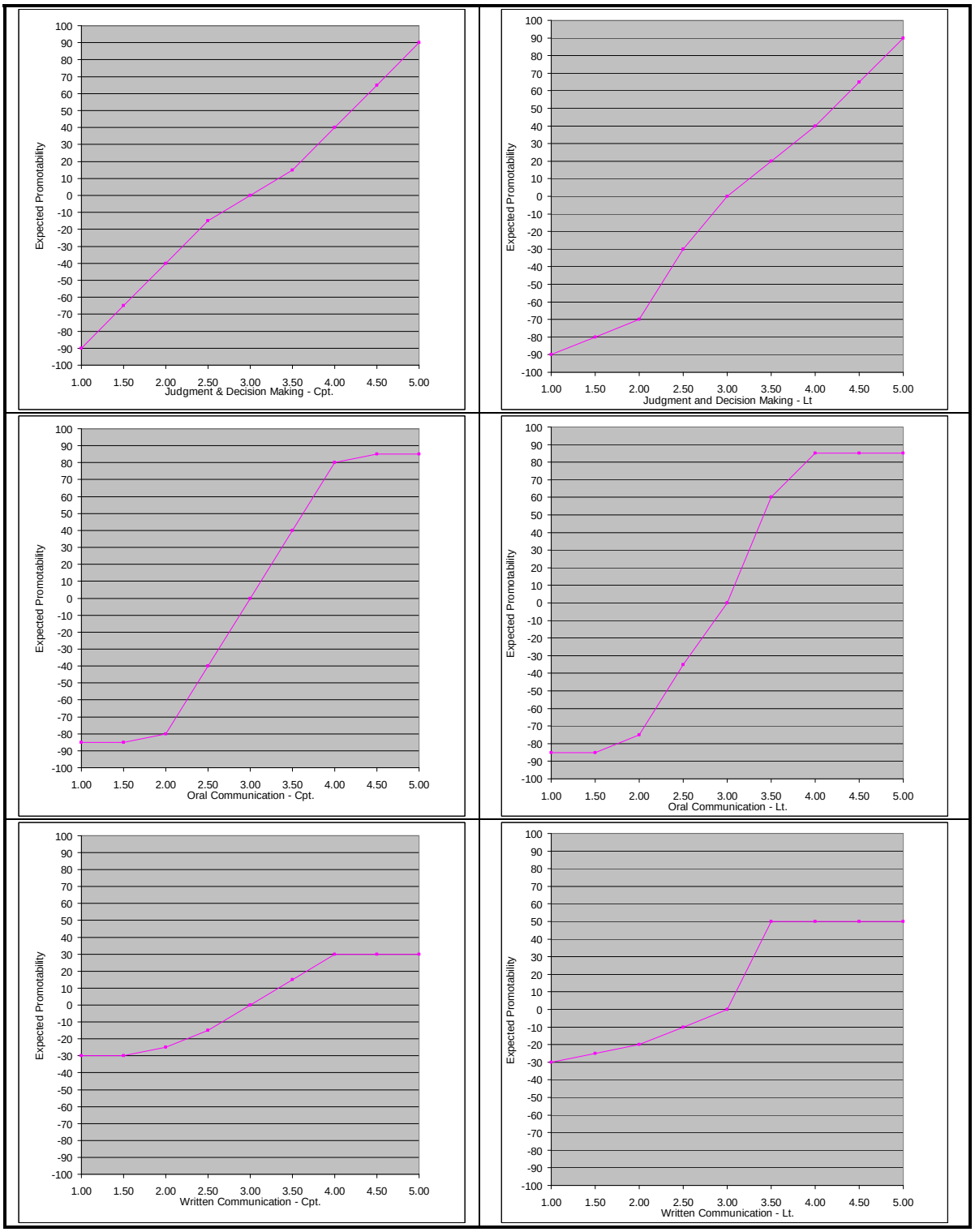


Figure 8 Similar Contingencies for the Two Job Levels

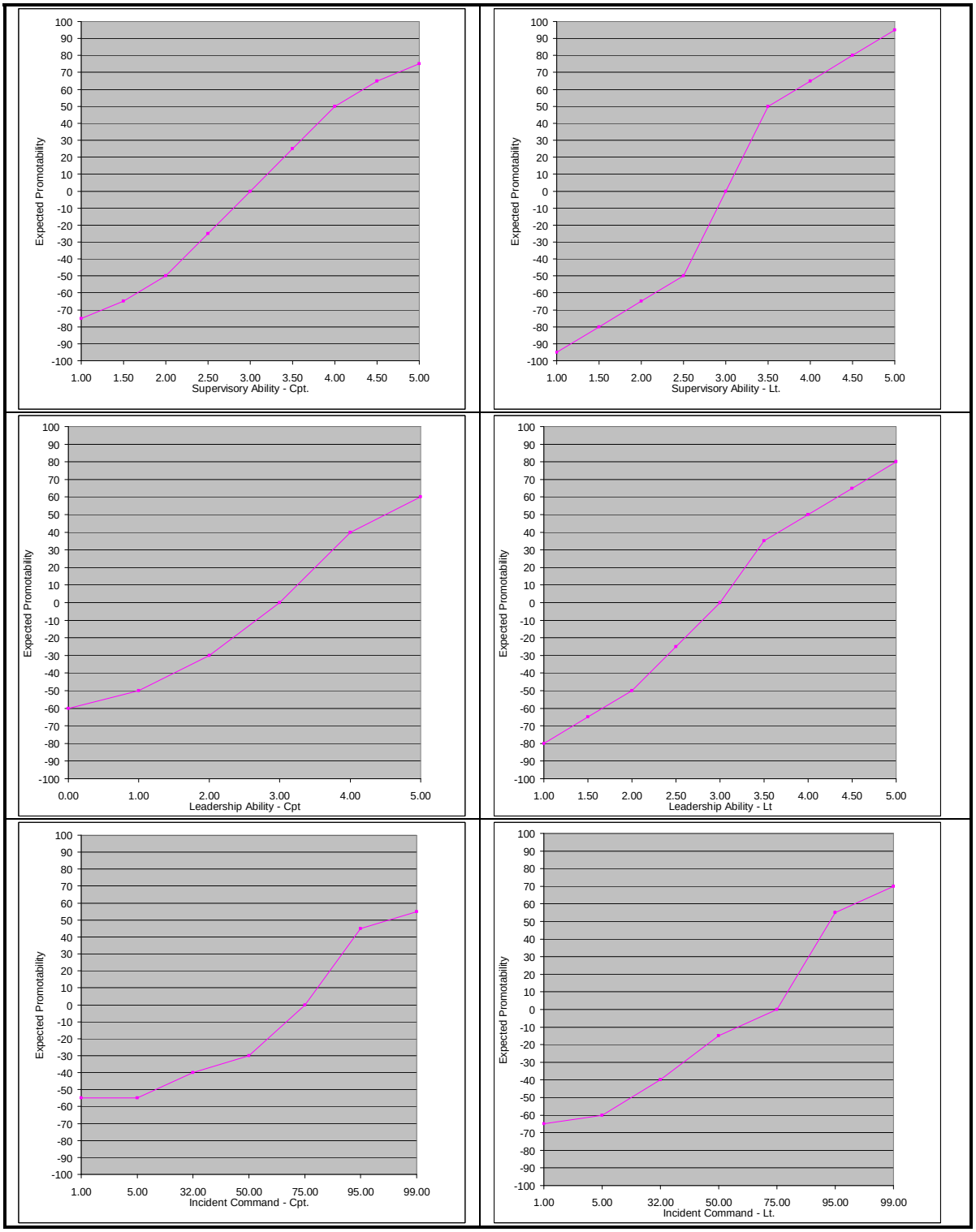


Figure 8 Similar Contingencies for the Two Job Levels

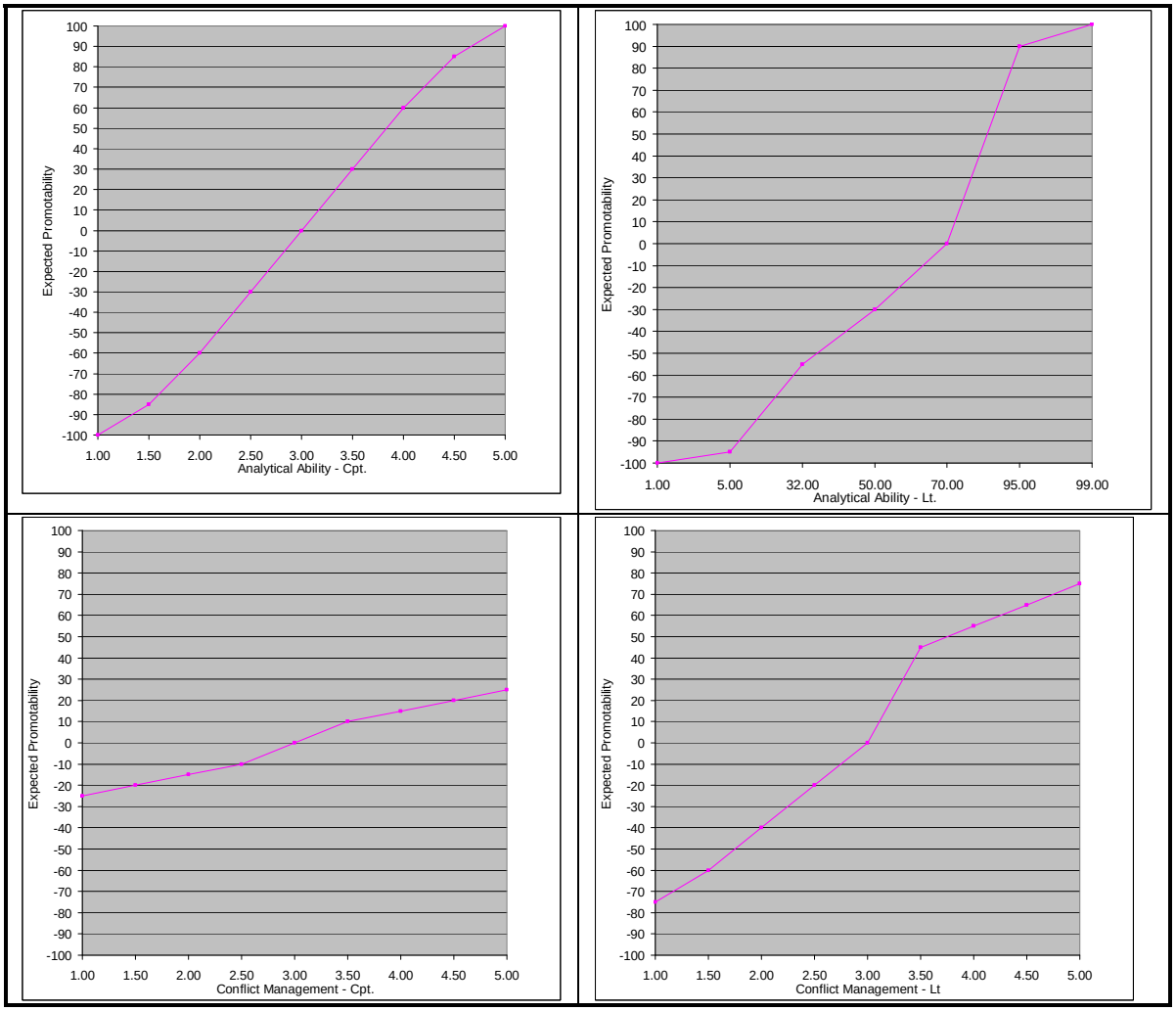


Figure 9 Dissimilar Contingencies for the Two Job Levels

Moreover, a reverse contingency method was implemented. Specifically, for the six dimensions that were common across job levels (i.e., judgment and decision making, oral communication, written communication, incident command, supervisory ability, and conflict management), contingencies developed for Fire Lieutenant were applied to the Fire Captain candidates and contingencies developed for Fire Captain were applied to the Fire Lieutenant candidates (i.e., reversed contingency method). An OAR and rank for the each candidate based on this reversed contingency method using the six available dimensions were also determined.

Subsequently, candidates at the previously specified cut-off ranks for both job levels were assessed to determine the percentage of candidates that was the same when applying this process as compared to the normal contingency data.

As seen in Table 9, results indicated that for the various rank cut-offs at both job levels, there are clear differences in the job candidates falling within the qualifying ranks when applying this reversed contingency method (i.e., applying contingencies for Fire Captain to the Fire Lieutenant candidates and applying the contingencies for Fire Lieutenant to the Fire Captain Candidates). For example, 20% of the candidates for Fire Lieutenant are different at cut-off rank 10 when applying the contingencies developed for Fire Captain in comparison to applying the contingencies developed specifically for Fire Lieutenant to the Fire Lieutenant job candidates. If the contingencies were the same, results would show that the same candidates are selected. However, these results show that different candidates are selected when the reverse contingency method is applied; therefore, the contingencies for the two groups are different from a practical standpoint. On average across all cut-off ranks, 13% of the candidates for Fire Lieutenant and 25% of the candidates for Fire Captain were different based on the reverse contingency approach. The differences in candidates at the various ranks based on the reverse contingency method support the hypothesis that the contingencies developed differ based on job level.

Table 9 Changes in Group Composition Between the Contingency Approach and the Reversed Contingency Approach

Reversed Contingency Method	
Fire Lieutenant (N = 326)	
Top 10	20%
Top 39	5%
Top 48	13%
Fire Captain (N = 118)	
Top 10	30%
Top 22	18%
Top 31	26%

Hypothesis 3 Results

The third hypothesis stated that the final results obtained for the contingency approach are expected to be different than the final results obtained for the traditional approaches in that the people remaining in the top ranks will change depending on the combination method used. To first gain an understanding of how the various combination methods correlated with one another, Pearson *r* correlation coefficients were calculated. Table 10 shows the correlation coefficients among the various combination methods for Fire Lieutenant and Fire Captain. As expected, results indicated that the various combination methods are highly correlated with correlation coefficients ranging from .92 to .99. This high degree of correlation is consistent

with previous findings comparing weighted and unweighted composites (Arthur, Doverspike, & Barrett, 1996; Ree, Carretta, & Earles, 1998; Wilks, 1938).

Table 10 Correlation Coefficients for the Five Methods

Method	JA by Component	Job Analysis by KSA	Unit Weighting by KSA	Unit Weighting by Component	Contingency Method
JA by Component	--	.92	.97	.98	.95
Job Analysis by KSA	.98	--	.97	.94	.96
Unit Weighting by KSA	.99	.99	--	.97	.97
Unit Weighting by Component	.99	.97	.98	--	.96
Contingency Method	.96	.97	.97	.95	--

Note. Correlation coefficients for Fire Lieutenant are presented below the diagonal. Correlation coefficients for Fire Captain are resented above the diagonal.

