The Development of a Hybrid Scoring Key for a Situational Judgment Test Designed for Training Evaluation

Rolanda Findlay

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE

in

Psychology

Neil M.A. Hauenstein, Ph.D., Chair

Roseanne J. Foti, Ph.D.

Lee D. Cooper, Ph.D.

Defended April 20, 2007

Blacksburg, VA 24061

Keywords: Situational Judgment Test, Training Evaluation, Scoring Key

The Development of a Hybrid Scoring Key for a Situational Judgment Test Designed for Training Evaluation

Rolanda Findlay

(ABSTRACT)

As a low fidelity work simulation, Situational Judgment Tests (SJTs) are an affordable and practical way of empirically linking training and on-the-job performance, thereby providing a viable means of evaluating training effectiveness. An issue, when utilizing SJTs, is deciding the appropriate manner in which the SJT should be scored. Traditional SJT scoring methodologies, while successfully utilized for selection and prediction, pose specific challenges when applied to a SJT designed to evaluate the effectiveness of a training program.

This study discusses the shortcomings of traditional SJT scoring methodologies when used in the evaluation context. To overcome these challenges, an innovative scoring methodology, the Hybrid methodology, is presented. This study provides the detailed description of the Hybrid scoring key creation, and compares the Hybrid scoring key with two traditional scoring keys (Subject Matter Expert (SME) and Respondent-based scoring keys). Responses from a military training program are utilized to illustrate the distinctive effects of using the three different scoring approaches. The superiority of the hybrid scoring key, due to increased confidence in the key’s accuracy, and findings regarding training evaluation are discussed. Future research directions and practical applications of the research are also discussed.

Acknowledgements

I would like to thank my advisor and mentor, Neil M.A. Hauenstein, for his continued patience, thoughtful guidance, and unwavering support through every stage of this thesis. Thank you for believing in my ability and entrusting me with this amazing research opportunity. Your vote of confidence means more to me than words can convey.

I would also like to thank my committee members, Roseanne J. Foti and Lee D. Cooper, for their valuable feedback and insight, which caused me to dig deeper and think about the big picture. Your assistance has undoubtedly made this thesis a stronger finished product.

My sincerest thanks are extended to DEOMI, the Directorate of Research, Daniel McDonald, Jerry Scarpate, Rebecca Marcum, William McGuire, and the diligent EOA trainers and trainees, whose contributions all made this thesis possible. Thank you for supporting the SJT vision and adopting me into the DEOMI family.

Last but certainly not least, I would like to thank my family & friends. Your encouragement, love, and support remind me each day that I am truly blessed. Thank you for never letting me give up and always pushing me to reach for the sky.

Table of Contents

Introduction

Traditional models of training effectiveness argue that a well designed and successfully implemented training program will elicit positive reactions from trainees and will lead to knowledge acquisition. Effective training programs will lead to behavior changes and performance improvements in the workplace, which ultimately leads to improvements in organizational performance (Kirkpatrick, 1976; Ostroff, 1991). It is necessary to consider each of these expected outcomes of effective training to make a comprehensive and valid judgment of training success (Arvey & Cole, 1989; Kirkpatrick, 1976). A hindrance in the field of training evaluation is a recurring inability to accurately assess all of these outcomes; of most concern, behavior change and performance improvements in the workplace and ultimately improvement in organizational performance.

Every year, the United States Armed Forces invests considerably into the training and development of their human resources. Over the past few years, individual and collective training was estimated to cost upward of 30 billion dollars (Salas, Milham, & Bowers, 2003). In addition to the financial investment, military training requires significant time and energy to plan, develop, and implement. To be prepared and equipped for action, all military personnel, across all of the services, are required to participate in frequent and extensive training. During times of relative peace, military personnel spend 100% of their time on duty training and preparing for conflict (Salas *et al*., 2003). Considering the consequences of ineffective training and the magnitude of the investment necessary to train personnel, it is clear that frequent evaluation of the effectiveness of military personnel training is required.

This study is posited on the assumption that situational judgment tests (SJTs) traditionally used in selection contexts are a viable assessment strategy for evaluating training effectiveness.

SJTs measure both job knowledge and the integration of job knowledge which, in turn, reflects broader job-relevant competencies. Clearly, the assessment of broader competencies complements the goal of training evaluation. As such, the first purpose of this study is to demonstrate the use of the SJT assessment strategy in the training evaluation context. However, the training context provides a unique challenge for SJT scoring protocol development. Therefore, the second purpose of this study is to compare and evaluate different scoring protocol strategies when utilized for training evaluation.

Training Evaluation

Training is most commonly evaluated using trainees' affective opinion or reaction to training. Trainees' reactions can be measured via self-report surveys, questionnaires, or interviews. Examples of what is measured include perceptions of what was learned, perceptions of training effectiveness, and perceptions of training components/areas that should be improved. The self-report method is important because it allows recipients of training to give feedback on their training experience in an understandable and timely fashion. Trainee reaction data are not without flaws, though. Trainee perceptions can be inaccurate and biased. It is clear that trainee reactions can yield information about the training experience, but this information may be flawed, and trainee reactions will not yield information regarding the transfer of knowledge and skills to the work environment.

Another criterion commonly measured in training evaluation is learning outcomes; where training is evaluated by assessing trainees' demonstrated knowledge of the trained material. Demonstrated knowledge is usually measured via paper and pencil knowledge tests (e.g. proficiency tests, achievement tests). Similar to trainee reaction questionnaires, this is a relatively quick, convenient, and cost-effective methodology. Measuring learning outcomes is useful because the results allow trainers and evaluators to identify knowledge and skills that trainees successfully have learned.

In traditional models of training, it is presumed that training is ineffective if the trainees do not acquire the information imparted during training. However, the converse is not true. Acquisitions of knowledge and skills are a necessary but not sufficient condition for training transfer. Despite the obvious limitations regarding the fundamental issue of training transfer, trainee reactions and learning outcomes are commonly used as the sole evidence to conclude training effectiveness (Ostroff, 1991).

The most distal training criteria are organizational level outcomes, such as organizational profit/loss. The impact of training on these outcomes is usually correlated or estimated through simulation, as training can only be inferred or assumed to be causally linked to most organizational level outcomes. Particularly, it is difficult to pinpoint if benefits or declines at the organizational level are due to training, or to work environment variables, organizational leadership, individual employee factors, or the economy at large, which can have a substantial impact on an organization's bottom line figures.

Although most corporate decisions, including decisions on training, are based on broader organizational goals, the most important training evaluation criterion is that of behavioral transfer. Measures of behavioral transfer assess training effectiveness based on the trainee's ability to apply knowledge and skills acquired in training in the actual workplace. Behavioral transfer criteria are more proximal and individually-focused than organizational results outcomes; and simultaneously are designed to evaluate training outside of the training environment, unlike training reaction and learning outcomes.

When the transfer of training material is assessed based on the changes (improvements or declines) in trainee behavior on the job, it is most commonly accomplished through ratings of behavior and/or performance in the workplace by supervisors, co-workers, or subordinates. These methods often fail to show the effects of training or differences between trained and untrained groups (Ostroff, 1991). The lack of differences detected between trained and untrained groups could be due to a number of extraneous and confounding variables such as training design issues, trainee characteristics, work environment characteristics, and criterion issues (Ostroff, 1991). However even when measures are relevant and carefully developed, and confounding variables are controlled, there has still been a frequent failure to detect behavior and performance changes after training when rating scales are used as the criteria (Ostroff, 1991). This may be partly explained by the fact that performance and behavior ratings, while usually sufficient for examining and recording performance, by nature are limited to the observer's impressions, memory, and inferences.

Work samples designed to focus on the desired behavioral changes have been more successful at demonstrating transfer of training (e.g., Burnaska, 1976; Latham & Saari, 1979). Unfortunately, traditional work samples (such as those provided at assessment centers) often are expensive, and resource and access-dependent. Considering the success of measuring transfer of training using work samples, situational judgment tests (SJT), which are low fidelity work simulations, represent a viable alternative for measuring transfer of knowledge learned in training to later performance in the workplace.

SJTs are an excellent choice for training evaluation because they measure behavioral consistency, practical intelligence, and intentions of the respondent (Motowidlo, Dunnette, & Carter, 1990). Also, SJTs are a cost-effective and convenient way to predict performance and measure judgment, yielding similar results to traditional work samples and interviews (Ostroff, 1991). “The use of SJT methodology would provide a means of quantitatively evaluating the impact of training whose purpose was to improve judgment or skills [of pilots] using a pre-and posttest design and parallel forms of SJT measures”(Hunter, 2003, p. 383). Ostroff (1991) compared the use of SJTs and traditional ratings of performance in order to identify which methodology was more useful in measuring training effectiveness. Ostroff measured the effectiveness of a two-day training program designed to improve educator administrative and interpersonal skills. She compared the performance of trained and untrained personnel, before and after the training program. The measures of performance were supervisor ratings of behavior, supervisor ratings of skill, self-rating of skill, and a SJT. Results show that for both the trained and untrained groups, the mean scores on the rating measures were virtually identical before and after the training program. The SJT was the only measure of performance that exhibited a significant difference between trained and untrained groups over the course of the training program. Ostroff concluded that the SJT was more sensitive to detecting change due to training than on-the-job ratings of performance.