This hypothesis was then tested by examining the individuals comprising the various rank cut-offs for the different combination methods and assessing the percentage of people who were different when applying the contingency combination method. Support for this hypothesis would be found if there are clear differences in those candidates at the various cut-off ranks, thus indicating that different selection decisions would be made depending on the combination method used.

In support of hypothesis 3, results indicated that for the various rank cut-offs at both job levels, there are clear differences in the job candidates falling within the qualifying ranks when applying the contingency combination method compared with all four traditional combination methods (see Table11). For example, 40% of the candidates for Fire Lieutenant are different at cut-off rank 10 when applying the contingency method in comparison to the unit weighting by component method. Further, an average of 33% of the candidates was different at cut-off rank

10 when the differences for all four methods were averaged. For Fire Captain, results indicated that 16% of the candidates are different at cut-off rank 31 when applying the contingency method in comparison to the job analysis by component method. Further, an average of 14% of the candidates was different at cut-off rank 31 when the differences for all four methods were averaged. Overall, the grand mean, which represents the average across both jobs and all critical cut-offs, indicated that approximately 19% of candidates was different when applying the contingency method in comparison to all other methods.

Table 11 Changes in Group Composition Between Contingencies and Other Methods

	METHOD				
	Job Analysis by Component	Job Analysis by KSA	Unit Weighting by KSA	Unit Weighting by Component	Mean of the Four Methods
Fire Lieutenant (N = 326)					
Top 10 (N = 10)	30%	30%	30%	40%	33%
Top 39 (N = 39)	8%	15%	13%	13%	13%
Top 48 (N = 48)	17%	13%	15%	15%	15%
Mean across Cut-offs	18%	19%	19%	23%	20%
Fire Captain (N = 118)					
Top 10 (N = 10)	10%	10%	20%	30%	18%
Top 22 (N = 22)	23%	14%	18%	18%	18%
Top 31 (N = 31)	16%	10%	10%	19%	14%
Mean across Cut-offs	16%	11%	16%	22%	17%
Mean Across Both Jobs and All Cut-offs					
	17%	15%	18%	23%	19%

Hypothesis 4 Results

Prior to testing the fourth hypothesis stating that scoring the AC with the contingency-based method is expected to result in less adverse impact than scoring the AC using the traditional combination methods, independent samples *t*-tests were first conducted to gain an understanding of group differences based on race for each combination method. Table 12 presents the descriptive statistics by race applying each of the five combination methods for both jobs in addition to mean score differences and effect sizes (*d*). It should be noted that mean score differences were not calculated for sex because of the small number of females (i.e., N=16 for Fire Lieutenant; N=5 for Fire Captain) completing the selection process. Results indicated that mean differences between Blacks and Whites ranged from .22 to 9.67 for the various combination methods, effect sizes for these differences ranged from .03 to .22, but none of these differences were significant based on the t-tests. It is interesting to note that for Fire Lieutenant, Blacks scored higher than Whites in all combinations. For Fire Captain, the reverse was true; Blacks scored lower than Whites in all combinations.

Table 12 Descriptive Statistics and Effect Sizes

	RACE	N	Mean	Std. Deviation	Mean Difference	Pooled SD	*t*-test p-value	*d*
Lieutenant Contingency	W	208	-.67	21.50	-4.42	20.35	.07	-.22
	B	116	3.74	19.13				
Lieutenant Job Analysis by Component	W	208	-2.95	75.25	-9.67	71.72	.25	-.13
	B	116	6.72	68.00				
Lieutenant Job Analysis by KSA	W	208	-2.18	68.58	-7.29	65.21	.34	-.11
	B	116	5.10	61.64				
Lieutenant Unit Weighting by KSA	W	208	-.35	7.92	-1.14	7.57	.20	-.15
	B	116	.79	7.20				
Lieutenant Unit Weighting by Component	W	208	-.28	4.41	-.88	4.23	.08	-.21
	B	116	.60	4.05				
Captain Contingency	W	67	-7.80	15.75	1.82	16.72	.56	.11
	B	48	-9.62	17.62				
Captain Job Analysis by Component	W	67	1.77	65.22	4.66	68.82	.72	.07
	B	48	-2.90	72.23				
Captain Job Analysis by KSA	W	67	3.12	61.75	8.49	64.26	.48	.13
	B	48	-5.37	66.67				
Captain Unit Weighting by KSA	W	67	.03	7.47	.22	7.93	.88	.03
	B	48	-.19	8.36				
Captain Unit Weighting by Component	W	67	.24	5.38	.63	5.57	.55	.11
	B	48	-.39	5.76				

Adverse impact (AI) ratios were then calculated for each combination method based on the 4/5th Rule. This was done at the various cut-off ranks to see if the level of adverse impact differed at each rank (see Tables 13 and 14). If the selection rate for any group is less than 4/5 (or 80%) of the selection rate of the group with the highest selection ratio, this would be considered adverse impact. It should be noted that AI was not calculated for sex because of the small number of females (i.e., N=16 for Fire Lieutenant; N=5 for Fire Captain) completing the selection process. Results indicated that the level of AI differed at each rank for the various methods. The tables show selection ratios and AI statistics rounded to two decimals. For example, in the upper left cells of Table 13, Job Analysis Weighting by Component for Rank 10, the selection ratios for both Blacks and Whites are shown as .03 and the AI statistic is .77. This

is due to rounding. The actual selection ratio was .0259 for Blacks and .0337 for Whites. Dividing the selection ratio for the minority group by the selection ratio of the majority group (.0259/.0337) yielded an AI statistic of .77.

As Table 13 shows, there was no AI for Fire Lieutenant at cut-off rank 39 or cut-off rank 48 for any of the combination methods. For example, at cut-off rank 39 for Fire Lieutenant, the job analysis by component method shows a selection ratio, the proportion of individuals in the top 39, of .11 for Blacks and .13 for whites. Dividing the smaller value of .11 by the larger value of .13 yields an AI statistic of .90, which is well above the .80 level of adverse impact based on the $4/5^{th}$ rule. Therefore, there is no adverse impact. At the most stringent cut-off for Fire Lieutenant (rank 10), 3 of the methods showed adverse impact. Only 2 methods (contingency and unit weighting) did not show adverse impact. One would expect the most adverse impact to occur when standards are highest (i.e., the cut-off is strict), so this is a positive finding for the contingency method. At the most stringent cut-off for Fire Captain (rank 10), all methods showed adverse impact; however, at the next cut-off none of the methods showed adverse impact (AIs ranged from .80 to .97). Thus, there is no advantage of the contingency method there. At the most lenient cut-off (rank 31) for Fire Captain, the unit weighting by component method was the only approach resulting in adverse impact.

The contingency method had no adverse impact at either of the cut-off ranks for Fire Lieutenant, but did show evidence of adverse impact at cut-off rank 10 for Fire Captain. The unit weighting by dimension method was the only other combination method resulting in no adverse impact at either cut-off rank for Fire Lieutenant with adverse impact only observed at cut-off rank 10 for Fire Captain. However, the unit weighting by dimension method had an adverse impact statistic of .48 at cut-off rank 10 for Fire Captain, but the contingency method

resulted in an adverse impact statistic of .72, thereby indicating less adverse impact for the contingency method. Overall, across both jobs and all critical cut-off ranks, the contingency approach resulted in one situation of adverse impact while the other combination methods ranged from one to three in number of situations of adverse impact.

Table 13 Fire Lieutenant Adverse Impact Ratios at the Critical Cut-off Ranks

Cutoff Point	Minority Group	Job Analysis Weighting by Component		Job Analysis Weighting by KSA		Unit Weighting by KSA		Unit Weighting by Component		Contingency	
		Selection Ratio	AI Stat	Selection Ratio	AI Stat	Selection Ratio	AI Stat	Selection Ratio	AI Stat	Selection Ratio	AI Stat
Rank 10	Blacks	.03	.77	.03	.77	.03	.84	.04	.56	.03	.84
	Whites	.03		.03		.03		.02		.03	
Rank 39	Blacks	.11	.90	.12	1.00	.11	.90	.13	.89	.11	.90
	Whites	.13		.12		.13		.12		.13	
Rank 48	Blacks	.15	.98	.16	.93	.15	.98	.16	.85	.16	.93
	Whites	.15		.14		.15		.14		.14	

Note. All Adverse Impact calculations for Fire Lieutenant excluded two individuals reporting races other than Black or White.

Table 14 Fire Captain Adverse Impact Ratios at the Critical Cut-off Ranks

Cutoff Point	Minority Group	Job Analysis Weighting by Component		Job Analysis Weighting by KSA		Unit Weighting by KSA		Unit Weighting by Component		Contingency	
		Selection Ratio	AI Stat	Selection Ratio	AI Stat	Selection Ratio	AI Stat	Selection Ratio	AI Stat	Selection Ratio	AI Stat
Rank 10	Blacks	.10	.72	.10	.72	.13	.48	.10	.72	.10	.72
	Whites	.07		.07		.06		.07		.07	
Rank 22	Blacks	.19	.97	.17	.80	.21	.86	.19	.97	.21	.86
	Whites	.19		.21		.18		.19		.18	
Rank 31	Blacks	.27	.99	.25	.88	.25	.88	.21	.66	.27	.94
	Whites	.27		.28		.28		.31		.25	

Note. All Adverse Impact calculations for Fire Captain excluded three individuals reporting races other than Black or White.

Due to the fact that the data yielded an atypical level of adverse impact (i.e., a lower level for all of the combination methods than is normally seen), it is worthwhile to examine adverse impact at other cut-off points within the range of cut-offs used. A 2x5 one-way analysis of variance (ANOVA) was performed on the AI statistics at all ranks within the range of cut-off used to examine if there were significant differences in the mean AI statistics between any of the five combination methods. Results indicated that the job analysis by KSA method was

significantly different than all other combination methods for Fire Lieutenant. The unit weighting by KSA method was significantly different than all other combination methods for Fire Captain. There were no other significant mean differences in the AI statistics between any of the other methods.

Table 15 provides a summary of adverse impact (AI) statistics for each method at all ranks by indicating the percentage of time the contingency method had a higher (i.e., less likely to show adverse impact), equal (i.e., equally likely to show adverse impact), or lower (i.e., more likely to show adverse impact) AI statistic than the other methods. For example, there were a total of 39 ranks observed for Fire Lieutenant. The contingency method had a higher AI statistic than the job analysis by KSA method at 26 of those 39 ranks. Therefore, the contingency method had a higher AI statistic (i.e., less likely to show adverse impact) than the job analysis by KSA method in 67% of the cases for Fire Lieutenant. Also seen in Table 15, the contingency method had an equal or higher AI statistic in 71% of the cases over the other four methods For Fire Lieutenant and 84% of the cases for Fire Captain. The results clearly supports that the contingency method had higher overall AI statistics than the other methods, meaning that it was less likely to show adverse impact because the higher the AI statistic, the better. The contingency method only exhibited lower AI statistics (i.e., more likely to show adverse impact) than the other methods in 29% of the cases for Fire Lieutenant and 16% of the cases for Fire Captain.

Table 15 Summary of Adverse Impact Statistic Comparisons for All Ranks

	METHOD				
	Job Analysis by Component	Job Analysis by KSA	Unit Weighting by KSA	Unit Weighting by Component	Mean of the Four Methods
Fire Lieutenant (N Cut-offs = 39)					
Higher	41%	67%	33%	44%	46%
Equal	31%	15%	38%	15%	25%
Lower	28%	18%	29%	41%	29%
Fire Captain (N Cut-Offs = 22)					
Higher	23%	50%	73%	45%	48%
Equal	50%	36%	18%	41%	36%
Lower	27%	14%	9%	14%	16%

In addition to the inability to control for either Type I or Type II errors, the 4/5th rule of adverse impact is subject to considerable sampling error, especially when dealing with small sample sizes (Lawshe, 1987; Morris, 2001; Roth, Bobko, & Switzer, 2006). For example, the 4/5th rule may result in adverse impact even when it is not statistically and/or practically significant if the sample size is small. However, the 4/5th rule may fail to show adverse impact even when differences are statistically and/or practically significant. To overcome part of this obstacle and control Type I error by setting the alpha level at .05, I conducted two statistical tests, which are both mathematically equivalent to those recommended by the OFCCP (1993) to estimate adverse impact. These tests include the Pearson chi-square test of association and Fisher's exact test. Both tests were designed to test the hypothesis that there are significant differences in the pass rate of two groups. It should be noted that the power of these tests are

dependent upon sample size and selection rate. The independent variable in this study is race (Black or White), and the dependent variable is passing the cut-off. Due to the fact that the dependent variable is dichotomous, either of these tests is appropriate to test for significant differences.

The chi-square test estimates the probability of the sample results based on the association between group membership and test outcome by comparing the fit between observed frequencies and expected frequencies. In this study chi-square cut-offs for pass/fail were dependent upon the cut-off rank. For example, if chi-square is conducted at rank 39, all candidates receiving ranks 1-39 are categorized as passing. All candidates receiving any rank other than 1-39 are categorized as failing. Chi-square then tests whether the differences in the pass rate for Blacks and Whites are significant. If the resulting *p*-value from the chi-square test is less than .05 (i.e., the alpha value) then the pass/fail difference between the groups is statistically significant. The chi-square probability does approach the exact probability as the sample size increases (Hays, 1994). The chi-square probability statistic should not be used if the minimum expected frequency is less than 5 (Moore & McCabe, 1993) and should be interpreted cautiously when the minimum expected frequency is less than 10 (Hays, 1994).

Fisher's exact test examines the same hypothesis as chi-square. However, unlike chi-square there are no concerns about interpreting the results of Fisher's exact when there is minimum expected frequency less than 10, and Fisher's exact provides the exact probability of the sample results rather than an estimate. If the resulting *p*-value from Fisher's exact test is less than .05 (i.e., the alpha value) then the pass/fail difference between the groups is statistically significant. If the resulting *p*-value from Fisher's exact test is greater than .05 (i.e., the alpha value) then the pass/fail difference between the groups is not statistically significant. Although

the OFCCP (1993) recommends using Fisher's exact test when sample sizes are small, this test is appropriate for all sample sizes if the statistical software that is used allows for the calculation.

Results indicated there were no significant chi-square values for any of the methods for Fire Lieutenant or Fire Captain (see Tables 16 through 45) nor was the Fisher exact test significant for any of the methods. More specifically, these results indicated that there were no significant differences between the pass rate for Blacks and Whites across all ranks for all combination methods. Therefore, these results suggest that there is no significant adverse impact for any of the methods at any of the cut-off ranks. However, a lack of statistically significant differences between pass rates for Blacks and Whites does not mean that there is no practical significance of the levels of adverse impact shown with the $4/5^{th}$ rule.