Situational Judgment Tests (SJTs)

Situational judgment tests (SJTs) have been utilized in a number of domains since the 1920s (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Chan & Schmitt, 1997; Dalessio, 1994; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Weekley & Jones, 1997). SJTs are multi-dimensional, rationally-derived measurement methods. SJTs are usually written inventories that present complex, hypothetical problem-situations to respondents followed by alternative ways to address the situation. The respondent is then asked to select the alternative course of action that is most/least effective, or the behavior they are most/least likely to engage. SJTs are considered “low-fidelity simulations” (Motowidlo *et al*., 1990) because each item provides a simulation or sample of performance by presenting a situation similar to one that would be encountered in the workplace. The assumption underlying SJT use is that individual job performance can be predicted based on how the individual performs on a simulation of the job (McDaniel & Nguyen, 2001). This methodology allows measurement of complex, multi-dimensional constructs outside of the environment where the construct would naturally be expressed.

Situational judgment tests have received increased attention and have grown in popularity over the past ten years because they are practical, cost-effective, and have shown validity comparable to cognitive measures, while producing smaller subgroup differences (Chan & Schmitt, 2002; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Hanson & Ramos, 1996; Hunter, 2003; Motowidlo *et al*., 1990; Motowidlo & Tippins, 1993; Weekley & Jones, 1997). SJTs are also a valid predictor of job performance in a variety of contexts (Becker, 2005; Chan & Schmitt, 2002; McDaniel *et al*., 2001).

Situational judgment tests can be useful in the domain of training evaluation for a number of reasons. First, as a multi-dimensional measure, it can be utilized to measure multi-dimensional constructs that cannot be validly and parsimoniously measured with traditional one-dimensional tests (Hunter, 2003). Employee training is comprehensive and involves the learning of multiple constructs, skills, and/or abilities; thereby, demanding evaluation that utilizes a multi-dimensional measure. Second, SJTs are valid predictors of both task and contextual performance (Chan & Schmitt, 2002) providing a realistic and thorough performance preview. Similar to traditional work sample tests, SJTs are able to assess whether trainees are able to utilize task and contextual knowledge, skills, and abilities learned in training in a variety of situations. Third, SJTs have been shown to be more resistant to faking than traditional paper-pencil tests, since there should be no transparent or blatantly correct answers (Becker, 2005). This aids SJTs in avoiding ceiling effects, and allows the measure to discriminate between high and low performers (Krokos, Meade, Cantwell, Pond, & Wilson, 2004), and between trained and untrained groups. The final benefit of using SJTs for training evaluation is that SJTs can measure the transfer of knowledge in the sense that performance on the SJT is a valid indicator of on-the-job performance.

SJT Development

There are a variety of ways to create a situational judgment test or inventory. The standard protocol for creating this rationally-derived measure is to first identify and define the construct/s to be measured (e.g. area of knowledge or competencies targeted in training). Next, critical incidents/situations from the job are gathered using job analysis techniques and/or the assistance of subject matter experts (SMEs). Incidents are then grouped into similar content areas. The test developer, using these critical incidents, writes hypothetical scenarios that are representative of the applicable content areas and that will elicit information regarding the variable of interest. The scenarios are edited for length, format, interpretability, applicability, and appropriateness (Lievens, 2000; McDaniel & Nguyen, 2001).

SMEs are asked to identify effective responses to the edited scenarios. It is likely that the responses obtained from SMEs will range in effectiveness, which allows different response options to be collected for each scenario. Response options are edited for length, format, appropriateness, and interpretability. The items (each scenario and corresponding response options) are gathered and compiled to create a SJT (Lievens, 2000; McDaniel & Nguyen, 2001).

Traditionally, developers of SJTs rely solely on SMEs to judge the appropriateness of an item to be included in a SJT. Inclusion depended on the SME and test developer's judgments of the items relationship with the construct being measured (Paullin, 2003). The limitation with this type of test development is that it inherently assumes that the developer and SMEs have the insight necessary to judge the relationship between items and construct being measured (Paullin, 2003). It is becoming more popular for test developers to use empirical data to help them edit and refine inventories, such as editing or removing items that show no response variance.

Another limitation with SJTs is that the situation and purpose of the SJT will determine the most appropriate approach to its creation and scoring. This yields variation in how SJTs are created and scored. These variations make it difficult to compare the relationship between factors in SJT construction and scoring and SJT validity and utility across studies. This is a persistent concern with the use of SJTs and is evidenced, for example, in the wide variation of SJT validity coefficients (Krokos *et al*., 2004).

A possible explanation for the SJT validity discrepancy may lie beyond SJT development and more so in the various ways situational judgment tests can be scored. (Bergman *et al*., 2006). In Bergman et al. (2006), the authors created 11 different scoring keys for a video-based leadership SJT. The authors applied the 11 scoring keys to SJT responses from a sample of 181 non-academic university supervisors. The SJT yielded large differences in the validity coefficients (i.e. -.03 to .32), depending on the scoring key applied. Of the 11 scoring keys applied, only three, the empirical, SME, and one of the hybrid keys, showed significant relationships to the leadership criteria. The lesson learned from this study is that the inappropriate selection of a scoring methodology "could lead to a conclusion that the SJT's content is not valid when it may only be the scoring key that is not valid" (Bergman et al., 2006, p.231). As to date, no consensus has been reached as to which SJT scoring approach, of the many available, is best. There is also no consensus as to the possible boundary conditions of a scoring method's superiority.

Present Study

There have been a number of recent calls for research to compare scoring protocols for situational inventories, to determine whether scoring produces significantly better results in terms of validity (Green, Alter, & Carr, 1993; Lievens, 2000; McHenry & Schmitt, 1994; Ostroff, 1991). In Bergman *et al*. (2006), the authors reviewed and compared several different methods of scoring a video-based SJT. They found that the validity of a SJT does depend, in part, on the way it is scored. They also suggest that a scoring system for a SJT should be chosen based on the test's purpose, and with respect to the test's applicable theory and practical constraints (Bergman *et al*., 2006).

Considering the impact a chosen scoring protocol can have on test utility and validity, it is essential that researchers continue developing and investigating new and improved scoring keys for SJTs. This is especially true for SJTs designed to evaluate training. There are several unique issues that need to be taken into consideration when creating and implementing a scoring key for SJTs designed for training evaluation, as opposed to SJTs created for selection purposes.

The purpose of this study is to develop a hybrid scoring key for a situational judgment test that is specifically designed to evaluate training, and to compare, in terms of training evaluation, the different methods of scoring key development. It is the author's intention to create the hybrid scoring key by combining elements of the SME-based and respondent-based scoring methodologies. The goals of the hybrid scoring key are to maintain the strengths of these traditional scoring methods while abating their weaknesses.

SJT Scoring Keys

SJT items consist of situations for which there is typically more than one effective response. They are by design not supposed to have one “single irrefutable correct answer” (Krokos *et al*., 2004). This is an important issue that test developers must take into consideration when selecting a scoring protocol. According to McDaniel & Nguyen (2001), SJTs can be scored based upon SME opinion (SME-based scoring keys), criterion data (empirical keys), or based upon central tendency statistics of SJT data (respondent-based keys).

SME-Based Scoring Keys

Most SJT scoring keys are developed rationally using the insight of SMEs. To create a scoring key, SMEs are asked to reach acceptable levels of agreement on the effectiveness of an item’s alternative responses. For example, they are asked to decide which is the most effective and/or least effective option presented. Each test developer must determine their own decision rules regarding how much agreement is necessary to establish the most effective answer. A common criterion is that of 75% agreement (Lievens, 2000). When agreement among SMEs cannot be reached on an item, the item needs to be revised to create consensus or deleted from the inventory (McDaniel & Nguyen, 2001).

SME-based scoring, also referred to as rational scoring, is useful if the selected SMEs are capable of identifying the ‘best’ response option based on their experience and expertise. In addition, they should be able to generate and identify alternative response options that are less than optimal. This type of test development and scoring methodology allows content validity to be inferred. In selection validation research, the criterion validity of rationally scored SJTs varies, but the results are generally positive (Krokos *et al*., 2004).

There are limitations to SME-based scoring protocols. First, SME-based scoring is labor-intensive and it can be difficult to access SMEs in a particular field or domain. If able to solicit the participation of appropriate SMEs, then it can be difficult to gain consensus among them about what is the most and/or least effective options (Lievens, 2000). It is expected that SMEs will have difficulty reaching acceptable agreement on twenty-five percent of the items on any given SJT inventory (Lievens, 2000). The ambiguity that makes SJT items resistant to respondent faking also makes it difficult for SMEs to reach agreement. Items where SMEs cannot reach acceptable agreement are removed from the inventory, and this increases the probability that the SJT will include items where the best response is the most socially desirable response (Krokos *et al*., 2004).

SMEs may also have difficulty reaching agreement because their judgments are bound by their unique perspectives, experiences, and biases. Even with their expertise, it is impossible for SMEs to predict relationships among items or between items and responding subgroups (Lievens, 2000). In this vein, SMEs are not able to predict the discrimination of SJT items, forcing items to be given equal weights (Lievens, 2000).