The examination of AI statistics and the significance tests together suggests that while there were no statistically significant differences in pass rates for Blacks and Whites, the contingency approach was superior in terms of degree of adverse impact. It is worth noting that there seems to be a contradiction between the *d* statistic results and the AI results. Specifically, although the mean score differences between Blacks and Whites were not statistically significant, the largest effect size resulted with the contingency method. This finding may lead to the thought that the contingency method would then exhibit more adverse impact based on the largest effect size for group differences; however, that is not the case. The *d* statistic is not a measure of adverse impact and it takes into account overall mean group differences for a given method whereas the AI statistics focus on a given method at the specified rank cut-off. Therefore, it is understandable how the contingency method is still able to produce less overall adverse impact than the other methods despite having the larger effect size.

Table 16 Fire Lieutenant Statistical Significance Tests for the Contingency Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	6	4	10	Impact Ratio	.84	N/A	N/A
Failed Test	202	112	314	Fisher exact	N/A	.75	No
Total	208	116	324				
Pass Rate	.03	.03	.03				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 17 Fire Lieutenant Statistical Significance Tests for the Contingency Method at Rank 39

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	26	13	39	Impact Ratio	.90	N/A	N/A
Failed Test	182	103	285	Fisher exact	N/A	.86	No
Total	208	116	324	Chi-Square	.12	.73	No
Pass Rate	.13	.11	.12				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 18 Fire Lieutenant Statistical Significance Tests for the Contingency Method at Rank 48

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	30	18	48	Impact Ratio	.93	N/A	N/A
Failed Test	178	98	276	Fisher exact	N/A	.87	No
Total	208	116	324	Chi-square	.07	.79	No
Pass Rate	.14	.16	.15				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 19 Fire Lieutenant Statistical Significance Tests for the Job Analysis by Component Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	7	3	10	Impact Ratio	.77	N/A	N/A
Failed Test	201	113	314	Fisher exact	N/A	1.00	No
Total	208	116	324				
Pass Rate	.03	.03	.03				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 20 Fire Lieutenant Statistical Significance Tests for the Job Analysis by Component Method at Rank 39

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	26	13	39	Impact Ratio	.90	N/A	N/A
Failed Test	182	103	285	Fisher exact	N/A	.86	No
Total	208	116	324	Chi-Square	.12	.73	No
Pass Rate	.13	.11	.12				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 21 Fire Lieutenant Statistical Significance Tests for the Job Analysis by Component Method at Rank 48

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	31	17	48	Impact Ratio	.98	N/A	N/A
Failed Test	177	99	276	Fisher exact	N/A	1.00	No
Total	208	116	324	Chi-Square	.00	.95	No
Pass Rate	.15	.15	.15				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 22 Fire Lieutenant Statistical Significance Tests for the Job Analysis by KSA Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	7	3	10	Impact Ratio	.77	N/A	N/A
Failed Test	201	113	314	Fisher exact	N/A	1.00	No
Total	208	116	324				
Pass Rate	.03	.03	.03				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 23 Fire Lieutenant Statistical Significance Tests for the Job Analysis by KSA Method at Rank 39

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	25	14	39	Impact Ratio	1.00	N/A	N/A
Failed Test	183	102	285	Fisher exact	N/A	1.00	No
Total	208	116	324	Chi-Square	.00	.99	No
Pass Rate	.12	.12	.12				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 24 Fire Lieutenant Statistical Significance Tests for the Job Analysis by KSA Method at Rank 48

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	30	18	48	Impact Ratio	.93	N/A	N/A
Failed Test	178	98	276	Fisher exact	N/A	0.87	No
Total	208	116	324	Chi-square	.07	0.79	No
Pass Rate	.14	.16	.15				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 25 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	6	4	10	Impact Ratio	.84	N/A	N/A
Failed Test	202	112	314	Fisher exact	N/A	0.75	No
Total	208	116	324				
Pass Rate	.03	.03	.03				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 26 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 39

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	26	13	39	Impact Ratio	.90	N/A	N/A
Failed Test	182	103	285	Fisher exact	N/A	.86	No
Total	208	116	324	Chi-Square	.12	.73	No
Pass Rate	.13	.11	.12				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 27 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 48

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	31	17	48	Impact Ratio	.98	N/A	N/A
Failed Test	177	99	276	Fisher exact	N/A	1.00	No
Total	208	116	324	Chi-Square	.00	.95	No
Pass Rate	.15	.15	.15				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 28 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by Component Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	5	5	10	Impact Ratio	.56	N/A	N/A
Failed Test	203	111	314	Fisher exact	N/A	.34	No
Total	208	116	324				
Pass Rate	.02	.04	.03				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 29 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by Component Method at Rank 39

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	24	15	39	Impact Ratio	.89	N/A	N/A
Failed Test	184	101	285	Fisher exact	N/A	.72	No
Total	208	116	324	Chi-square	.14	.71	No
Pass Rate	.12	.13	.12				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 30 Fire Lieutenant Statistical Significance Tests for the Unit Weighting by Component Method at Rank 48

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	29	19	48	Impact Ratio	0.85	N/A	N/A
Failed Test	179	97	276	Fisher exact	N/A	.63	No
Total	208	116	324	Chi-square	.35	.55	No
Pass Rate	.14	.16	.15				

Note. The table does not include two individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 31 Fire Captain Statistical Significance Tests for the Contingency Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	5	5	10	Impact Ratio	.72	N/A	N/A
Failed Test	62	43	105	Fisher exact	N/A	0.74	No
Total	67	48	115				
Pass Rate	.07	.10	.09				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 32 Fire Captain Statistical Significance Tests for the Contingency Method at Rank 22

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	12	10	22	Impact Ratio	.86	N/A	N/A
Failed Test	55	38	93	Fisher exact	N/A	.81	No
Total	67	48	115	Chi-Square	.15	.69	No
Pass Rate	.18	.21	.19				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 33 Fire Captain Statistical Significance Tests for the Contingency Method at Rank 31

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	17	13	30	Impact Ratio	.94	N/A	N/A
Failed Test	50	35	85	Fisher exact	N/A	.83	No
Total	67	48	115	Chi-Square	.04	.84	No
Pass Rate	.25	.27	.26				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 34 Fire Captain Statistical Significance Tests for the Job Analysis by Component Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	5	5	10	Impact Ratio	.72	N/A	N/A
Failed Test	62	43	105	Fisher exact	N/A	.74	No
Total	67	48	115				
Pass Rate	.07	.10	.09				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 35 Fire Captain Statistical Significance Tests for the Job Analysis by Component Method at Rank 22

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	13	9	22	Impact Ratio	.97	N/A	N/A
Failed Test	54	39	93	Fisher exact	N/A	1.00	No
Total	67	48	115	Chi-square	.01	.93	No
Pass Rate	.19	.19	.19				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 36 Fire Captain Statistical Significance Tests for the Job Analysis by Component Method at Rank 31

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	18	13	31	Impact Ratio	.99	N/A	N/A
Failed Test	49	35	84	Fisher exact	N/A	1.00	No
Total	67	48	115	Chi-Square	.00	.98	No
Pass Rate	.27	.27	.27				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 37 Fire Captain Statistical Significance Tests for the Job Analysis by KSA Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	5	5	10	Impact Ratio	.72	N/A	N/A
Failed Test	62	43	105	Fisher exact	N/A	.74	No
Total	67	48	115				
Pass Rate	.07	.10	.09				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 38 Fire Captain Statistical Significance Tests for the Job Analysis by KSA Method at Rank 22

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	14	8	22	Impact Ratio	.80	N/A	N/A
Failed Test	53	40	93	Fisher exact	N/A	.64	No
Total	67	48	115	Chi-square	.32	.57	No
Pass Rate	.21	.17	.19				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 39 Fire Captain Statistical Significance Tests for the Job Analysis by KSA Method at Rank 31

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	19	12	31	Impact Ratio	.88	N/A	N/A
Failed Test	48	36	84	Fisher exact	N/A	.83	No
Total	67	48	115	Chi-square	.16	.69	No
Pass Rate	.28	.25	.27				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 40 Fire Captain Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	4	6	10	Impact Ratio	.48	N/A	N/A
Failed Test	63	42	105	Fisher exact	N/A	.32	No
Total	67	48	115				
Pass Rate	.06	.13	.09				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 41 Fire Captain Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 22

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	12	10	22	Impact Ratio	.86	N/A	N/A
Failed Test	55	38	93	Fisher exact	N/A	.81	No
Total	67	48	115	Chi-Square	.15	.69	No
Pass Rate	.18	.21	.19				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 42 Fire Captain Statistical Significance Tests for the Unit Weighting by KSA Method at Rank 31

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	19	12	31	Impact Ratio	.88	N/A	N/A
Failed Test	48	36	84	Fisher exact	N/A	.83	No
Total	67	48	115	Chi-square	.16	.69	No
Pass Rate	.28	.25	.27				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 43 Fire Captain Statistical Significance Tests for the Unit Weighting by Component Method at Rank 10

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	5	5	10	Impact Ratio	.72	N/A	N/A
Failed Test	62	43	105	Fisher exact	N/A	.74	No
Total	67	48	115				
Pass Rate	.07	.10	.09				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value. Chi-square analysis was not recommended based on expected frequencies less than 5.

Table 44 Fire Captain Statistical Significance Tests for the Unit Weighting by Component Method at Rank 22

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	13	9	22	Impact Ratio	.97	N/A	N/A
Failed Test	54	39	93	Fisher exact	N/A	1.00	No
Total	67	48	115	Chi-square	.01	.93	No
Pass Rate	.19	.19	.19				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Table 45 Fire Captain Statistical Significance Tests for the Unit Weighting by Component Method at Rank 31

	Sample Frequencies			Adverse Impact Test Results			
	White	Black	Total	Test type	Test Value	*p*-value	Significant?
Passed Test	21	10	31	Impact Ratio	.66	N/A	N/A
Failed Test	46	38	84	Fisher exact	N/A	.29	No
Total	67	48	115	Chi-square	1.57	.21	No
Pass Rate	.31	.21	.27				

Note. The table does not include three individuals reporting races other than Black or White. The impact ratio does not produce a p-value or significance result because it is not a statistical test. Fisher's exact test does not produce a test value.

Finally, in an attempt to understand why the contingency approach did have less adverse impact overall than the other combination methods based on the 4/5th rule, I examined the contingencies of the dimensions requiring higher levels of cognitive ability to see if they had non-linear shapes. The rationale was that if the cognitively loaded dimensions had non-linear shapes (e.g., diminishing return), the lower levels of adverse impact may be due to minorities scoring lower on those particular dimensions, which are known to produce more adverse impact, without lowering their overall expected promotability score because they performed at least at the level providing the maximum gain in expected promotability. For example (see Figure 8), if written communication was deemed a cognitively loaded dimension, and minority candidates are likely to perform lower overall on cognitively loaded dimensions, this dimension is likely to produce adverse impact. Thus, minority OARs would be impacted by their lower performance on this dimension and other cognitively loaded dimensions. However, based on the contingency method, as long as the candidate scored at the level of 4 on the written communication dimension, they would receive the maximum gain in expected promotability for this dimension even if they are performing lower than majority group members. This logic would carry over to all other cognitively loaded dimensions with non-linear functions and ultimately result in lower levels of adverse impact.

To see if this was a plausible explanation for the lower levels of adverse impact with the contingency method, a determination was made about which dimensions were cognitively loaded and which were not. This determination was made using two SMEs. They first provided independent judgments regarding what dimensions required higher levels of cognitive ability and with discussion reached a final consensus with total agreement on which dimensions were cognitively loaded. The dimensions of policies and procedures, firefighting tactical knowledge,

fire behavior knowledge, analytical ability, and judgment and decision making were considered cognitively loaded based on the assumption that they rely more heavily on cognitive ability skills. Results indicated that although neither of the two cognitively loaded dimensions for Fire Captain had non-linear contingencies, four of the five cognitively loaded dimensions for Fire Lieutenant had non-linear contingencies (see Figures 10 and 11). This result supports the suggested explanation that the non-linear contingencies for four of the five cognitively loaded dimensions for Fire Lieutenant may have played a pivotal role in the lower adverse impact observed.