Despite these limitations, SJT developers have successfully used the SME-based scoring approach for prediction and selection (Hunter, 2003; Morath, Curtin, Brownstein, & Christopher, 2004; Motowidlo *et al*., 1990). Only one study (Ostoff, 1991) could be located that used SMEs to create a scoring key for a SJT designed specifically to evaluate a training program.

Beyond the comparison of traditional performance ratings to SJT scores, Ostroff (1991) also compared a SME-based scoring key against an empirical scoring key for the SJT. Thirteen SMEs were utilized to create the SME-based key. Each item was scored on a scale of 0 to 10 using this key. Each item’s response options were weighted depending on SME endorsement (i.e. a score of 10 was given if 100% of the SMEs endorsed this response option). This protocol is somewhat different than a traditional SME-based scoring protocol like that described in McDaniel & Nguyen (2001), where one correct answer is chosen for each item and response options are weighted equally.

The procedure for developing the empirical key in Ostroff's study was rather complex in comparison to the creation of a SME-based scoring key. Briefly explained, to create the empirical key, the author split a sample of 57 trainees into high and low criterion groups based on supervisors' behavior and skill ratings of said trainees. Based on the trainees' responses in the high and low criterion groups, percentages were formed for each item's response options. Based on these percentages, a weight was calculated for each item's response options. These results were successfully cross-validated using the same procedure with another group of 51 trainees. The net weights for each response option were then used to score the SJT. In comparing this empirical scoring key to the SME-based scoring key, Ostroff only observed a training effect when using the empirical scoring key. This finding is difficult to interpret, considering Ostroff's protocol for creating the SME-based scoring key was not traditional. Each response option was weighted based on the percentage of SME endorsement versus SMEs reaching a preset consensus for each item, and deciding the best and/or worst equally weighted response option. Nonetheless, Ostroff's findings indicate that rational scoring keys may not work as well as scoring keys developed from alternative methods.

Empirical Scoring Keys

For empirical scoring keys, assessee performance on relevant job criterion (criteria) is used to create the SJT scoring protocol. There are many different methods to develop an empirical scoring key, such as the correlational method, vertical percent method, horizontal percent method, mean criterion method, configural method, deviate keying method, rare response method, etc. The common theme across all empirical methods is that SJT items are scored and weighted according to empirical evidence that the items differentiate between individuals on a criterion (Paullin, 2003). Therefore, valid criterion data are necessary to compute any empirical scoring scheme (Legree, Psotka, Tremble, & Bourne, 2005).

There are a number of benefits when using an empirical scoring technique. Empirical scoring procedures are popular because they are more cost-efficient and less labor-intensive than SME-based scoring procedures. In addition, selection validation research has shown that empirical scoring protocols are equal or superior to SME-based scoring protocols when used for predictive purposes (Krokos *et al*, 2004; Paullin, 2003). Also, the larger sample sizes that are often afforded when creating empirical keys can bolster confidence in the key's reliability more so than with SME-based keys (Waugh & Russell, 2004).

Considering that the item responses that have the highest predictive validity may not be the response that is most intuitive (Paullin, 2003), it is advantageous that empirical scoring keys do not select correct responses based on rationale. By using criterion data, empirical scoring allows the inclusion of predictive, non-obvious items that normally would be removed when item inclusion is dependent upon on SME consensus. This inclusion of non-obvious items increases SJT resistance to respondent social desirability responding. As an alternative function, empirical scoring procedures can be utilized to complement SME-based scoring, serving as a validity check on SME ratings of the correct response option (Krokos et al., 2004).

The empirical scoring approach also has limitations. A concern with empirical keys is that they may lack face validity and be difficult to understand since they are not theoretically based (Paullin, 2003). Another major concern with empirical scoring keys is questionable generalizability across population due to statistical shrinkage (Krokos *et al*., 2004; Paullin, 2003; Reiter-Palmon & Connelly, 2000). In a recent study comparing predictive validity of a SME-based scoring key and various empirical scoring keys, Krokos *et al*. (2001), found that the empirical scoring protocols showed substantial statistical shrinkage during cross-validation. Only one of the empirical scoring methods, the correlational method, out of six empirical methods examined, retained its predictive accuracy in the cross-validation sample. Although generalizability from sample to sample is a major concern associated with empirical keys, it is assumed that through careful maintenance of the scoring key, utilization of a large and representative sample, selection of valid and reliable criterion variables, and assurance of item construct validity, instruments scored by empirical keys can show both stability and generalizability across samples (Paullin, 2003).

Another critical issue of concern when creating an empirical key lies in determining and obtaining appropriate criterion measurements. Difficulties in obtaining valid measurements of actual on-the-job performance can lead to deficiencies in the criterion. The level of inaccuracy due to deficiency in the criterion is not easily identifiable, but is something that should be taken into consideration when selecting measures of on-the-job performance and creating criterion keys. This potential criterion deficiency in criterion scoring keys can lead to inaccuracies in SJT measurement, which can lead to the miscalculation of trainee performance. Miscalculation will ultimately impact the judgment of whether a training program is effective.

Although biodata has a long history of using empirical scoring procedures and these procedures are adaptable to SJT items (Krokos *et al*., 2004), there are only a handful of published studies that have employed an empirical scoring key for SJTs (Becker, 2005; Dalessio, 1994; Gillespie, Oswald, Schmitt, Manheim, & Kim, 2002; Lievens, 2000; Ostroff, 1991;

Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004). Ostroff (1991) is the only published study that evaluated a training program using an empirical key to score a SJT. As described earlier, Ostroff's findings were in favor of the empirical key being the only SJT scoring key able to detect change in trainees over the course of training.

The findings of Ostoff (1991) provide food for thought regarding the impact of a chosen criterion and the limitations of interpreting the results of an empirical scoring key for the purpose of training evaluation. In Ostroff's study, the criterion utilized for the creation of the empirical scoring key was supervisory ratings of trainee behavior and skill, before and after training. Ostroff found that there was no difference, based on these supervisor ratings, for trained and untrained personnel. She also found no difference in trained and untrained personnel when utilizing the SJT scored by a SME-based key. Based on these findings the training program could be deemed ineffective. Beyond this, the same training program, the two-day training program for educators, was evaluated in a prior study (Schmitt, Noe, & Ostroff, 1986). Schmitt *et al*. (1986) attempted to measure the training's effectiveness through traditional evaluation measures, such as an in-basket measure of performance (work simulation), and behavior and performance ratings completed by trainees (self-rating), supervisors, and peers. The ratings were completed before training, three months after training, and six months after training. All of the measures utilized by Schmitt *et al*. (1986) failed to show significant positive training effects, for trained and non-trained groups. Again, based on these findings the training program could be deemed ineffective. The only evidence, after all of this training evaluation, of training effectiveness for the two-day educator training program studied in Ostroff (1991) was the SJT scored by an empirical scoring key.

It brings to bear the question of whether the SJT scored by the empirical key was the only measurement method capable of measuring the difference between trained and untrained groups and no other measure was sensitive enough to capture the trainee improvement. Conversely, it could be possible that based on the other measurement methods, there was a genuine lack of difference between trained and untrained groups and the empirical key utilized in Ostroff's study was based on a deficient criterion. Particularly, for the purpose of training evaluation, this is a situation that must be clarified, as the goal of training evaluation is to determine adequacies or deficiencies in a training program.

In the selection context, empirical keys are designed to maximize the predictive accuracy of the SJT scores. In one sense, an empirical key would be the best strategy for keying an SJT in that improved scores on the SJT as a function of training indicates that training is effective. However, there is an added complexity in using SJTs for training evaluation. In the selection context, items that do not exhibit predictive accuracy are simply dropped from the test. In the training context, though, the elimination of items that do not show improvement as a function of training is inappropriate. In the evaluation of training, SJT items that do not show an effect can reflect a problem with the trainees (i.e., they did not learn what they needed to learn) *or* a problem with the training program (i.e., the material was not covered or was not taught in a manner conducive to acquisition).   Thus, a purely empirical SJT key that drops items that do not show training effects will overestimate training effectiveness.

Respondent-Based Scoring Keys

There are many situations where valid and reliable criterion data are not available. It is difficult to consistently measure job performance when job variables, such as duties, assignments, and/or environment are not uniform across individuals in a position. Performance is also difficult to measure when job responsibilities are poorly and/or vaguely defined. More so, job performance is extremely difficult to measure when the aspects of work performance in question are internal and difficult to observe (e.g., diversity awareness/sensitivity). These difficulties are especially relevant in the realm of training evaluation where accurate and reliable measures of individual job performance must be collected before and after training. For situations where the measurement of criterion data is not appropriate or available, respondent-based scoring keys can be created.

In the respondent-based scoring approach, response patterns to the SJT from applicants / trainees are used to create the scoring key instead of the judgments of SMEs. According to Schulze and Roberts (2005), in the selection context, respondent distributions are a suitable alternative data source when SMEs and criterion data are not available. The respondent sample (e.g. incumbents, trainees, potential employees, students) utilized should be determined based on the purpose of the test. Item response options are then weighted and scored according to the respondent data (McDaniel & Nguyen, 2001). There are three different strategies for developing a respondent-based scoring protocol: 1) Most frequently endorsed response option, 2) Largest effect size, and 3) Combining the most frequently endorsed response option and largest effect size.