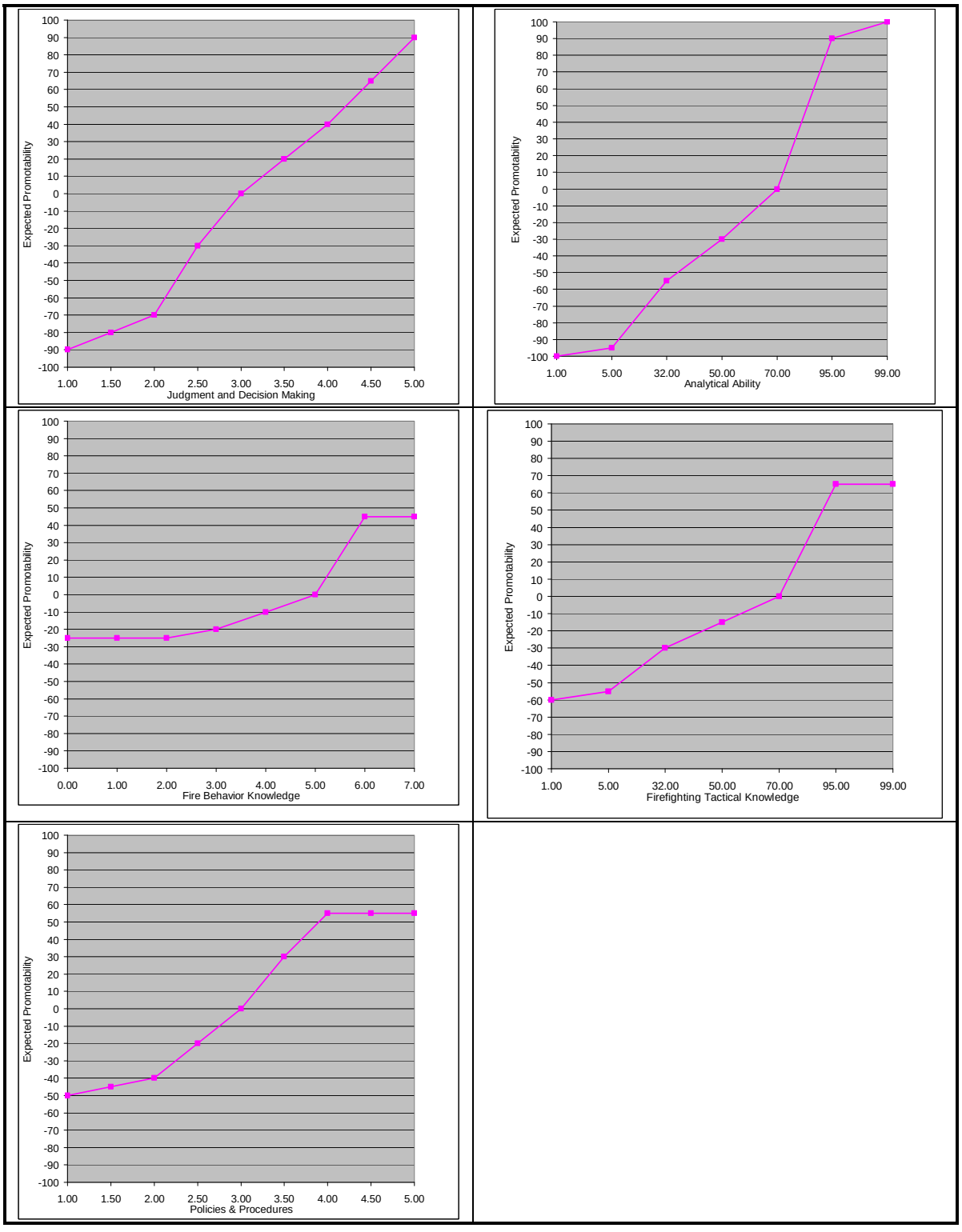


Figure 10 Fire Lieutenant Cognitively Loaded Dimensions

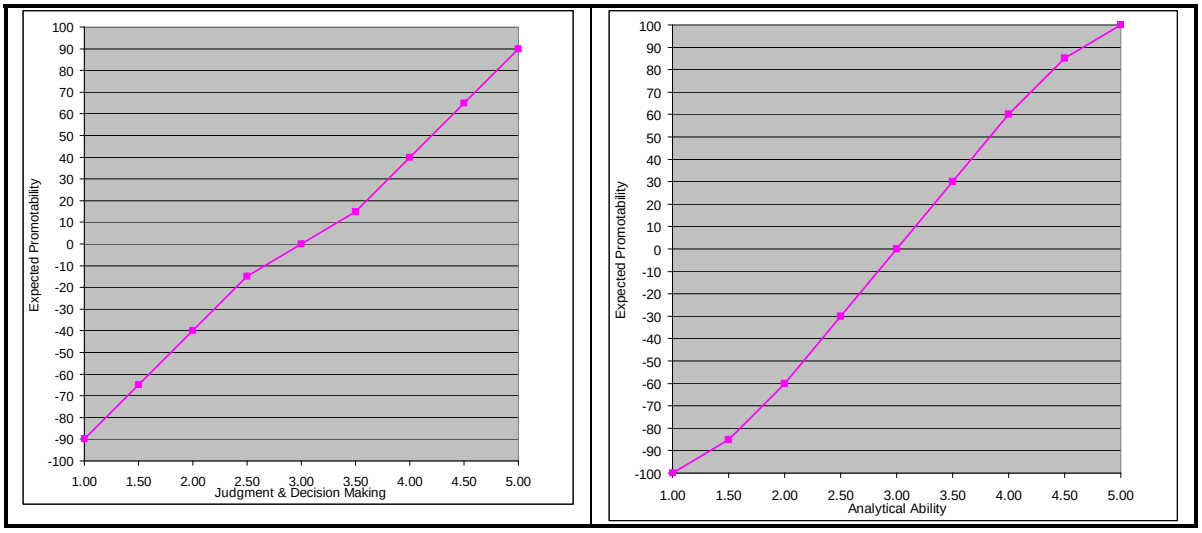


Figure 11 Fire Captain Cognitively Loaded Dimensions

CHAPTER FIVE: CONCLUSION

Discussion

This study examined the viability and potential usefulness of applying the contingency approach from the organizational productivity literature to assessment center scoring. The first hypothesis examined whether SMEs would be able to successfully develop contingencies. There was a high level of agreement between my notes and ratings provided by the SMEs indicating that SMEs performed each step of contingency development and reached total consensus throughout the process with little, if any, difficulty. Further, SMEs were highly involved throughout the process and agreed that the resulting contingencies provided accurate representations of the targeted position. In addition, the contingencies reflected clear differential importance for the various dimensions for both jobs with distinct non-linearity for several of the contingencies. Based on these findings, I conclude that SMEs were able to successfully develop the contingencies.

The second hypothesis examined whether the contingencies developed for a given dimension would differ based on the level of the job. Visual inspections of the graphs showed that the only two of the eight dimensions that were common across both jobs had dissimilar contingencies (e.g., linear from one job and non-linear for the other job) in addition to dissimilar ranges. Furthermore, the reverse contingency approach (i.e., contingencies developed for Fire Lieutenant were applied to the Fire Captain candidates and contingencies developed for Fire Captain were applied to the Fire Lieutenant candidates) showed clear differences in the job candidates falling within the qualifying ranks; thus, indicating there were meaningful differences

in the contingencies developed for the two jobs. Based on these findings, I conclude the contingency approach led to different sets of contingencies for the two jobs.

The third hypothesis examined whether there will be different people in the top ranks if the different combination methods are used to calculate final scores. When the contingency approach was used to calculate OARs, different applicants made the eligible cut-off ranks than when other mechanical combination methods were used to calculate OARs. Specifically, approximately 19% of candidates across both jobs and all critical cut-offs was different when applying the contingency method in comparison to all other methods. Based on this finding I conclude that people remaining in the top ranks will change and different selection decisions will be made when different combination methods are used. Thus, the contingency approach is providing different information than the other methods.

The fourth hypothesis examined whether scoring the AC with the contingency-based method will result in less adverse impact than the traditional combination methods. An important first observation is that there was much less adverse impact in this study with all traditional combination methods than typically observed. This makes it difficult to observe any further reduction in adverse impact based on the contingency method due to the fact that adverse impact is already low. In this study across both jobs and all critical cut-off ranks, the contingency approach resulted in one situation of adverse impact while the other combination methods ranged from one to three in number of situations of adverse impact. Moreover, the contingency method still had much less adverse impact than the only other combination method that also resulted in only one situation of adverse impact. I also found that the contingency method was equally or less likely to show adverse impact in 71% of the cases for Fire Lieutenant and 84% of the cases for Fire Captain. Based on these findings I conclude that there is a

consistent pattern showing the contingency approach is less likely to show adverse impact. Despite a lack of statistically significant differences in the pass/fail rate of Blacks and Whites across the various combination methods and critical cut-off ranks, there was adverse impact based on the 4/5th rule, which indicates practically significant differences in the pass/fail rates of the two groups.

There are a number of reasons why these results are interesting. First, the use of the contingency approach has been extended beyond its original purpose in ProMES (e.g., providing feedback to increase productivity) and utilized in an operational AC as a scoring technique. Not only was it interesting to see that this could be done, it is of even more value to realize that it could be implemented successfully and produce unique information beyond that already gained by other scoring methods.

Contingencies also have other advantages. Contingencies clearly identify the minimum expected level for each dimension assessed in addition to highlighting the extent to which each level of the dimension is good or bad. This information is not obtained with any of the other scoring methods; however, this information can be quite useful from a research and practical standpoint. Researchers can take this information and expand upon it to see if the information provided can be used to better predict candidates who are successful. For example, researchers can test whether candidates performing at, above, or below minimum expectations in this AC translate into employees who perform at, above, or below minimum expectations respectively on the job.

Another advantage is the application of contingencies for different job levels within a job family, which reflect differences in the importance of each dimension. The key issue is that the same assessment center can be used for different jobs without doing a new assessment center.

The idea is to use the contingencies to capture differential importance of the various dimensions at each job level when scoring the AC data. For example, if we assume that managerial jobs have largely the same dimensions of performance such as planning, budgeting, technical knowledge, subordinate development, etc., one AC can be developed for these jobs in general (this is what most consulting firms do) and the contingencies can be used to customize that AC for the different types of jobs and for the specific priorities of different organizations. This has tremendous cost saving benefits.

Another advantage of contingencies is their ability to identify non-linearities, which are not captured by the other scoring methods. Identifying these non-linearities and taking them into account when scoring the AC data increases the job-relatedness of the contingency approach. For example, once you have identified the point at which any further increase in performance on a given dimension fails to produce any increase in expected promotability, this information is used in the scoring process. Specifically, there would be no increase in a candidate's overall expected promotability score for performing beyond the level providing the maximum gain in expected promotability on a given dimension. This process increases the job-relatedness of the contingency approach due to the fact that candidates are not rewarded or penalized for levels of performance that are not required to obtain the maximum expected promotability score.

Examining non-linearity also provides the advantage of identifying priorities for improvement. Priorities here means identifying which dimensions one should focus on if looking to make the greatest increase in overall expected promotability. For example, the first step would be to identify the levels of performance on each dimension. The next step would be to determine the expected promotability score associated with those levels of performance. The final step would be to look at the shape of the graph and calculate expected gain in expected

promotability at various levels of improvement on the different dimensions. The idea is to focus attention on those dimensions that would result in the greatest gain in overall expected promotability. The non-linearities will help to clarify dimensions that should or should not be focused on given the individual's current level of performance on that particular dimension. This is of value because it is a very beneficial method to avoid wasting time focusing attention on areas that will not increase the OAR. For example, if a candidate did not qualify for the eligible ranks, this information could be used to help the candidate make the most effective decisions regarding where attention should be focused to improve the OAR if the candidate goes through the AC again.

The contingency approach also resulted in less adverse impact overall at the critical cut-off ranks with better adverse impact statistics overall at all ranks, not just the critical cut-offs. This can probably be explained by the job-relatedness of the contingencies. Specifically, the contingency approach directly links to important aspects of the job without overweighting cognitively loaded dimensions beyond the level of performance actually needed on that dimension to reap the maximum gain in expected promotability. It is worth noting that unlike the level of Fire Captain, the contingency method did not yield any adverse impact at the level of Fire Lieutenant. This may be due to the fact that five of the six cognitively loaded dimensions for Fire Lieutenant had non-linear contingencies. Consequently, minorities scoring lower on those particular dimensions, which are known to produce more adverse impact, would not have been penalized on their overall expected promotability score if they performed at least at the level providing the maximum gain in expected promotability. Therefore, overall scores for minorities would have been comparable to their majority member counterparts even if the

majority group members performed better on the cognitively loaded dimensions. This would have definitely resulted in less adverse impact for the contingency method.

Practical Implications

A major practical implication of these finding is that practitioners should be mindful of not only the manner in which ACs are developed, conducted, and scored, but also how the resulting data are combined to make selection and promotion decisions. It is important for to keep in mind the scoring technique used, regardless of its high correlation with other scoring methods, results in the selection of different candidates with different levels of adverse impact shown for the various methods. Implementing a technique that can reduce adverse impact in any way has practical significance for ACs based on the large number of individuals affected by even a small amount of adverse impact. This also benefits the organization from a legal standpoint, such that using the contingency approach is a job-related technique that is legally defensible and minimizes group differences.

Another practical implication is that after identifying cognitively loaded versus non-cognitively loaded dimensions assessed in the AC, contingencies can be utilized to increase the job-relatedness of the scoring process. Specifically, the contingencies could be used to identify what level of performance is needed for the cognitively loaded dimensions to provide the maximum gain in the overall criterion. At this point practitioners have the ability to utilize this information in a manner that rewards candidates for levels of performance that are required to obtain the maximum overall score and not beyond that level that is not related to the criterion of

interest. As stated previously, this process would carry over to all cognitively loaded dimensions with non-linear functions and ultimately result in lower levels of adverse impact.

Limitations and Future Research

One limitation of this study was the small number of candidates at each cut-off rank (i.e., sampling error), which made it difficult to determine if the differences in candidates at the various cut-off ranks were true differences or simply differences that would have ceased with a larger number of candidates. This may be the reason for mixed findings regarding second hypothesis that contingencies would differ based on the job level. Future researchers should determine if this finding would hold in a situation where there are larger numbers of candidates at the various cut-off ranks.

Another limitation of this study was having SMEs for contingency development that differed from the SMEs developing the job analysis weights for the dimensions of the AC. As stated previously, this resulted in some disagreement regarding the importance of each dimension. To facilitate this process in the future, it would be better to utilize the same SMEs for contingency development as the SMEs used for development of the AC components, tasks, and job analysis weights. This would avoid any confusion and disagreement regarding previously assigned job analysis weights and the need to determine what weights should be utilized. If different SMEs are used, future researchers may also want to replicate this study without providing job analysis weights to the SMEs ahead of time to see what differences would exist in the final contingencies. Future researchers should also use outside observers to provide ratings of the ability of SMEs to successfully develop the contingencies in addition to ratings

from all SMEs involved in the contingency development process. Using the approach would allow for more quantitative analyses in addition to the objectivity of an outside expert with no prior knowledge of the hypotheses for the study. Future researchers may want to compare the contingency approach with not only other statistical approaches, but clinical approaches as well.

This study has established that contingencies provide a viable option for calculating OARs in AC contexts, and now an important next step is to examine whether OARs developed with contingencies are more highly related to performance criteria than are other methods. Having this information would allow for much more powerful conclusions. As stated previously, even though the various methods resulted in different decisions being made from a selection standpoint, I am unable to say if any one decision is better than the other in its ability to predict performance on the job without criterion data. Future researchers should conduct a criterion-related validity study to determine what method better predicts job performance of the candidates at the various cut-off ranks. Specifically, further research is needed to test whether these people selected with contingencies are actually higher performers. At that point, conclusions can be drawn regarding which method is recommended over the others.

Conclusion

In conclusion, ACs affect many people in our society, so even relatively small amounts of adverse impact can be important. Many organizations suffer major lawsuits based on the use of selection instruments that adversely impact protected groups when there are other selection tools with less adverse impact that can be implemented. Although there were some instances where another method may have resulted in less adverse impact than the contingency method at a given

cut-off rank, there was less adverse impact overall with the contingency method. Any contribution that can be made to reduce this discrimination is quite valuable. The effects of less adverse impact start with the individual AC candidates and extend to the organizations as well as society as a whole given that AC are often used in large-scale testing programs for many civil services jobs across the world. Consequently, an insurmountable number of people can benefit tremendously from any efforts made to reduce adverse impact and promote fair selection procedures for all groups of people. Implementing a contingency-based approach to scoring AC data is one small step in reaching this goal. Overall, the contingency approach yielded a new scoring method that resulted in less adverse impact overall with different people qualifying for the eligible cut-off ranks compared to the other scoring methods. To the extent the contingencies are valid, better candidates would be selected using this approach.

APPENDIX A: FIRE LIEUTENANT ASSESSOR DEMOGRAPHICS

Assessor #	Rank	Race	Sex	Assessor #	Rank	Race	Sex
1	Lieutenant	Black	Female	19	Captain	Other	Female
2	Captain	Black	Female	20	Lieutenant	White	Male
3	Lieutenant	Black	Female	21	Lieutenant	White	Male
4	Captain	Black	Female	22	Lieutenant	White	Male
5	Captain	Black	Female	23	Captain	White	Male
6	Lieutenant	Black	Male	24	Lieutenant	White	Male
7	Captain	Black	Male	25	Engineer/FEO	White	Male
8	Lieutenant	Black	Male	26	Lieutenant	White	Male
9	Chief	Black	Male	27	Lieutenant	White	Male
10	Lieutenant	Black	Male	28	Captain I	White	Male
11	Platoon Commander	Black	Male	29	Captain	White	Male
12	Lieutenant	Black	Male	30	Lieutenant	White	Male
13	Lieutenant	Black	Male	31	Lieutenant	White	Male
14	Lieutenant	Black	Male	32	Captain	White	Male
15	Lieutenant	Black	Male	33	Lieutenant	White	Male
16	District Chief	Black	Male	34	Lieutenant	White	Male
17	Lieutenant	Black	Male	35	Captain	White	Male
18	Lieutenant	Black	Male	36	Lieutenant	White	Male

APPENDIX B: FIRE CAPTAIN ASSESSOR DEMOGRAPHICS

Assessor #	Rank	Race	Sex	Assessor #	Rank	Race	Sex
1	Lieutenant	Black	Female	18	Captain	Black	Male
2	2nd Deputy Fire Commissioner	Black	Female	19	Captain	Hispanic	Male
3	Battalion Chief	Black	Female	20	Captain	Other	Male
4	Captain	Black	Female	21	Captain	White	Female
5	Captain	Black	Female	22	Captain	White	Male
6	Battalion Captain	Black	Male	23	Division Chief	White	Male
7	Fire Captain	Black	Male	24	Fire Section Chief	White	Male
8	Captain	Black	Male	25	Fire deputy chief	White	Male
9	Captain	Black	Male	26	Major	White	Male
10	Chief	Black	Male	27	Major	White	Male
11	Lieutenant	Black	Male	28	Assistant Chief	White	Male
12	Platoon Commander	Black	Male	29	Captain	White	Male
13	Deputy Chief	Black	Male	30	Captain	White	Male
14	Captain	Black	Male	31	Battalion Chief	White	Male
15	Captain	Black	Male	32	Captain	White	Male
16	Fire Captain II	Black	Male	33	Fire Marshall Chief	White	Male
17	Captain	Black	Male	34	Battalion Chief	White	Male

APPENDIX C: CONFIDENTIALITY AGREEMENT

I, _____ will observe the following rules while serving as a Subject Matter Expert for this examination:

- I understand that my work as a Subject Matter Expert is of a highly confidential nature. I will not discuss any information presented or discussed in any test development meetings/activities to any business or professional associates, superiors, subordinates, friends, relatives, or anyone else not specifically authorized.
- I will perform all my assigned work as a Subject Matter Expert in the work spaces designated. I understand that I am not permitted to take any test-related materials, including personal notes, from the designated work area.
- I will take all precautions necessary to safeguard the integrity of the testing process and prevent any candidate from gaining any information regarding the examination process. I will consult the PBJC staff if any question or problem arises; no matter how minor it seems, concerning test security or the propriety of any matter relating to the examination.
- I will not duplicate or reproduce, in any form, any materials, including personal notes used in service as a Subject Matter Expert.
- I will not, in any way, directly or indirectly, help others prepare for the examination, and I will not advise others who may be helping candidates prepare.
- I certify that I do not have any relative who is a candidate in the examination process for which I am serving as a subject matter expert. I will notify the Director immediately if I discover that I have any relative(s) who is a candidate in the testing process. I understand that I will not be permitted to serve in this process if I have a relative who is a candidate for employment for the job in which I am serving as a subject matter expert.
- If any member or representative from a Fire Department located within the county attempts to contact me, either formally or informally, I agree to fully withhold the nature of my work with the PBJC from the individual(s). I will also immediately provide the date, time, and nature of the contact attempt, as well as any information identifying the contact (e.g., name, phone number).
- I understand that these rules are designed to protect the integrity of the testing procedures and I recognize that failure to adhere to these rules has significant consequences on the integrity of the testing process. I further understand that should I fail to adhere to the terms of this confidentiality agreement that disciplinary action will be sought to the fullest extent possible.
- I certify that a staff member has discussed the importance of test security and the importance of maintaining all information regarding the examination confidential, including the consequences associated with any breach on my part of this confidentiality agreement.

_____ _____

SIGNATURE OF SUBJECT MATTER EXPERT DATE

_____ _____

SIGNATURE OF ANALYST DATE

APPENDIX D: DETAILS REGARDING AC SCENARIOS AND ASSESSMENT

Supervisory Exam

Scenario 1: Candidates were given detailed information about an ongoing conflict in the department regarding several firefighters on their shift making rude remarks about a veteran firefighter. Each candidate responded as a recently promoted Fire Lieutenant acting as shift commander with the responsibility of supervising their former peers. Candidates were asked to indicate how they would address this situation? Additionally, candidates were asked to answer a follow-up question related to the scenario.

Scenario 2: Candidates observed video clips of sexual harassment incidents involving two firefighters. Candidates were asked to respond verbally to both of the firefighters together with corrective measures in addition to providing written documentation of the incident.

Scenario 3: This scenario was the written component of the Fire Lieutenant Supervisory Examination where candidates completed an incident report, including a narrative description of a car accident they witnessed. This part of the exam was designed to measure candidates' ability to accurately complete forms and communicate in writing.

At the end of the video-based test on the first day, candidates were escorted as a group to a large classroom to complete the written exercise. Candidates were equipped with a Participant Manual, pens, pencils, scratch paper, and a dictionary for the written exercise. The Participant Manual provided detailed information regarding the administration of the exam. Candidates were once again provided with video-based instructions in addition to the presence of a test monitor who remained available throughout the test session to answer procedural questions. Candidates were given 30 minutes to complete the written exercise and subsequently escorted to a check-out area by the test monitor where candidate materials were collected and checked to ensure that each candidate had turned in all test materials. Candidates completing the test in the

morning were held in a large waiting area and were not allowed to leave until all afternoon candidates had checked in. This precaution was taken to ensure that candidates who took the test early in the day could not leave the test site and share information about the content of the test with a candidate who was scheduled to take the test in the afternoon.

On the second day of the exam, candidates checked-in for the test by providing photo identification and signing the candidate roster. Each candidate received a unique identification number in addition to a color-coordinated ticket upon entry to the test room that designated the order in which he or she would complete the testing process. Candidates were once again grouped into "waves" of 18 individuals based on their order of arrival.

After completing the sign-in process, waves were escorted to a Preparation Room where the candidates were provided with written and video-based preparation material. Candidates were also allowed preparation time, which was based on job-related expectations of the amount of time job incumbents might have to consider and respond to situations arising in the work environment. For the same reasons specified for the first day of the exam, certain information was not disclosed until candidates were in the actual test situation. Candidates were given 20 minutes preparation time for the structured interview. Detailed video-based instructions were provided in addition to the presence of a test monitor to answer procedural questions during the preparation period and ensure integrity of the testing process.

At the conclusion of the Preparation Period, candidates retrieved their Technical Exam Participant Manual, which gave detailed information regarding the tasks to be completed, in addition to any exercise-related notes taken. Candidates were then escorted to individual testing rooms and seated at a desk facing the video monitor in view of the mounted video camera. Once

the test video was started candidates responded orally to each of three test scenarios. Each candidate's responses to the test scenarios were recorded for scoring at a later date.

Technical Exam

Scenario 1: Candidates responded to a fire call at a residence and acted as incident commander. Candidates were given specific information (e.g., time of day, temperature, size of the water pumper on the Engine, members of the crew the candidate will be working with, location of other units en route to the scene, whether or not possible citizens are trapped in the residence, etc.). Candidates responded aloud to several questions regarding the actions they would take in this situation.

Scenario 2: The candidates received detailed information regarding an incident where smoke had been observed coming from a restaurant (e.g., whether or not an evacuation seemed to be in progress, location of visible flames, location of the smoke, etc.). Candidates were asked to respond to several questions indicating how they would address this situation. Candidates subsequently responded aloud to a variety of follow-up questions regarding the information provided in the scenario.

Scenario 3: Candidates received detailed information regarding a car accident (e.g., weather, location, size of the water pumper on the Engine, members of the crew the candidate will be working with, location of other units en route to the scene, number of vehicles involved, description of possible injuries). Candidates were first asked to discuss the considerations they would take when en route to the scene. Candidates were subsequently provided specific information after arriving on the scene (e.g., exact number of vehicles involved, exact location and position of all vehicles, number of bystanders on the scene) and answered a follow-up

question based on that information. Finally, candidates were provided with specific information regarding the individuals involved in the accident (e.g., location, extent of injuries, etc.) and responded to another follow-up question as incident commander.

At the end of the video-based test on the second day, candidates were escorted to a check-out area by the test monitor where candidate materials were collected and checked to ensure that each candidate had turned in all test materials. Once again, candidates completing the test in the morning were held in a large waiting area and were not allowed to leave until all afternoon candidates had checked in.

Fire Captain Test Administration Procedure

The Fire Captain Exam consisted of three unique phases, which simulated a single shift for a Fire Captain and required candidates to complete several tasks that may be performed during a shift by a Fire Captain (see Table 6 for the specific dimensions assessed by each task). The Fire Captain promotional exam was administered at a large testing facility over the course of one day. Candidates checked-in for the test by providing photo identification, signing the candidate roster, and signing a Confidentiality Agreement for the exam. Each candidate received a unique identification number in addition to a color-coordinated ticket upon entry to the test room that designated the order in which he or she would complete the testing process. Candidates were grouped into “waves” of 18 individuals based on their order of arrival.

Table 46: Fire Captain Test Plan

KSAO Dimension	Assessment Method								
	Phase I		Phase I					Phase III	
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9
Dimension 1: Judgment and Decision Making			X						
Dimension 2: Management Ability	X	X							
Dimension 3: Oral Communication					X				
Dimension 4: Written Communication								X	X
Dimension 5: Professionalism				X	X				X
Dimension 6: Incident Command						X	X		
Dimension 7: Supervisory Ability				X					
Dimension 8: Leadership							X		
Dimension 9: Conflict Management				X					
Dimension 10: Analytical Skills									X
Dimension 11: Departmental/Jurisdictional Knowledge			X		X				
Dimension 12: Technical Knowledge – Emergency Response						X	X		
Dimension 13: Technical Knowledge – Firefighting						X	X		

After completing the sign-in process, waves were escorted to a group testing room where the candidates were provided with detailed written and video-based instructions in addition to the presence of a test monitor to answer procedural questions during the preparation period and ensure integrity of the testing process. Furthermore, all candidates were equipped with a Fire Department's Policy Manual, all of the test materials contained in the in-basket, color-coded scratch paper (for test security), a dictionary, pens and pencils. The Policy Manual provided detailed information for completing the tasks regarding the policy of the fire department. Candidates were then given one hour and fifteen minutes to complete their responses to Phase I

of the exam, which included the work sample in-basket exercises (i.e., Tasks 1 and 2). During this time period candidates were also allowed to prepare for the second and third phases of the exam. A timer was provided to candidates so that they could effectively manage their time during the examination period.

Phase I – In-Basket

Phase I of the Captain exam included an in-basket exercise consisting of two tasks that were designed to simulate some of the administrative activities associated with the job of Fire Captain. As such, these tasks were deigned to assess candidates' ability to manage various tasks in an organized and precise fashion. Before beginning the in-basket tasks, candidates were provided general instructions, which were both provided to the candidate in written form and read by a narrator on a video. They were also provided an in-basket envelope, which contained a copy of information about the fictitious fire department used for the AC and its policies, written instructions for each of the two in-basket tasks, all of the materials candidates needed to complete the in-basket tasks, and some materials that were related to tasks in the other phases of the exam. After candidates were read the instructions for the in-basket phase of the exam, they were given one hour and fifteen minutes to complete both of the tasks. Candidates were able to decide when to perform each task and how much time to dedicate to each. The two tasks are described in more detail below.

Task 1: Daily Schedule – Several people who were scheduled to work Shift A do not make it to work that day for various reasons. The candidate's job was to ensure all of the stations in Jefferson City meet minimum staffing requirements. Candidates are provided forms

in their in-basket to record and explain the steps they would take to maintain minimum staffing at all of the stations. This task is performed for the Assistant Fire Chief in his the absence.

Task 2: Processing Request for Leave – The candidates received several requests for leave for the following month's schedule. They were asked to process these requests and to determine which should be approved and which should be rejected. They were also provided forms in their in-basket to record and explain their decisions on each of the requests for leave. This task was also being performed for the Assistant Fire Chief in his absence.

Candidates were provided all of the materials they needed to complete these tasks in their in-basket envelope. Additionally, they were provided labeled forms on which to record their responses to each of the exercises. After the one hour and fifteen minutes provided to complete the in-basket tasks elapsed, candidates were asked to place all of their materials in their in-basket envelope to take to their individual test rooms. The video work sample tasks for Phase II of the AC, which are described next, were completed in the individual test rooms.

After completing Phase I of the exam, candidates listened to an audio clip of a voicemail forwarded to them by the Fire Chief, which detailed a citizen complaint about a Paramedic who worked at the candidate's station. The purpose of this voicemail was to provide additional information regarding the counseling and citizen complaint tasks, which the candidate was to complete in Phases II and III, respectively.

At the conclusion of Phase I, candidates gathered all of their test materials in addition to and exercise-related notes taken and placed them in an envelope labeled with the candidate's identification number. Candidates were then escorted to individual testing rooms and seated at a desk facing the video monitor in view of the mounted video camera where they were instructed to remove their test materials from the envelope. Once the test video was started candidates

viewed Tasks 3-7 on the video monitors. Candidates were allowed to refer to notes that they had taken during Phase I, as well as use any of the information provided to them about the fire department (i.e., Policy Manual), to complete Phase II of the exam.

Phase II - Video Work Sample Tasks

Phase II of the Fire Captain exam was the video work sample tasks, which consisted of five tasks that were administered through a video format on the video monitor. Each task required candidates to respond verbally to various questions. Candidates were allowed to use any materials or notes from their in-basket materials. Their responses were video and audio recorded to two DVDs.