Most Frequently Endorsed Response Option. If the sample is carefully selected, the majority’s endorsement of a response option should be indicative of the most effective answer and predictive of the desired criterion. This may not always be the case, as this technique is limited to the selected respondents’ competency, knowledge, skills, and abilities. Even with a carefully selected sample, this technique confounds item difficulty with item validity. Utilizing the respondent endorsements, without accurately gauging item difficulty, can lead to incorrectly scored items and uncertain test interpretation.

Largest Effect Size. Another technique for determining the best response option is to select the response option that best discriminates between focal and comparison group(s) (e.g. high and low ability groups or trained and untrained groups). Once the focal comparison is determined, for each SJT item, the response option(s) that produces the largest effect size is scored as the correct answer for the SJT item. In contrast to the most frequent response option technique, the largest effect size technique does not confound item difficulty with item validity. That is, the response option that produces the largest effect size does not necessarily have to be the most frequently endorsed response option.

In the selection context, the validity of SJT scores using the largest effect size technique is function of how strongly the variable used to make the focal comparison (e.g., high ability group versus low ability group) is related to the desired job performance criterion. As long as the developer of the SJT has selected a job-relevant variable upon which to base the focal comparison, the resultant validity should be strong. Of course, it could be asked why the variable used for the focal comparison is not being used as the predictor instead of the SJT scores.

In the training context, the largest effect size technique is inappropriate, assuming the focal comparison is untrained versus trained. The logic of the largest effect size technique presumes that the focal variable distinguishes the respondents on the desired job criterion / criteria. That is, to use the largest effect size technique, it must be presumed that the training program is effective! Clearly, such a presumption is inappropriate in the context of training evaluation.

Combination technique. The third technique combines the most frequent response option and the largest effect size techniques such that only those SJT items with response option that are both the most frequently endorsed option and the option that produces the largest effect size are retained. SJT items that do not meet both criteria are not included in the scoring protocol. On the surface, it may appear that the combination approach overcomes the limitations of the most frequently endorsed technique and the largest effect size technique. Closer examination indicates that this is not the case. The combination technique does not disentangle item difficulty and item validity, nor does it eliminate the logical flaw of presuming training is effective associated with the largest effect size technique. As such, the combination technique cannot be used to develop an SJT scoring protocol in the training evaluation context.

Hybrid Scoring Keys

Hybrid scoring keys combine two independently created scoring keys (Bergman *et al.*, 2006). In Bergman *et al*. (2006) (the only published study utilizing a hybrid scoring key for a SJT), the authors created three hybrid scoring keys for a video-based SJT used to assess leadership skills. Each hybrid scoring key was created by additively combining an empirical scoring key with a leadership theory-based scoring key. The theory based scoring key was a scoring key that was based on a specific leadership theory. The additive key combined the scores for each response option, allowing the positive scores on one key to cancel out a negative or zero score for the same response option on the other key (Bergman *et al*., 2006). Another method of creating hybrid keys mentioned in Bergman *et al*. (2006) is the substitution of zeros. One key is designated as the primary key and this primary key is the basis for the hybrid key. The secondary key is then used to replace the zero scores in the primary key (Bergman *et al.*, 2006). A third method mentioned in Bergman *et al*. (2006) is the differential weighting method, where one key is based on full scores and the other key is fractionally weighted. The authors did not elaborate on the details of the possible methods for hybrid key creation and they were extremely vague in their hybrid key description. The bottom line is that there is not a universally-accepted protocol for creating a hybrid scoring key, for scoring key creation is dependent on a test's purpose and practical constraints. The common goal across all of these hybrid scoring keys is to incorporate the strengths of an individual scoring key methodology, while eliminating or substantially reducing the related weaknesses, in order to create a superior scoring key.

A hybrid key would be especially beneficial for the scoring of a SJT designed to evaluate a training program. The ideal scenario for a hybrid SJT scoring protocol in the training context might appear to include the availability of criterion data for trainees. With criterion data, a hybrid scoring key could be based on SME's judgments combined with knowledge of how SJT items differentiate between trainees on the job. However, if the criterion data for trainees are available, it is not clear why the SJT is needed for training evaluation. The underlying logic of using an SJT for training evaluation is that the SJT simulates the work performance.

In the training context, the hybrid key for a SJT would combine SME's judgments with a respondent-based scoring protocol. Such a hybrid key would be useful because it would allow evaluators to gain insights into both the performance of the trainees and the training program. As explained earlier, the SME-based scoring key is, by itself, limited because it relies solely on human rationale which can be biased, inaccurate, and/or limited. It also may lead to the exclusion of SJT items which are informative of training but do not receive the level of consensus necessary when administered to SMEs. On the other hand, respondent-based keys are limited because the keys do not have a rational base and can lead to scoring keys that are difficult to interpret and will likely over-predict training effectiveness. Both SME and respondent-based keys may also lead to the exclusion of informative items for training evaluation, due to lack of SME consensus or an observed positive effect. An optimal key to be used in training evaluation would combine the rationale provided by SME-based keys and the data provided by respondent-based keys. Items that receive the same treatment across the SME and respondent-based keys are considered to be treated correctly as they have separately reached the prescribed criteria of both scoring keys. However, items that show a discrepancy between the SME and respondent-based keys require further investigation. Through the systematic investigation of these conflicting items, the hybrid key will ensure the appropriate treatment of these *questionable* items, and ensure the inclusion of SJT items that are informative about training but may not be included using the SME or respondent-based keys individually. As such, the hybrid key is a more comprehensive scoring key, allowing greater confidence to be placed in the interpretation of SJTs designed for training evaluation.

The goal of this study is to describe the creation of a hybrid scoring key and demonstrate the hybrid key's superiority for evaluating training when compared to a SME-based scoring key and a respondent-based scoring key. The hybrid key was created by strategically combining the SME-based key and the respondent-based key. All three keys were created to score an Equal Opportunity Situational Judgment Test which was designed specifically to evaluate a military training program.

Method

Equal Opportunity Advisor (EOA) Training Course

Equal Opportunity Advisors (EOAs) are military personnel that inform and counsel unit commanders regarding issues of equal opportunity and diversity awareness. This position exists in all five branches of the U.S. Armed Forces. Active military personnel can either volunteer or be assigned based on need to become a military EOA in their particular service branch. EOA training is conducted three times a year at the Defense Equal Opportunity Management Institute (DEOMI), located on Patrick Air Force Base in Cocoa Beach, Florida.

The EOA training course was formerly an intensive 15-week program. EOA trainees spent eight hours a day in either classroom lecture or small group discussion sessions. The first 12 weeks of the course was focused on broad equal opportunity and diversity issues that are applicable across all branches of the military. The final three weeks of training covered specific service branch policies and complaint procedures. Recently, starting with the Spring 2006 cohort, the training program was condensed to an intensive 10-week program. The first seven weeks of the course are now focused on broad equal opportunity and diversity issues and the final three weeks of training cover specific service branch policies and complaint procedures. The Equal Opportunity Situational Judgment Test targets the complaint processing aspect of the EOA position and that portion of the training remained constant, during the final three weeks of EOA training, across the 15 and 10 week training courses.

Equal Opportunity (EO) Situational Judgment Test (SJT)

To assess the transfer of EOA training to the work environment, the Virginia Tech Research Team, under the leadership of Neil Hauenstein, Ph.D., developed the Equal Opportunity (EO) Situational Judgment Test (SJT) in June of 2005. The EO SJT contains thirty hypothetical problem scenarios EOAs in the field will likely encounter. Each scenario is followed by a question prompt and three response options.

The first step in the development of the EO SJT was to review job descriptions for the EOA position in the different branches of the military. Both a generic EOA job description, applicable to all branches of the military, and a service-specific job description of the EOA position, was reviewed from the Air Force Occupational Measurement Squadron. After examining the critical duties, it was decided that the initial focus of the EO SJT would be in reference to "Complaint Processing". Records regarding both formal and informal EO complaints were readily available and provided a rich source of data from which to create scenarios for the SJT. Examination of the complaints archive indicated that the majority of complaints revolved around race, sex, religious affiliation, and extremist group activity. To establish the content validity of the scenarios, the researchers ensured adequate representation of demographic groups, formal and informal complaints, as well as all phases of the complaint process (i.e., from intake to resolution and follow-up).

The SJT scenarios were constructed by selecting a demographic group and a core incident from the archive of complaints. Next, a series of scenarios was created based on the core incident to reflect the different complaint processing tasks required of an EOA. In the end, 150 scenarios were created. From the 150 scenarios, a subset of 60 representative scenarios was vetted for realism by 16 SMEs via an online survey. All SMEs were instructors in the EOA training program. Fourteen of the SMEs (instructors) were former EOAs in the field/fleet. The SMEs represented all branches of the military and all SMEs had extensive experience with both the EOA job requirements and the EOA training curriculum. There was general agreement regarding acceptable realism of 40 of the 60 scenarios that were vetted. For each of these 40 scenarios, utilizing an online survey, the SMEs were asked to select the most appropriate question prompt (e.g., "The course of action you would take next is"). After choosing the most appropriate question prompt, the SMEs were asked to compose what he/she believed to be the best response to the situation. For each scenario, the SME responses to the online survey were reviewed and adapted by test developers to create three separate response options.

Utilizing this SME feedback, the test developers created the first version of the EO SJT. The first version of the EO SJT contained 30 of the 40 SJT items. A second version of the EO SJT was implemented in November 2005, in which ten of the SJT items included in the first version were replaced with ten new SJT items. The second version of the EO SJT contains 30 SJT items, 20 items which are identical to those included in the first version of the EO SJT.

Scoring Key Development

Method

Participants

Participants were 46 EOA trainees from the Fall 2005 training cohort, and 21 EOA trainers. All of the trainees included in this study completed the EO SJT both before and after completing the EOA training course. Trainees who did not complete both the pre and post SJT were not considered in this study. All of the trainers, who participated as SMEs, had in addition to their experience as an EOA trainer, extensive EOA field experience. All of the participants, trainees and trainers, were diverse in respect to race, sex, military rank, and representative of all five branches of the military; Army, Navy, Coast Guard, Marines, and the Air Force.

Measures

Equal Opportunity (EO) Situational Judgment Test (SJT). The EO SJT, as detailed above, contains thirty hypothetical problem scenarios EOAs in the field will likely encounter. Each scenario is followed by a question prompt and three response options. Only the responses to the common items presented on version 1 and 2 of the EO SJT will be considered in this study. Common EO-SJT items are presented in Appendix A.

Questionable Item Questionnaire. The *Questionable* Item Questionnaire (Appendix B) is a 13-item open-ended survey created by the author. It was designed to determine, in a uniform manner, if the material covered by items on the EO-SJT was 1) covered in the EOA training program and 2) if trainees completing the course should know the material in order to successful complete their duties as an EOA and 3) if items were yielding discrepancies due to specific item issues (e.g. semantics, item clarity, item difficulty, etc.).

Procedure

SME-Based Scoring Key. To create the SME-based scoring key, 21 EOA trainers were selected by DEOMI's Director of Research based on trainer availability, past experience as an EOA, and performance as a current trainer. The selected EOA trainers were solicited in person and via email to act as SMEs and complete an online version of the EO SJT. This online EO SJT included all items presented on the first and second versions of the EO SJT, however, for the purposes of this study, only the 20 common items shared on version 1 and 2 were considered. For each presented SJT item, the SMEs were asked to select the most appropriate response option from the three presented choices.

To create the SME-based scoring key, the percentage of SME endorsement was calculated for each response option, for each SJT item. There were 20 SJT items under consideration, and each SJT item has three response options. Hence, a total of 60 SME endorsement percentages were calculated.

Using the SME endorsement percentages, the ‘best’ response option was chosen for each SJT item. Specifically, the ‘best’ SJT response option was the response option that 70% or more of the SMEs endorse as the ‘best’ response option. If for any SJT item, 70% or more of the SMEs did not endorse one of the presented response options, the SJT item was flagged for lack of consensus, and was not scored in the key. That is, unless, the SMEs equally endorsed two of the presented response options at a percentage greater than or equal to 45% for both response options. In this circumstance, both equally endorsed response options were considered to be ‘best’ response options.

Using these decision rules, the SJT items were separated into two groups: scored and non-scored SJT items. Twelve SJT items were scored based on SME agreement as to the ‘best’ response option. Eight SJT items were not scored due to lack of SME consensus on the ‘best’ response option. Table 1 shows which SJT items were scored and non-scored using the SME-based scoring key.

To score the SJT using the SME–based key, respondents received 1 point for selecting the ‘best’ response option for each scored SJT item. The total points earned for selecting the ‘best’ response option were summed to create a raw score. This raw score was then divided by twelve (the total number of SJT items that were scored) to create a percentage score.

Respondent-Based Scoring Key. To create the respondent-based key, the author utilized the pre and post EO SJT responses for the Fall 2005 EOA trainee cohort. The Fall 2005 cohort completed a version of the online EO SJT on two occasions. The trainees completed the EO SJT version 1 in August of 2005, prior to beginning EOA training (pretest). The trainees then completed the EO SJT version 2 in November of 2005 after completing the EOA core training program (post-test). Versions 1 and 2 of the EO SJT share 20 identical items; therefore, as with the SME-based key, only the responses to the 20 identical SJT items were considered to create the respondent-based scoring key.

Utilizing the pre and post SJT results, effect sizes, in the form of odds ratios, were calculated for each response option. An odds ratio was calculated by dividing the odds of trainee response option endorsement in the post-test by the odds of trainee response option endorsement in the pre-test. For example, if for a given response option, 43 out of 46 trainees endorsed the option in the post-test, the odds of post-test endorsement for this option are 43 to 3, or 14.3 : 1 = 14.3. If for the same response option, only 27 of 46 trainees endorsed the option in the pre-test, the odds of pre-test endorsement for this option are to 27 to 19; or 1.42 : 1 = 1.42. Since 14.3 / 1.42 = 10, 10 is the odds ratio. Thus, the post-test response option has 10 times the odds of trainee endorsement than the same response option in the pre-test. An odds ratio was calculated for each of the SJT items' response options (n = 60). The odds ratio was utilized to determine if a response option exhibited a positive training effect by having a greater likelihood of trainee endorsement in the post-test.

For each SJT item, the combination technique was used to determine if a SJT item had one 'best' response option. The 'best' response option had to have an observed odds ratio greater than or equal to 1.50 AND receive the highest trainee endorsement in the SJT post-test. Odds ratios greater than 1 indicate that a response option is more likely to be endorsed by trainees in the post-test than in the pre-test. Considering the sample size used to create this key (n = 46), an odds ratios greater than 1.5 indicated a response option was more likely to be endorsed by at least 10% more trainees in the post-test than in the pre-test. The ten percent increase was assumed the appropriate minimum increase necessary to assume a positive training effect for an item in this study. Requiring the highest trainee endorsement ensured that only one response option can be considered the ‘best’ option for any SJT item. Items that had no response options meeting the criteria detailed above were flagged, and were not scored in the scoring key.

Using these decision rules, the SJT items were separated into two groups: scored and non-scored SJT items. Fourteen SJT items were scored based on the respondent-based key criteria. Six SJT items were not scored because these items did not meet the respondent-based key criteria. Table 2 shows which SJT items were scored and non-scored using the respondent-based scoring key. To score the SJT using the respondent–based key, respondents received 1 point for selecting the ‘best’ response option for each included SJT item. The total points earned for selecting the ‘best’ response option were summed to create a raw score. This raw score was then divided by fourteen (the total number of SJT items that were scored) to create a percentage score.

Hybrid Scoring Key. To create the hybrid scoring key, the SME and respondent-based scoring keys were revisited. The twenty SJT items were separated into four categories based on the following criteria. Items that reached SME-consensus, based on the decision rules provided for the SME-based key, AND met the criteria for inclusion in the respondent-based key were placed in Category A. Items that were flagged for lack of SME consensus, based on the decision rules provided for the SME-based key, BUT met the criteria for inclusion in the respondent-based key, were placed in Category B. Items that reached SME-consensus, based on the decision rules provided for the SME-based key, BUT did not meet the criteria for inclusion in the respondent-based key, were placed in Category C. Items that were flagged for lack of SME consensus, based on the decision rules provided for the SME-based key, AND did not meet the criteria for inclusion in the respondent-based key, were placed in Category D. Figure 1 shows the relationship of SME and respondent-based keys (effect and SME consensus) for each category. Table 3 shows how many SJT items fell into each category. Table 4 shows into which category each SJT was classified for the hybrid key.

Seven SJT items were classified into Category A. Items in Category A were *scored* items in both the SME and respondent-based keys. These items were scored identically in the hybrid key as in the SME and respondent-based scoring keys. Two SJT items were classified into Category D. Items in Category D were *non-scored* items, as they did not meet the criteria outlined in either the SME-based or respondent based keys. These items were not scored in the hybrid key.

Six SJT items were classified into Category B and five SJT items were classified into Category C. Items in Categories B and C were considered *questionable* items, as they have met the criteria outlined in either the SME-based key OR the respondent-based key. These items underwent further investigation to determine the cause of the observed discrepancy between the two scoring keys and if, and how, they should be scored in the hybrid key. To make this determination, data was obtained via two SME focus groups held at DEOMI during July of 2006.

The first SME focus group was approximately three hours in length and had seven SMEs participants. The second SME focus group was approximately three hours in length and had two SME participants. Both focus groups discussed the same SJT items. Each SME focus group was facilitated by the author, who adhered to the following protocol. First, for each SJT item, that SJT scenario, question prompt, and corresponding response options were presented by overhead projector, and read aloud by the author. Following the presentation of each item, SMEs were solicited for feedback using the *Questionable* Item Questionnaire (Appendix A). This instrument was utilized to ensure that each item was investigated in a uniform manner. The author verbally posed each question from the *Questionable* Item Questionnaire, and the SMEs verbally responded until a group consensus was reached for each question presented. The author documented SME responses for each item. This method was repeated for each *questionable* item.

Based upon the SME feedback received in these focus groups, the decision was reached whether the *questionable* items in Categories B and C were appropriate for inclusion in the scoring key. An item was deemed appropriate for inclusion in the scoring key if 1) the material covered in the item is also covered in the EOA training program, 2) the trainees completing the course should know the material by the end of training in order to successfully complete their EOA responsibilities, 3) the scenario, question prompt, and response options are clear, realistic, and applicable, and 4) the SMEs can select and justify one of the response options as superior to the other response options. A 'best' response option was chosen for each item meeting the above criteria. These items were then included in the hybrid scoring key.