Task 3, the first video work sample task, asked candidates to counsel an employee and was designed to measure candidates' judgment and decision making, supervisory ability, analytical skills and departmental/jurisdictional knowledge. Task 4 required candidates to respond to a conflict between two firefighters and measured professionalism, supervisory ability and conflict management. Task 5, which measured candidates' oral communication skills, professionalism and departmental/jurisdictional knowledge, called for candidates to make a presentation to the new Deputy Mayor and then field questions from the Deputy Mayor and her staff. Task 6 asked how candidates would respond to two evolving emergency situations and was intended to measure candidates knowledge of incident command, firefighting and emergency response. Task 7, the final video work sample task, measured the same three KSA clusters as Task 6, in addition to Leadership and Analytical Skills, but required candidates to answer several questions as they watched a video of a fire scene unfolding from the perspective

of an Incident Commander. Each of the five video work sample tasks is described in more detail below.

Task 3: Counseling a Firefighter – Candidates dealt with a firefighter who was accused by a citizen of being rude during a medical call. Candidates were asked to respond aloud to questions about how they would handle the situation.

Task 4: Conflict Management – After candidates counseled the firefighter, candidates were asked about how they would respond to a conflict between two of their employees.

Task 5: Presentation to Deputy Mayor – The Fire Chief asked the candidates to make a five minute presentation on the fictitious fire department to the new Deputy Mayor. As part of their preparation for the Fire Captain Exam, candidates were asked to prepare their five minute presentation before the day of the exam. They were not allowed to bring notes for their presentation to the exam, but they do have time during the in-basket to prepare new notes for their presentation. After they made their presentation, they were asked to respond aloud to some questions from the Deputy Mayor and her staff.

Task 6: Complete Training Needs Assessment Test – After making the presentation to the Deputy Mayor, candidates were asked to complete a training needs assessment test that had been provided to them by the training officer of the department. The exercise provided them a description of two emergency situations (e.g., fires, medical calls) and asks them to respond aloud to some questions about how each situation should be handled.

Task 7: Immersive Scenario – After completing the Training Needs Assessment Test, candidates responded to a fire call and acted as incident commander. The incident was presented in the format of a video. The video was filmed from the perspective of Captain Candidate. Therefore, everyone responded to the camera as if they were speaking to the candidate. In

addition, candidates were asked to respond aloud to questions about how they would respond to the changing conditions at the scene of the fire. The pre-fire plan for the building to which candidates were responding was provided to candidates in the individual test room in an envelope.

After completing Phase II, candidates were then escorted to a second group testing room to complete Phase III of the exam, which included the write-up tasks. To complete Phase III of the exam, all candidates were given a participant manual and supplemental documents relevant to Task 9 (Citizen Complaint Task). Candidates were provided with detailed video-based instructions in addition to the presence of a test monitor to answer procedural questions and ensure integrity of the testing process. Comparable to Phase I, a timer was provided to candidates so that they could effectively manage their time during the examination period.

Phase III - Write-up Tasks

Phase III of the exam consisted of two write-up tasks. The first write-up task, which was designed to assess candidates written communications skills, required them to complete an incident report that described the fire incident they just completed during the video work sample. The second write-up task was designed to measure candidates' written communication skills, professionalism, analytical skills and departmental/jurisdictional knowledge. For this task candidates were asked to conduct an investigation into a citizen complaint. At the conclusion of their investigation, they were told to write a memo to the Fire Chief and a letter to the complaining citizen at the conclusion of their investigation. The two write-up tasks are described in more detail below.

Task 8: Write Incident Report – Candidates were asked to complete an incident report for the immersive scenario that they completed earlier in the exam.

Task 9: Written Response to Citizen Complaint – The candidates' Chief asked them to handle a complaint he received from a citizen. Candidates were asked to investigate the complaint and then write a memo to the Fire Chief and a letter to the citizen that explained how they planned on handling the complaint.

At the end of Phase III, candidates were escorted to a check-out area by the test monitor where candidate materials were collected and checked to ensure that each candidate had turned in all test materials. Once again candidates completing the test in the morning were held in a large waiting area and were not allowed to leave until all afternoon candidates had checked in.

AC Dimension Measures

There were several dimensions the candidates were rated on throughout the ACs. There was some overlap in the dimensions assessed for Fire Lieutenant and Fire Captain; however, there were several dimensions only assessed for one of the two job levels. All dimensions were rated using either a behavioral checklist, a five-point Likert scale ranging from 1 = unacceptable to 5 = outstanding, or a combination of both rating styles. Lists of specific work behaviors were used as benchmarks to capture candidates' performance on each dimension. Example work behaviors associated with the various dimensions are provided.

Policies and Procedures Example work behaviors that were utilized to rate the candidates' performance on the dimension of policies and procedure include **'**Candidate doesn't mention that the issue must be documented,' 'Candidate explains that he/she

must document the incident/send up chain of command,' and 'Candidate explains how the subordinate's actions pertain to harassment policy'.

Safety and Life Preservation Example work behaviors that were utilized to rate the candidates' performance on the dimension of safety and life preservation include 'Establishes secondary crew for back-up of primary rescue team,' 'Informs rescue 2 personnel to proceed to scene and establish a temporary safe haven for emergency medical care,' and 'Ensures that all crewmembers are accounted for'.

Firefighting Tactical Knowledge Example work behaviors that were utilized to rate the candidates' performance on the dimension of firefighting tactical knowledge include **'**Orders ladder crew to roof with line for ventilation,'
'Begins horizontal/vertical venting,' and 'Provides future tactics for crew once rescue is complete'.

Supervisory Ability Example work behaviors that were utilized to rate the candidates' performance on the dimension of supervisory ability include **'**Candidate explains which actions are being taken toward the situation at this point,' 'Candidate emphasizes the importance of professionalism in the workplace,' and 'Outlines disciplinary actions that will be taken'.

Leadership Ability Example work behaviors that were utilized to rate the candidates' performance on the dimension of leadership ability include 'Candidate spoke to firefighter in a calm tone,' 'Candidate uses words that are neutral and non accusatory,' and 'Candidate explains/reviews what occurred at the scene concerning the comment/inappropriate behavior'.

Conflict Management Example work behaviors that were utilized to rate the candidates' performance on the dimension of conflict management include 'Candidate refuses to address the issue with the citizen.' 'Candidate apologizes to the citizen for the inappropriate comment made by the firefighter.' and 'Candidate attempts to calm the citizen'.

Fire Behavior Knowledge Example work behaviors that were utilized to rate the candidates' performance on the dimension of fire behavior knowledge include 'Orders crews to limit the amount of damage to the facility while still ensuring no spreading of fire,' 'If not already performed gives impact weather could have on fire scene during size up,' and 'Recognizes and communicates the building construction type during size up'.

Analytical Ability Example work behaviors that were utilized to rate the candidates' performance on the dimension of analytical ability include 'Candidate identifies the need for proper ventilation,' 'Candidate describes the size up considerations he/she would make, including the following: Construction, Occupancy, Life Safety, Water supply, Exposures, Location/extent,' and 'Candidate identifies the structure's fire protection system'.

Judgment and Decision Making Example work behaviors that were utilized to rate the candidates' performance on the dimension of judgment and decision making include 'Candidate assigns a crew to evacuation,' 'Candidate states that he/she will preserve evidence of cause or origin,' and 'Candidate states that he/she would pass or establish command'.

Oral Communication Example work behaviors that were utilized to rate the candidates' performance on the dimension of oral communication include 'Candidate's statements

were clear and demonstrated well thought out ideas,' 'Ideas were communicated clearly,' and 'Varied pitch to maintain attention of the listener'.

Written Communication Example work behaviors that were utilized to rate the candidates' performance on the dimension of written communication include 'Provides accurate information in appropriate blanks,' 'Sentences are grammatically correct,' and 'Writes coherent, logical thoughts down in narrative blanks'.

Incident Command/IMS Example work behaviors that were utilized to rate the candidates' performance on the dimension of incident command include 'Candidate states that he/she would contact dispatch, call for additional companies, and provide dispatch with an on scene report,' 'Candidate addresses the on scene command issue,' and 'Candidate states that he/she would request that dispatch contact the police'.

Professionalism Example work behaviors that were utilized to rate the candidates' performance on the dimension of professionalism include 'Candidate indicates that the citizen's concern is valid,' 'Candidate maintains confidentiality of material uncovered during the investigation,' and 'Candidate assures the citizen will be notified of the results of the investigation'.

Departmental/Jurisdictional Knowledge Example work behaviors that were utilized to rate the candidates' performance on the dimension of departmental/jurisdictional knowledge include 'States he/she will coordinate the investigation into the incident,' 'States he/she will interview as many participants in the incident as necessary to determine what occurred at the incident,' and 'States he/she will determine which unit or units were involved in the incident'.

Technical Knowledge – Emergency Response Example work behaviors that were utilized to rate the candidates' performance on the dimension of emergency response technical knowledge include 'Establish a safe zone for civilians and personnel, evacuates people the appropriate distance away from the structure,' 'Ensures that residents are sheltered,' and 'Ensures medical needs of residents are met'.

Assessment

The assessments of the Fire Lieutenant and Fire Captain exams were conducted over the course of one week each. All assessors underwent specialized training sessions conducted by experienced job analysts which included information and practice opportunities for observing and recording behavior, categorizing behavior, evaluating behavior, and making ratings of behavior, as well as how to appropriately take notes and categorize notes in terms of performance dimensions. Assessors also received training regarding the administrative aspects of the assessment, such as completing rating forms and entering ratings into the computer system. Assessors were trained on common rater errors (e.g., halo error) and the remedies for those errors. Additionally, all assessors were provided with a detailed description of the job they were assessing (e.g., Fire Lieutenant or Fire Captain), including knowledge, skills, abilities and other characteristics that were measured by the exercises.

As a critical part of training, assessors were given multiple opportunities to practice making ratings based on observations of candidate responses to each component of the exam. Each assessor made independent ratings of performance using the benchmarks developed for the exercise. Analysts preceded to conduct a calibration session, in which individual ratings were summarized on a large flip-chart and ratings were identified and discussed with assessors.

Consensus was then reached through group discussion and additional review of the benchmarks demonstrated by the candidate.

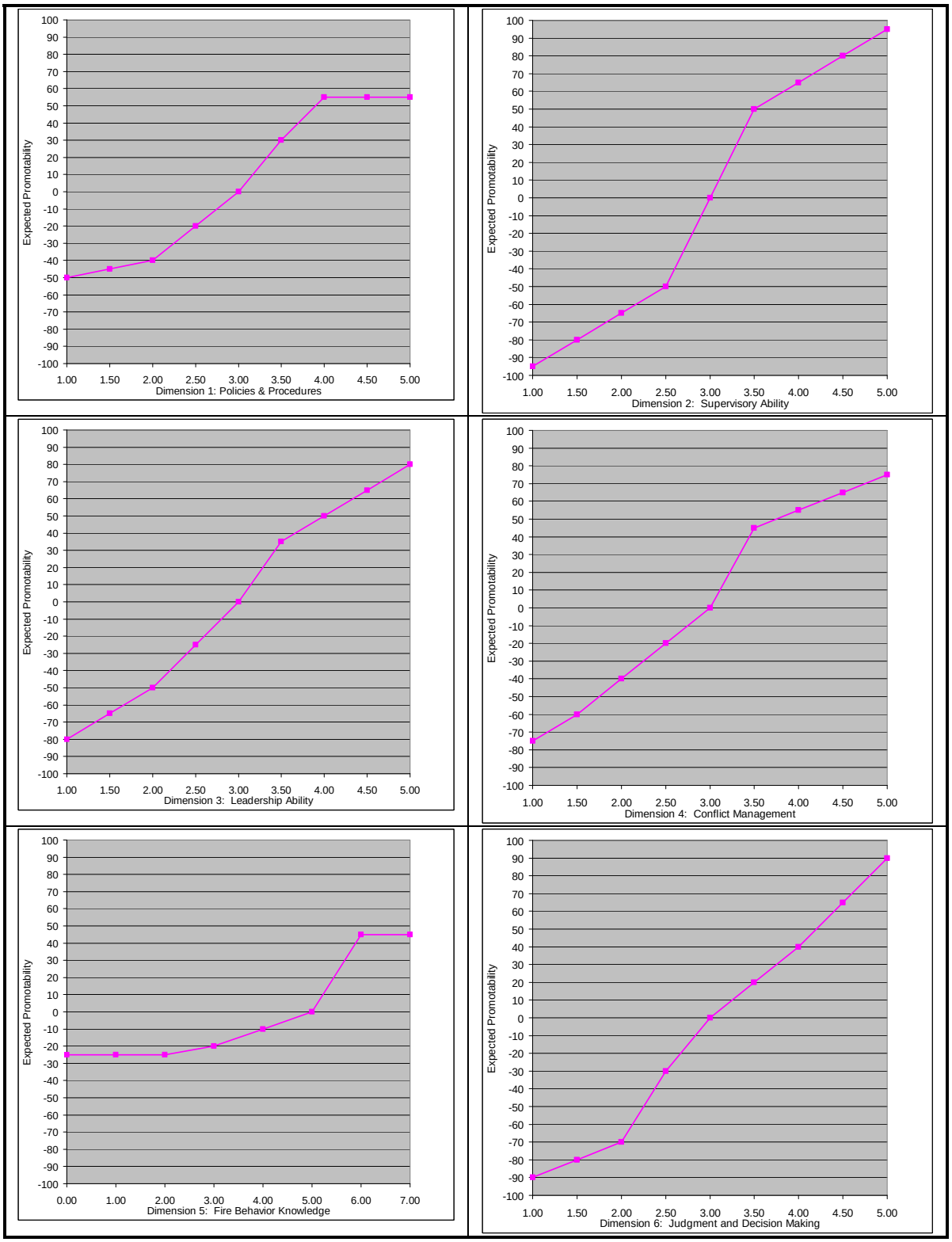
Benchmarks for many of the exercises were anchored to three points of a five-point rating scale: (1) Unacceptable, (3) Acceptable, and (5) Outstanding. All assessors were carefully trained in making fine distinctions on the rating scale by considering all benchmarks observed by candidates. For other exercises, a checklist rating scale was used in which candidates were given one point for each benchmark hit. Assessors were required to reach 100% agreement on each benchmark on such checklists throughout the assessment.