Eighteen SJT items were ultimately scored based on this hybrid protocol. Two SJT items remained non-scored items as prescribed by both the SME and the respondent-based keys. Table 5 shows which SJT items were scored and non-scored using the hybrid scoring key. To score the SJT using the hybrid key, respondents received 1 point for selecting the 'best' response option for each included SJT item. The total points earned for selecting the 'best' response option were summed to create a raw score. This raw score was then divided by eighteen (the total number of SJT items that had been scored) to create a percentage score.

Comparison of Questionable Item Treatment Across Keys. Eleven items were classified into Categories B and C as *questionable* items, meaning that there was a discrepancy between the SME and respondent-based scoring keys. Table 6 shows the scored/non-scored treatment of *questionable* items across the three scoring keys. The SME and respondent-based scoring keys scored only a portion of the SJT items as prescribed by the SME focus groups and the hybrid scoring key. Specifically, using the SME-based scoring key ten out of the twenty SJT items were scored as determined best in the hybrid scoring key. Six of these SJT items were not scored at all in the SME-based scoring key due to lack of SME consensus. The remaining four SJT items were scored differently on the SME-based key than later prescribed by the SME focus groups. Using the respondent-based scoring key fourteen out of the twenty SJT items were scored as determined best in the hybrid scoring key. Four of these items were not scored in the respondent-based scoring key because the respondent-based scoring key criteria were not satisfied. The remaining two items were scored differently than later prescribed by the SME focus groups. Table 7 displays the items that were incorrectly scored when utilizing the SME and the respondent-based scoring keys. The hybrid scoring key scored all eleven *questionable* items presented in Categories B and C based on the insight gained in the SME focus groups, thus, all eleven *questionable* items were appropriately scored using the hybrid key.

Training Evaluation Study

Method

Participants

Participants were 55 EOA trainees from the Spring 2006 training cohort. All of the participants were diverse in respect to race, sex, military rank, and representative of all five branches of the military; Army, Navy, Coast Guard, Marines, and the Air Force.

Measures

Equal Opportunity (EO) Situational Judgment Test (SJT). The EO SJT, as detailed above, contains thirty hypothetical problem scenarios EOAs in the field will likely encounter. Each scenario is followed by a question prompt and three response options. Only the responses to the common items presented on version 1 and 2 of the EO SJT will be considered in this study. Common EO-SJT items are presented in Appendix A.

SJT Scoring Keys. The SME-based scoring key, the respondent-based scoring key, and the hybrid scoring key, as detailed above, were created by the author to score the EO SJT. Each key represents a unique methodology for scoring a SJT designed to evaluate a training program.

Procedure

The EO SJT version 2 was administered in a paper-pencil format to 55 EOA trainees from the Spring 2006 cohort on three occasions, once before training (pre-test), once after core DEOMI training (mid-term), and once after service specific training (post-test).

Results

Comparison of SJT Scoring Keys

Utilizing the SME-based scoring key, the Spring 2006 EOA trainee cohort post test results (M = 8.53, SD = 1.71) were significantly higher than the Spring 2006 EOA trainee cohort pre test results (M = 7.71, SD = 1.63), t(54) = -3.34, p <.01 (two-tailed), d = .49. The 95% confidence interval for t(54)= -3.34 is -5.30 to -1.38. According to Cohen (1988), the effect size of .49 can be considered a medium effect size. Table 8 presents the paired sample *t*-test statistics for the Spring 2006 EOA trainee cohort utilizing the SME-based scoring key.

Utilizing the respondent-based scoring key, the Spring 2006 EOA trainee cohort post test results (M = 8.93, SD = 2.28) were significantly higher than the Spring 2006 EOA trainee cohort pre test results (M = 7.00, SD = 2.18), t(54) = -6.49, p = < .01 (two-tailed), d = .87. The 95% confidence interval for t(54)= -6.49 is -8.45 to -4.53. According to Cohen (1988), the effect size of .87 can be considered a large effect size. Table 9 presents the paired sample *t*-test statistics for the Spring 2006 EOA trainee cohort utilizing the respondent-based scoring key.

Utilizing the hybrid scoring key, the Spring 2006 EOA trainee cohort post test results (M = 9.13, SD = 1.87) were significantly higher than the Spring 2006 EOA trainee cohort pre test results (M = 8.11, SD = 2.09), t(54) = -3.62, p = < .01 (two-tailed), d = .52. The confidence 95% interval for t(54)= -3.62 is -5.58 to -1.66. According to Cohen (1988), the effect size of .52 can be considered a medium effect size. Table 10 presents the paired sample *t*-test statistics for the Spring 2006 EOA trainee cohort utilizing the hybrid scoring key.

Respondent-Based Scoring Key Cross-Validation.

Stability of effects is an issue in the respondent-based scoring key. To examine this issue, the Spring 2006 effect sizes were compared to the Fall 2005 effect sizes. The pre and post-test responses from the Spring 2006 EOA trainee cohort was utilized to calculate effect sizes, in the form of odds ratios. Results from the Spring 2006 EOA cohort show that eighteen of the twenty SJT items (90%) exhibited the same effect size pattern exhibited by the Fall 2005 cohort. Specifically, for each of these eighteen SJT items, it was observed that the same response options received the highest trainee endorsement, and greater endorsement in the post-test, utilizing both the Fall and Spring cohort data. Table 11 shows response option endorsement and corresponding effect sizes for both the Fall and Spring EOA trainee cohorts.

Two items did not demonstrate pattern stability across the Fall 2005 and Spring 2006 cohorts. These were SJT items 14 and18. Item 14 was not scored in the respondent-based scoring key because although response option 1 received the greatest trainee endorsement in the post test (n = 31), the effect size for option 1 did not meet the criteria for the respondent-based key. However, using the Spring 2006 data, Item 14 would be scored on the respondent-based scoring key with response option 2 chosen as the ‘best’ response option (post test endorsement *n* = 27; *OR* = 2.67). Item 18 was also not scored in the respondent-based scoring key for lack of an effect using the Fall 2005 cohort responses. However, utilizing the Spring 2006 cohort responses, Item 18 would be scored on the respondent-based key with response option 1 chosen as the ‘best’ response option (post test endorsement *n* = 31; *OR* = 2.09)

Utilizing the Spring 2006 EOA cohort data (instead of the Fall 2005 EOA cohort data) to create the respondent-based key would have an insignificant impact on the hybrid scoring key. Item 14, initially excluded from the respondent-based and the hybrid key, would be re-classified as a *questionable* item in the hybrid key creation requiring further investigation. Item 18 would still be considered a *questionable* item in the hybrid key creation regardless of which EOA cohort data was utilized.

Discussion

The present study highlights the utility of SJTs as a viable methodology for evaluating training transfer. SJTs provide an affordable means of empirically linking training and performance (i.e. the impact of training to outcomes). There was one difficulty hindering the use of SJTs for training evaluation which needed to be addressed before a SJT could be utilized to evaluate a training program. The obstacle was creating an unbiased and accurate key to use to score the SJT. The objective of this study was to overcome the scoring key challenge by developing an innovative scoring key methodology that addresses the specific needs of SJTs designed for training evaluation. To address this issue, the hybrid scoring key was created and applied to a SJT designed to evaluate the effectiveness of a military training program. The logical justification and intuitive creation of the hybrid key was described in detail.

The utility and overall benefit of using the hybrid key when evaluating training with a SJT was further demonstrated when the hybrid key was compared with two traditional SJT scoring keys, the SME and respondent-based scoring keys. The SME-based scoring key, when applied to the SJT responses of the EOA training cohort, exhibited the smallest training effect. This finding could be interpreted to mean that EOA training will have a moderate impact on later trainee performance. On the other hand, the respondent-based key exhibited the largest training effect. This large effect could be interpreted to mean that the EOA training will have a large impact on later trainee performance. This illustrates how two scoring keys can deliver results that differ and can lead to different interpretations. Based on the hybrid key, the training was moderately successful. The increased confidence in the accuracy of the hybrid key, due to the stringent methodology upon which it was created, allows increased confidence in the conclusions drawn from the use of the hybrid key.

The hybrid key exhibited a medium effect only slightly larger than that exhibited by the SME-based key. It could be noted that while it is clear that the respondent-based scoring key overestimated the effectiveness of the training program, there is not a substantial difference between using the SME-based or the hybrid scoring key. The response to this contention is that both scoring keys demonstrate that the training has a moderate effect on performance. However, the key issue is that the researcher has more confidence in the conclusion of moderate training effectiveness based on the hybrid key than the SME key because diagnostic items excluded by the SME protocol have been included in the hybrid protocol.

Diagnosis of Training Problems

The trainee responses to the SJT can be utilized to diagnose areas in training that need improvement based on items that exhibit poor trainee performance. Due to the fact that the hybrid scoring key allows the inclusion of more of the SJT items that are informative about training into the scoring of the SJT, the hybrid key allows for additional insights to be obtained regarding areas where trainees are not benefiting from training. The SME-based key excluded six SJT items that were concluded to be informative about training, thereby not allowing investigation into these items and the areas they represent. The respondent-based key excluded four items that were informative about training in addition to overestimating the effectiveness of the training program. The hybrid key scored 18 out of the 20 SJT items, excluding only two items while accurately estimating the effectiveness of the program. The additional items scored on the hybrid key not only ensure the accuracy of the overall finding, it enables the identification of specific areas that require improvement in a training program.

Overall the SJT should be regarded as a valid and useful tool for procuring empirical evidence of effectiveness and, thus, evaluating training. Moreover, it should be understood that

SJTs are only as valid as the accompanying scoring key. Increasing the soundness of the scoring key ensures that the SJT can be successfully utilized to evaluate a training program. Increased confidence is placed in the hybrid scoring key because the hybrid key incorporates both rational and empirical information in a logical and organized fashion. The hybrid key methodology allows the combination of the independent strengths of traditional scoring keys while avoiding the possible weaknesses posed by utilizing traditional keys. Unlike the respondent-based key, the hybrid key is protected against the overestimation of training, and unlike the SME-based key, the hybrid key does not exclude SJT items which are informative of training. For these reasons, the hybrid key's superiority was illustrated over the SME and respondent-based scoring keys when utilized for the purpose of training evaluation.

In addition to reporting a technique that has promise for practitioners and test developers, the current study contributes to the SJT literature by investigating which SJT scoring strategy is best and under what conditions this assertion will hold true. This study has contributed to the research by identifying the hybrid scoring key as the most appropriate SJT scoring key methodology when used for the purpose of training evaluation. This study also adds to the training evaluation literature by demonstrating an affordable means of empirically evaluating training transfer. This study shows that SJTs, while regularly utilized in practice for the purpose of selection, can also be useful in the training evaluation context, when a valid scoring key is applied. This is important, as connecting the process of training to later performance outcomes is a difficult relationship to capture empirically, and the hybrid scoring key is allowing SJTs to be utilized to accomplish that feat.

In addition to linking effective training to specific outcomes, the hybrid scoring key also allows the SJT to be used to 1) legitimize training initiatives if found effective 2) allow flaws in training to be uncovered if training is found ineffective 3) allows flaws in training to be corrected and improved when training is found ineffective. The information obtained through SJTs designed for training evaluation is valuable to the organization and shareholders because it provides evidence of the utility and effectiveness of a training program. This evidence can then be utilized to make a case for future program funding. This is true for any training program that is the subsidiary of a larger organization and depends upon the larger organization for funding/support. The evidence also allows the organization to confidently invest in a training program, once its effectiveness has been determined. For this reason, these findings are of practical importance for training evaluators, coordinators, designers, and curriculum developers.

An example of this is the military EOA training program in which this study was embedded. With governmental budget cuts taking place every quarter, showing the effectiveness, and performance impact, of the EOA training program is essential for continued operational funding. The results are used to legitimize the funding received by the government and used to argue for increased budgets in the future.

Utilizing the hybrid key for SJTs designed for training evaluation also allow training evaluators, coordinators, designers, and curriculum developers to obtain a better understanding of what is effective and what is not effective in current training design. Subject areas on the SJT that show poor trainee performance allow practitioners to identify these as subject areas where training needs to be revised, re-implemented, or removed. For example, the SJT results received for the military EOA training program showed that for certain SJT items the majority of the trainees endorsed the incorrect response option, meaning they did not handle the situation in the optimal manner. After identifying which subject areas were discussed in the SJT scenarios, the training personnel realized that some of the subject areas in question were presented less and differently than usual in their current training program design. The subject areas were important for the EOA position, and therefore for future classes, training in those subject areas has been improved (i.e. will be covered more in-depth to ensure greater trainee skill acquisition). The information regarding areas that need improvement allowed EOA training decision-makers the opportunity to make the necessary changes to the training to impact the later performance of the trainees.

Finally, the hybrid scoring key methodology can be beneficial for SJT developers. Using the hybrid key methodology for creating the SJT scoring key, test developers can rely less on SME input during the creation of the SJT and instead focus on obtaining the vital SME input during the creation of the scoring key. Through the process of completing this study, it is clear that the traditional process for SJT creation can be improved when the SJT is created to evaluate a training program. Simply explained, traditional job knowledge tests are developed based on the training and curriculum (i.e. the process). SJTs are developed based upon on-the-job performance (i.e. the outcome). Traditionally, to determine the critical dimensions of on-the-job performance and ensure construct validity of the instrument, test developers were dependent upon SME input at several stages. SME input was necessary for subject area identification, scenario creation and vetting, and response option creation and vetting. This is a labor-intensive process that requires several iterations before a final SJT can be created. A SJT designed to evaluate training can be created utilizing information provided during job analysis and critical information regarding job requirements and expectations. Thus once, the developer has an understanding of the job requirements, it is possible for the developer to write the SJT items and response options independently of SMEs. Instead of the SMEs vetting the SJT items at every stage of development, the items can be critiqued during the SME focus groups utilized to develop the hybrid scoring key. Utilizing less SME input in the SJT development stages, can lead to an instrument that is developed faster and is less labor intensive. SME focus groups held during the hybrid key creation will identify items that are flawed and not informative about training. While some SME insight will still be required during the construction of any SJT, utilizing the hybrid methodology allows developers more autonomy the initial SJT development stages, and allows the SME attention to be focused at evaluating the end result, the overall SJT, not the individual SJT components.

Hybrid Key Considerations

A consideration when creating a hybrid scoring key is that both SME and respondent data are needed. These are vital resources which are not usually within the locus of control of the test developer. Therefore, the test developer is usually dependent upon third parties for access to these resources. Limited or no access to SMEs and/or respondent data, will bar the test developer from creating a hybrid scoring key. It is critical that test developers recognize and plan accordingly beforehand regarding the additional requirements necessary for the creation of the hybrid scoring key.

Another consideration when creating a hybrid key is the inherent nature of the hybrid key to assume the flaws present in the ‘parent’ scoring keys. For that reason, test developers must make great efforts to control the level of inaccuracy that is introduced into the multi-stage creation process of the hybrid scoring key. The scoring keys utilized to create the hybrid key, the SME and respondent-based keys, must be prudently developed and assiduously maintained, as errors in the creation of either key will be manifested in the hybrid key.

For example, the stability of the respondent-based scoring key will affect the stability of the hybrid scoring key, which in turn will have an impact on the validity and reliability of the hybrid scoring key. To ensure the stability of both scoring keys, both keys must be carefully developed and the effect sizes utilized to create both keys must be checked regularly. This study confirmed that the respondent-based key was stable across SJT administrations to Fall and Spring cohorts, by comparing the effect sizes of both cohorts. Only two SJT items exhibited a different effect size pattern than that observed for the first cohort. Some of the factors that may have had an impact on the effect size stability of these two items are 1) items may have been compromised/corrupted over time, 2) items may be inherently unstable, 3) and training and/or trainees may vary drastically across administrations of the test. Test developers should set aside time to investigate the cause of items' effect size deterioration over time as this will provide additional insight into whether items should be retired from a SJT or whether the impact of training has truly changed over time.

Study Limitations

In this study criterion data was not available, as such a respondent-based key was utilized in place of an empirical scoring key to create the hybrid scoring key. In the most optimal situation criterion data would have been available to create an empirical scoring key to use in place of the respondent-based key. It is the case, often in practice, that criterion data are not readily available, and thus, an empirical key was not an option. This study, while not relying on criterion data, did allow the example to be presented of what can be done when criterion data is not available. One cannot deny, however, that utilizing criterion data would allow the classification of high and low performers in the empirical key, versus accepting the assumption that trained and untrained personnel (i.e. pre and post trainees) represent high and low performers respectively in the respondent-based key.

As a next step in understanding the scoring protocols of SJTs used for training evaluation, it would useful to have access to criterion data on the trainees. As mentioned earlier, if criterion data are available on trainees, it is not necessary to administer the SJT for training evaluation because the SJT is simulating work performance. That being said, access to criterion data would allow for external validation for each of the scoring protocols. Such validation evidence could further delineate the best characteristics of the different scoring protocols. For this study, no criterion data has been gathered to measure EOA performance in the field. The SJT is the first attempt at examining proficiency of EOA graduates and the effectiveness of the EOA training program. The fact is that SJTs are most likely to be utilized in situations when valid criterion data is extremely difficult or impossible to obtain. While having this information would be beneficial to validate the overall effectiveness and accuracy of the scoring key, practically speaking, this may be an unattainable goal when attempting to measure performance on constructs that are confounded or immeasurable.

Finally, the utility of the hybrid scoring approach could be even more convincing if there were a greater number of SJT items and respondents in question. While the benefits are obvious regardless of trainee sample size or item pool size, one might question the cost-benefits of utilizing the hybrid scoring key in the current study. The analysis of cost-benefits would be particularly apparent if the SJT item bank was larger than the 20 items utilized in this study. With large SJT item pools, detailed investigation into each SJT item would be impractical and inefficient. The hybrid scoring approach allows investigation into the appropriate subset of the items (i.e. questionable items), which results in the most accurate, practical, and efficient scoring approach. While this study hopefully conveyed the utility of this method, having more SJT items and more respondents may have further highlighted this important point.