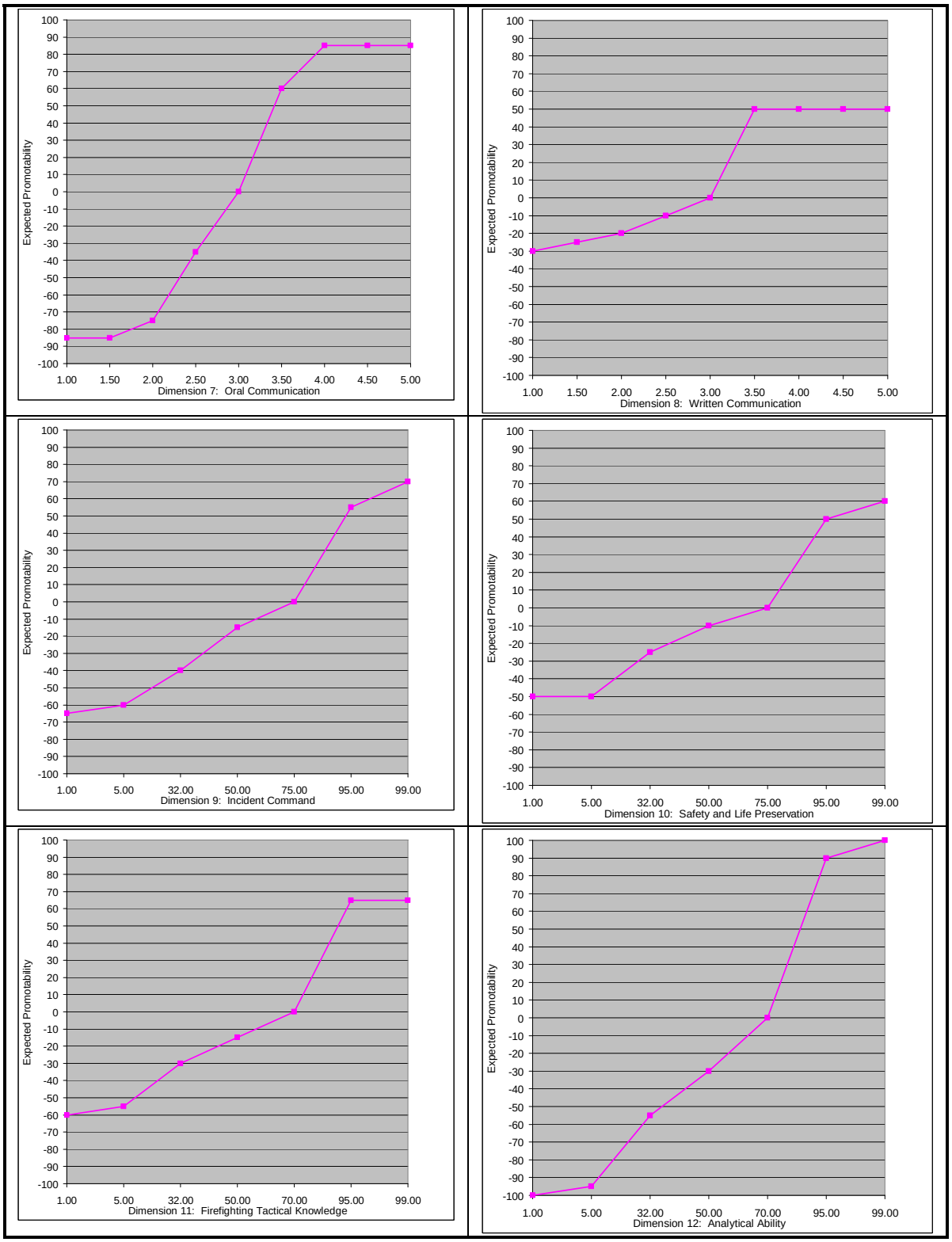
Assessors worked in appropriately diverse pairs to rate candidate performance on each component of the examination to ensure that each candidate was scored by at least one individual who was demographically similar (e.g., match on race, sometimes gender) to him/herself, thus avoiding potential rater biases based on race or gender. Additionally, for all components, assessor panels were rotated frequently throughout the assessment to avoid the risk of panels getting comfortable with each other to the extent that this could introduce inappropriate variance to ratings of performance.

For the job of Fire Lieutenant assessors observed the performance of 326 candidates on the video-based components and read photocopies of written exercise response materials. All assessors made independent, preliminary ratings of performance using benchmark rating forms provided. Once preliminary ratings were obtained, assessors discussed their ratings and made independent, final ratings with the requirement that final ratings had to fall within one scale point of each other for Likert rating scale, and checklist ratings had to match perfectly. In special cases where discrepancies could not be reconciled, the candidates were rated by a second panel to ensure appropriate consensus ratings were reached.

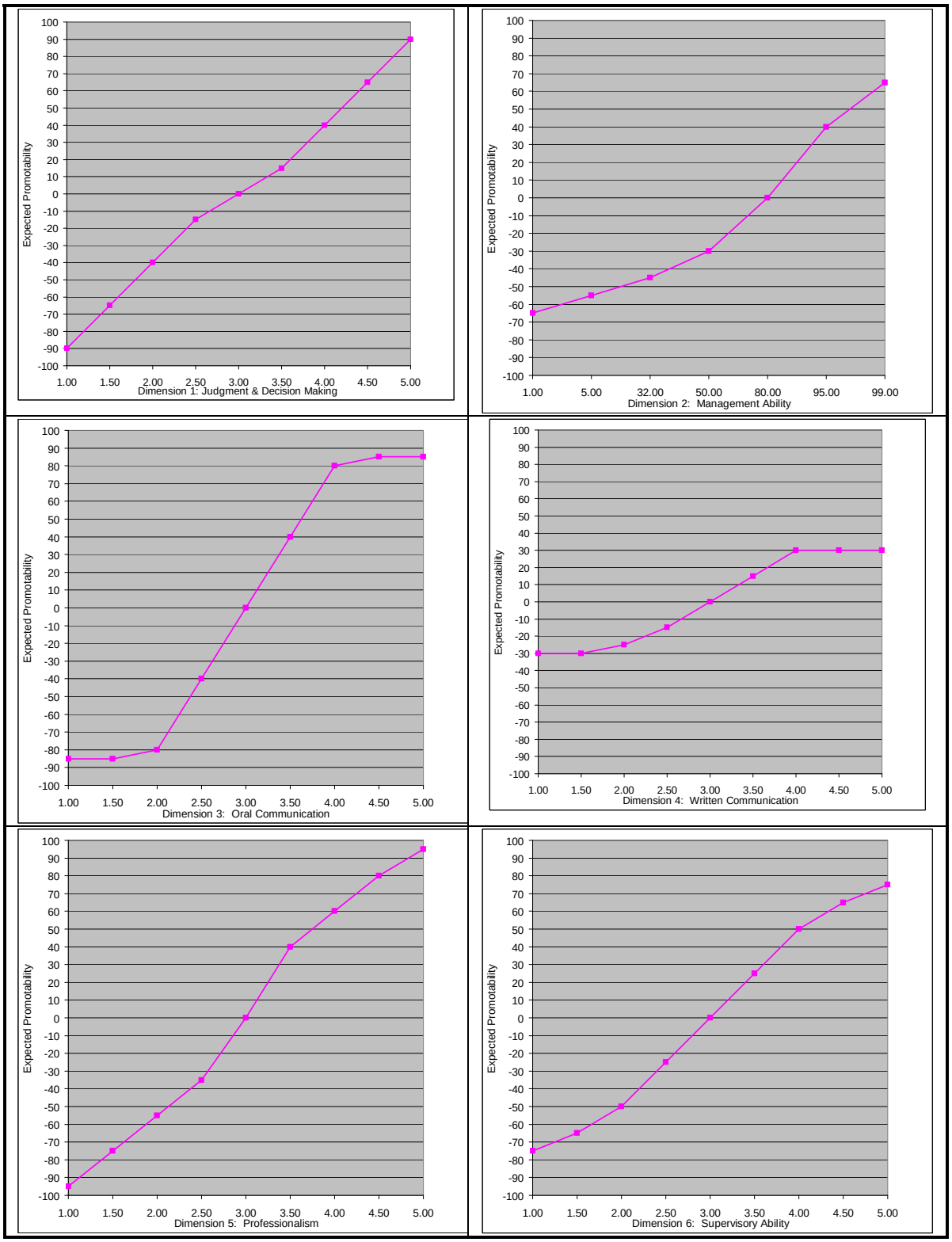
For the job of Fire Captain the in-basket exercises for all 118 candidates were scored by trained analysts at the testing facility. Each candidate's in-basket task (i.e., Tasks 1 and 2) was scored by two analysts where candidates' responses were objectively scored by comparing their answers to an answer key. Prior to scoring the in-basket tasks, the three assessors utilized in this scoring process reviewed the rating forms and score keys for each task to ensure they understood how to score the tasks and how to correctly interpret candidates' responses. The assessor pairs were not demographically balanced because the written responses did not disclose any demographic information that could potentially lead to rater bias. Assessors observed the performance of 118 Fire Captain candidates on the video-based components (Tasks 3-9) and read photocopies of write-up tasks (i.e., Task 8 and 9). All assessors made independent, preliminary ratings of performance using benchmark rating forms provided. Once preliminary ratings were obtained, assessors discussed their ratings and made independent, final ratings with the requirement that final ratings had to fall within one scale point of each other for Likert scale benchmarks, and checklist ratings had to match perfectly. In special cases where discrepancies could not be reconciled, thee candidates were rated by a second panel to ensure appropriate consensus ratings were reached.

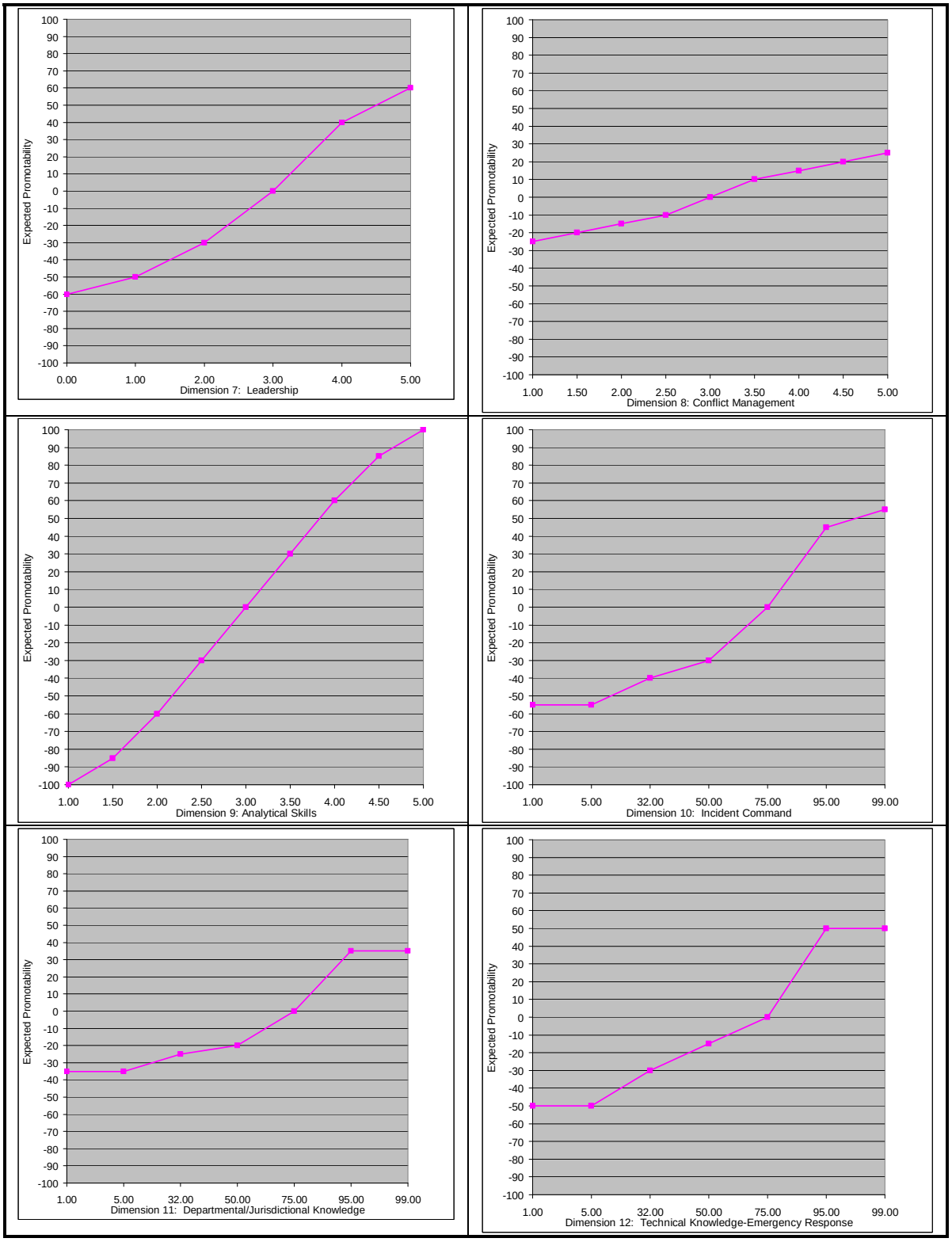
APPENDIX E: FINAL CONTINGENCIES FOR FIRE LIEUTENANT

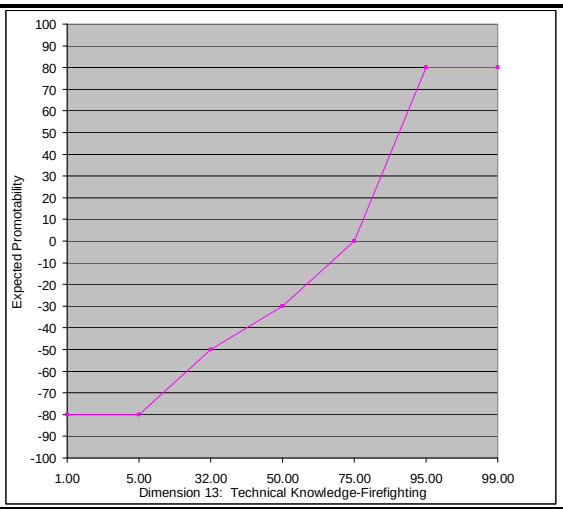

Expected Promotability
Dimension 1: Policies & Procedures
Dimension 2: Supervisory Ability
Dimension 3: Leadership Ability
Dimension 4: Conflict Management
Dimension 5: Fire Behavior Knowledge
Dimension 6: Judgment and Decision Making


Expected Promotability
Dimension 7: Oral Communication
Dimension 8: Written Communication
Dimension 9: Incident Command
Dimension 10: Safety and Life Preservation
Dimension 11: Firefighting Tactical Knowledge
Dimension 12: Analytical Ability

APPENDIX F: FINAL CONTINGENCIES FOR FIRE CAPTAIN


Expected Promotability
Dimension 1: Judgment & Decision Making
Dimension 2: Management Ability
Dimension 3: Oral Communication
Dimension 4: Written Communication
Dimension 5: Professionalism
Dimension 6: Supervisory Ability


Expected Promotability
100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100
0.00 1.00 2.00 3.00 4.00 5.00
Dimension 7: Leadership
Expected Promotability
100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100
1.00 1.50 2.00 2.50 3.00 3.50 4.00 4.50 5.00
Dimension 8: Conflict Management
Expected Promotability
100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100
1.00 1.50 2.00 2.50 3.00 3.50 4.00 4.50 5.00
Dimension 9: Analytical Skills
Expected Promotability
100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100
1.00 5.00 32.00 50.00 75.00 95.00 99.00
Dimension 10: Incident Command
Expected Promotability
100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100
1.00 5.00 32.00 50.00 75.00 95.00 99.00
Dimension 11: Departmental/Jurisdictional Knowledge
Expected Promotability
100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100
1.00 5.00 32.00 50.00 75.00 95.00 99.00
Dimension 12: Technical Knowledge-Emergency Response


100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
Expected Promotability
1.00
5.00
32.00
50.00
75.00
95.00
99.00
Dimension 13: Technical Knowledge-Firefighting

APPENDIX G: FIRE CAPTAIN SME RATING FORM

Fire Captain SME Rating Form

1a. Were the SMEs able to perform each step of contingency development? Yes or No.

1b. All SMEs appeared to be able to perform each step of the process. Please rate your level of agreement with this statement on the scale provided below.

1 = strongly disagree 3 = agree 5 = strongly agree

2. Was there disagreement at each step, how much, how was it resolved?

3a. Was total consensus obtained during the process or did SMEs appear to agree just to move along with the process? Yes or No.

3b. Total consensus was obtained during the contingency development process. Please rate your level of agreement with this statement on the scale provided below.

1 = strongly disagree 3 = agree 5 = strongly agree

4a. Did SMEs appear to be involved during the process? Yes or No.

4b. SMEs appeared to be involved during the process. Please rate your level of agreement with this statement on the scale provided below.

1 = strongly disagree 3 = agree 5 = strongly agree

5a. Did SMEs agree resulting contingencies were accurate reflections of the Fire Captain position? Yes or No.

5b. SMES agreed that the resulting contingencies were accurate reflections of the Fire Caption position. Please rate your level of agreement with this statement on the scale provided below.

1 = strongly disagree 3 = agree 5 = strongly agree

APPENDIX H: IRB LETTER

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Exempt Review Status

From: **UCF Institutional Review Board**
FWA00000351, Exp. 5/07/10, IRB00001138

To: **Keisha Wicks**

Date: **April 25, 2008**

IRB Number: **SBE-08-05627**

Study Title: **The Effects of a Contingency-Based Method for Combining Individual Assessment Center Dimension Ratings on Overall Assessment Ratings**

Dear Researcher:

Your research protocol was reviewed by the IRB Vice-chair on 4/24/2008. Per federal regulations, 45 CFR 46.101, your study has been determined to be **minimal risk for human subjects and exempt** from 45 CFR 46 federal regulations and further IRB review or renewal unless you later wish to add the use of identifiers or change the protocol procedures in a way that might increase risk to participants. Before making any changes to your study, call the IRB office to discuss the changes. **A change which incorporates the use of identifiers may mean the study is no longer exempt, thus requiring the submission of a new application to change the classification to expedited if the risk is still minimal.** Please submit the Termination/Final Report form when the study has been completed. All forms may be completed and submitted online at https://iris.research.ucf.edu.

The category for which exempt status has been determined for this protocol is as follows:

4. Research involving the collection or study of existing data, documents, records, pathological specimens or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. ("***Existing***" means already collected and/or stored before your study starts, not that collection will occur as part of routine care.)

All data, which may include signed consent form documents, must be retained in a locked file cabinet for a minimum of three years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained on a password-protected computer if electronic information is used. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

On behalf of Tracy Dietz, Ph.D., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 04/25/2008 08:51:48 AM EDT

Joanne Muratori

IRB Coordinator

REFERENCES

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125-154.

Arthur, W., Doverspike, D., & Barrett, G. V. (1996). Development of a job analysis-based procedure for weighting and combining content-related tests into a single test battery score. *Personnel Psychology, 49,* 971-985.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67,* 3-9.

Boyle, S., Fullerton, J., & Yapp, M. (1993). The rise of the assessment centre: A survey of AC usage in the UK. *Selection and Guidance Review, 9,* 1-4.

Bray, D. W. & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs, 80* (17, Whole No. 625).

Buster, M., Bobko, P., & Roth, P. (June, 2003). *Content valid composites: The elegance of unit weights.* Paper presented at the Annual IPMAAC Conference on Personnel Assessment, Baltimore, MD.

Byham, W. C. (1980). Starting an assessment center the correct way. *Personnel Administrator,* 27-32.

Cascio, W. F. & Silbey, V. (1979). Utility of the assessment center as a selection device. *Journal of Applied Psychology, 64,* 107-118.

Cohen, S. L. (1980). The bottom line on assessment center technology. *Personnel Administrator,* 50-56.

Cascio, W. F. (1987). *Applied Psychology in Personnel Management.* Englewood Cliffs, NJ: Prentice-Hall.

Collins, M. W. & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93,* 463-471.

Coulton, G. F. & Field, H. S. (1995). Using assessment centers in selecting entry-level police officers: Extravagance or justified expense? *Public Personnel Management, 24,* 223-254.

Dean, A. M., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93,* 685-691.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M. & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality, 12,* 85-108.

Einhorn, H. J. & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171-192.

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C. (2006, May). *Assessment center: Current practices in the United State.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127-148.

Feltham, R. (1988). Assessment centre decision making: Judgmental vs. mechanical. *Journal of Occupational Psychology, 61,* 237-241.

Feltham, R. (1988). Validity of a police assessment centre: A 1-19 year follow-up. *Journal of Occupational Psychology, 61,* 129-144.

Filer, R. J. (1979). Assessment center method in the selection of law enforcement officers. In C. D. Spielberger (Ed.), *Police Selection and Evaluation.* New York: Hemisphere.

Gatewood, R. D. & Feild, H. S. (2001). *Human Resource Selection.* Harcourt, Inc. New York.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1985). *Meta-analyses of Assessment Center Validity.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 93-511.

Gaugler, B. B. & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74,* 611-618.

Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51,* 357-374.

Hardison, C. M. & Sackett, P. R. (2004, April). *Assessment center criterion-related validity: A meta-analytic update.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Chicago.

Hays, W. L. (1994) *Statistics* (5th ed.). New York: Holt, Rinehart & Winston.

Hennessy, J., Mabey, B., & Warr, P. (1998). Assessment center observation procedures: An experimental comparison of traditional, checklist and coding methods. *International Journal of Selection and Assessment, 6,* 222-231.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15,* 405-411.

Hinrichs, J. R. & Haanpera, S. (1976). Reliability of measurement in situational exercises: An assessment of assessment center method. *Personnel Psychology, 29,* 31-40.

Hoffman, C. C. & Thornton, G. C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50,* 455-470.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative prediction of job performance. *Psychological Bulletin, 96,* 72-98.

Industrial Relations Services (1997). The state of selection: An IRS survey. *Employee Development Bulletin, 85,* 8-18.

Jones, A. (1981). Inter-rater reliability in the assessment of group exercises at a UK assessment centre. *Journal of Occupational Psychology, 54,* 84-96.

Keil, E. C. (1981). *Assessment Centers: A Guide for Human Resource Management.* Reading, MA: Addison-Wesley.

Klimoski, R. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40,* 243-260.

Klimoski, R. J. & Strickland, W. J. (1977). Assessment centers-valid or merely persistent. *Personnel Psychology, 30,* 353-361.

Klimoski, R. J. & Strickland, W. J. (1981). *A Comparative View of Assessment Centers: A Case Analysis.* Unpublished manuscript.

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20,* 345-362.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89,* 377-385.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13,* 323-353.

Landy, F. J. & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Lawshe, C. H. (1987). Adverse impact: Is it a viable concept? Professional Psychology Research and Practice, 18, 492-497.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86,* 255-364.

Lievens, F. & Klimoski, R. J. (2001). Understanding the assessment centre process: Where are we now? *International Review of Industrial and Organizational Psychology, 16,* 246-286.

Linnane, J. (1985). National survey of police recruiting practice. Internal report, Police Extended Interview Office.

Michel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60,* 573-579.

Meehl, P. E. (1954a). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology, 6,* 102-109.

Moore, D. S., & McCabe, G. P., (1993). *Introduction to the Practice of Statistics* (2nd ed.). New York: W. H. Freeman and Company.

Morris, S. B. (2001). Sample size required for adverse impact analysis. *Applied HRM Research, 6,* 13-22.

Morris, S. B. & Lobsenz, R. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53,* 89-111.

Office of Federal Contract Compliance Programs (1993). *Federal contract compliance manual.* Washington, D.C.: Department of Labor, Employment Standards Administration, Office of Federal Contract Compliance Programs (SUDOC# L 36.8: C 76/1993).

O'Leary, L. R. (1979). *The Selection and Promotion of the Successful Police Officer.* Springfield, IL: Thomas.

Pritchard, R. D. (1990). *Measuring and improving organizational productivity: A practical guide*. New York: Praeger.

Pritchard, R. D. (1992). Organizational productivity. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 3* (2nd ed., pp. 443-471). Palo Alto, CA: Consulting Psychologists Press.

Pritchard, R. D. (Ed.). (1995). *Productivity measurement and improvement: Organizational case studies*. New York: Praeger.

Pritchard, R. D., Harrell, M., DiazGranados, D. & Sargent, M.J. (2007). The Productivity Measurement and Enhancement System: A Meta-Analysis. Accepted to *Journal of Applied Psychology*.

Pritchard, R. D., Holling, H., Lammers, F., & Clark, B. D. (Eds.). (2002). *Improving organizational performance with the Productivity Measurement and Enhancement System: An international collaboration*. Huntington, NY: Nova Science.

Pritchard, R. D., Jones, S. D., Roth, P. L., Stuebing, K. K., & Ekeberg, S. E. (1988). The effects of feedback, goal setting, and incentives on organizational productivity. *Journal of Applied Psychology Monograph Series, 73*(2), 337-358.

Pritchard, R. D., Paquin, A. R., DeCuir, A. D., McCormick, M. J., & Bly, P. R. (2002). Measuring and improving organizational productivity: An overview of ProMES, The

Productivity Measurement and Enhancement System. In R. D. Pritchard, H. Holling, F. Lammers, & B. D. Clark (Eds.), *Improving organizational performance with the Productivity Measurement and Enhancement System: An international collaboration* (pp. 3-50). Huntington, NY: Nova Science.

Pritchard, R. D. & Roth, P. J. (1991). Accounting for non-linear utility functions in composite measures of productivity performance. *Organizational Behavior and Human Decision Processes, 50,* 341-359.

Pritchard, R. D., Watson, M. D., Kelly, K., & Paquin, A. (1998). *Helping teachers teach well: A new system for measuring and improving teaching effectiveness in higher education.* San Francisco: New Lexington Press.

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1*, 407-420.

Reynolds v. Alabama Department of Transportation, Civil Action No. 85-T-665-N, 1994 U.S. Dist. LEXIS NEXIS ACADEMIC.

Ross, J. D. (1980). Determination of the predictive validity of the assessment center approach to selecting police managers. *Journal of Criminal Justice, 8,* 89-96.

Roth, P.L., Bobko, P., & Switzer, F. (2006). Modeling the behavior of the 4/5th rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91,* 507-522.

Sackett, P. R. & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401-410.

Sackett, P. R. & Hakel, M. D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. *Organizational Behavior and Human Performance, 23,* 120-137.

Sackett, P. R. & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology, 67,* 10-17.

Sackett, P. R. & Tuzinski, K. (2001). The role of dimensions and exercises in assessment center judgments. In M. London (Ed.), *How People Evaluate Others in Organizations.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sarbin, T. R., Taft, R., & Bailey, D. E. (1960). *Clinical Inference and Cognitive Theory.* New York: Holt, Rinehart, & Winston.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178-200.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87,* 735-746.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407-422.

Schneider, J. R. & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77,* 32-41.

Shackleton, V. (1991). Management selection. A comparative survey of methods used in top British and French companies. *Journal of Occupational Psychology, 64,* 23-36.

Silverman, W. H., Dalessio, A., Woods, S. B., & Rudolph, L. J. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39,* 565-578.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in the United States. *Personnel Psychology, 50,* 71-91.

Stewart v. Rubin, Civil Action No. 90-2841, 1996, LEXIS NEXIS ACADEMIC.

Stillman, J. A. & Kirkley, W. (2007). When traits are behaviors: The relationship between behavioral responses and trait-based overall assessment ratings. *Human Performance, 20,* 415-432.

Thornton, G. C. (1992). *Assessment Centers in Human Resource Management.* New York: Addison-Wesley.

Thornton, G. C. & Byham, W. C. (1982). *Assessment Centers and Managerial Performance.* London: Academic Press.

Thornton, G. C. & Mueller-Hanson, R. A. (2004). *Developing Organizational Simulations: A Guide for Practitioners and Students.* Mahwah, NJ: Lawrence Erlbaum Associates.

Thornton, G. C., Murphy, K. R., Everest, T. M., & Hoffman, C. C. (2000). Higher cost, lower validity, and higher utility: Comparing the utilities of two tests that differ in validity, costs, and selectivity. *International Journal of Selection and Assessment, 8,* 61-75.

Tielsch, G. P. & Whisenand, P. M. (1977). *The assessment center approach in the selection of police personnel.* Santa Cruz, CA: Davis.

Tinsley, H. E. & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358-376.

Turnage, J. J. & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job

performance: Interpretive implications of criterion distortion for the assessment center. *Journal of Applied Psychology, 69,* 595-602.

Tziner, A. & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67,* 728-736.

U. S. Equal Employment Opportunity Commission. (n.d.). *Uniform employee selection guidelines on employee selection procedures. Federal Register, 43,* 38290-38315.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23-40.

Woehr, D. J. & Arthur, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29,* 231-258.

Wollowick, H. B. & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53,* 348-352.

Zedeck, S. & Cascio, W. F. (1984). Psychological issues in personnel decisions. *Annual Review of Psychology, 35,* 481-492.