Future Directions

Although only one type of hybrid key was created in this project, it is clear that other hybrid keys could be created (by combining different scoring keys and developing different scoring key criteria) depending on the purpose of the SJT and contextual restraints. It is the author's belief based on this and past research that the effect will remain consistent with the SME having the smallest effect, the empirical/respondent-based having the largest, and the hybrid lying in the middle, since it is a combination of the two parent scoring protocols. It is also the author's belief that the results would confirm the hybrid scoring key as superior when utilized for training evaluation. It would be interesting to more specifically define the boundaries of training evaluation in which the hybrid key will be superior (i.e. add specific types of training evaluation) and see if there are any different outcomes depending on the type of evaluation in question. It would also be interesting to investigate the additional contexts in which the hybrid scoring methodology would remain superior to traditional scoring protocols. Additionally, obtaining criterion data to further support the accuracy of the SJT, the hybrid scoring key, and the overall findings regarding training impact and effectiveness would be significant contributions to this study.

Considering the role that time plays on measurement accuracy, it would be a significant research contribution to investigate how hybrid scoring key accuracy and stability is affected over time. Creating new strategies that can be used to measure the effects of time on the hybrid key, such as monitoring stability of effect sizes, and developing ways to diagnose and address the various reasons for observed instability, such as creating a protocol for diagnosing whether observed changes across classes are due to fluctuations in trainee ability, teaching, or over-exposure to test, would be insightful. In the vein of test over-exposure, developing a methodology for creating parallel forms of SJTs, and developing an identification system to alert test developers when SJT items need to be retired, would prove to be an influential line of research that would benefit and inform both science and practice.

References

Arvey, R. D., & Cole, D. A. (1989). Evaluating change due to training. In Goldstein, IL & Associates (Eds.), Training and Development in Organizations (pp. 89-117). San Francisco: Josey-Bass.

Becker, T. (2005). Development and validation of a situational judgment test of employee integrity. International Journal of Selection and Assessment, 13(3), 225-232.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. International Journal of Selection and Assessment, 14, 223-235.

Burnaska, R. F. (1976). The effects of behavioral modeling training upon managers' behaviors and employees' perceptions. Personnel Psychology, 29, 329-335.

Byham, W. C., Adams, D., & Kiggins, A. (1976). Transfer of modeling training to the job. Personnel Psychology, 29, 345-349.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. Journal of Applied Psychology, 82, 143-159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. Human Performance, 15(3), 233-254.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. Journal of Applied Psychology, 86, 410-417.

Dalessio, A. I. (1994). Predicting insurance agent turnover using a video-based situation judgment test. Journal of Business and Psychology, 9, 23-32.

Gillespie, M. A., Oswald, F. L., Schmitt, N., Manheim, L., & Kim, B. (2002). Validation of a situational judgment test of college student success. Paper presented at the 17th Annual Convention of the Society for Industrial and Organizational Psychology.

Green, P. C., Alter, P., & Carr, A. F. (1993). Development of standard anchors for scoring generic past-behavior questions in structured interviews. International Journal of Selection and Assessment, 1, 203-212.

Hanson, M. A., & Ramos, R. A. (1996). Situational judgment tests. In R. S. Barrett (Ed.), Fair employment strategies in human resource management (pp. 119-124). Westport, CT: Quorum Books/Greenwood Publishing Group, Inc.

Hauenstein, N. M. A., Sinclair, A. L., Robson, V., Quintella, Y., & Donovan, J. J. (2003). Performance dimensionality and the occurrence of ratee race effects. Paper presented at the 18th Annual Conference of the Society for Industrial-Organizational Psychology, Orlando, FL.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. The International Journal of Aviation Psychology, 13(4), 373-386.

Kirkpatrick, D. (1976). Evaluation of training. In R. L. Craig (Ed.), Training and development handbook (2nd ed., pp. 301-319). New York: McGraw-Hill.

Kirkpatrick, D. L. (1978). Evaluating in house training programs. Training and Development Journal 38, 32-37.

Konradt, U., Hertel, G., & Joder, K. (2003). Web-based assessment of call center agents: Development and validation of a computerized instrument. International Journal of Selection and Assessment, 11(2/3), 184-193.

Krokos, K. J., Meade, A. W., Cantwell, A. R., Pond, S. B., & Wilson, M. A. (2004). Empirical keying of situational judgment tests: Rationale and some examples. Paper presented at 19th annual meeting of the Society for Industrial/Organizational Psychology.

Latham, G. P., & Saari, L. M. (1979). Application of social learning theory to training supervisors through behavior modeling. Journal of Applied Psychology, 64, 239-246.

Legree, P. J., Psotka, J., & Tremble, T. (2005). Using consensus based measurement to assess emotional intelligence. Ashland, OH: Hogrefe and Huber Publishers.

Lievens, F. (2000). Development of an empirical scoring scheme for situational inventories. European Review of Applied Psychology, 50, 117-124.

Lievens, F. (2005). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), Situational Judgment Tests (Vol. SIOP Frontier Series).

MacCann, C., Roberts, R. D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance based emotional intelligence tests. Personality and Individual Differences, 36(3), 645-662.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: a clarification of the literature. Journal of Applied Psychology, 86(4), 730-740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of Practice and Constructs Assessed. International Journal of Selection and Assessment, 9(1/2), 103-113.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. J. Rumsey, C. D. Walker & J. Harris (Eds.), Personnel Selection and Classification Research (pp. 193-222). Mahwah, NJ: Lawrence Erlbaum Publishers.

Morath, R., Curtin, P., Brownstein, E., & Christopher, C. (2004). Situational judgment tests: Recent innovations in development and scoring: Caliber Associates.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. Journal of Applied Psychology, 75(6), 640-647.

Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form a situational inventory. Journal of Occupational and Organizational Psychology, 66, 337-344.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. International Journal of Selection and Assessment, 13(4), 250-260.

Ostroff, C. (1991). Training effectiveness measures and scoring schemes: A comparison. Personnel Psychology, 44(2), 353-374.

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. Organizational Research Methods, 8(2), 149-164.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. Journal of Applied Psychology, 89(2), 187-207.

Paullin, C. J. (2003). *Investigating the impact on validity of applying empirical scoring procedures to rationally derived inventories.* Unpublished doctoral dissertation, University of Minnesota.

Reiter-Palmon, R., & Connelly, M. S. (2000). Item selection counts: A comparison of empirical key and rational scale validities in theory-based and non-theory based item pools. Journal of Applied Psychology, 85(1), 143-151.

Salas, E., Milham, L. M., & Bowers, C. A. (2003). Training evaluation in the military: Misconceptions, Opportunities, and Challenges. Military Psychology, 15(1), 3-16.

Schmitt, N., Noe, R. A., & Ostroff, C. (1986). Evaluation of the Springfield development project: Unpublished report prepared for the National Association of Secondary School Principals. Reston, VA.

Sorcher, M., & Spence, R. (1982). The interface project: Behavior modeling as social technology in South Africa. Personnel Psychology, 35, 557-581.

Waugh, G. W., & Russell, T. L. (2004). Comparison of situational judgment tests formats, scoring key developers and scoring algorithms: Human Resources Research Organization (HumRRO).

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. Personnel Psychology, 50(1), 25-49.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: antecedents and relationships with performance. Human Performance, 18(1), 81-104.

Table 1

SME-Based Key Item Treatment

	Item Treatment	
SJT item #	Scored	Non-Scored
1	X	
2	X	
3	X	
4	X	
5		X
6		X
7	X	
8		X
9	X	
10	X	
11	X	
12	X	
13		X
14		X
15		X
16		X
17		X
18	X	
19	X	
20	X	
Total	12	8

Table 2

Respondent-Based Key Item Treatment

	Item Treatment	
SJT item #	Scored	Non-Scored
1	X	
2	X	
3	X	
4	X	
5	X	
6	X	
7		X
8	X	
9	X	
10	X	
11		X
12	X	
13		X
14		X
15	X	
16	X	
17	X	
18		X
19		X
20	X	
Total	14	6

Table 3

SJT Items Per Category

Category	# of SJT Items
A	7
B	6
C	5
D	2

Table 4

Item Categorical Classification for Hybrid Key

	Category			
SJT Item #	A	B	C	D
1	X			
2	X			
3	X			
4	X			
5		X		
6		X		
7			X	
8		X		
9	X			
10	X			
11			X	
12			X	
13				X
14				X
15		X		
16		X		
17		X		
18			X	
19			X	
20	X			
Total	7	6	5	2

Note. X signifies into which category the SJT item was placed.

Table 5

Hybrid Scoring Key Item Treatment

	Item Treatment	
SJT item #	Scored	Non-Scored
1	X	
2	X	
3	X	
4	X	
5	X	
6	X	
7	X	
8	X	
9	X	
10	X	
11	X	
12	X	
13		X
14		X
15	X	
16	X	
17	X	
18	X	
19	X	
20	X	
Total	18	2

Table 6

Treatment Across Keys of Questionable Items (Scored/Non-scored)

	Scoring Key		
Questionable Items	SME	Respondent	Hybrid
5		X	X
6		X	X
7	X		X
8		X	X
11	X		X
12	X	X	X
15		X	X
16		X	X
17		X	X
18	X		X
19	X		X
Total	5	7	11

Note. X signifies that the item was scored.

Table 7

Incorrectly Scored SJT Items

	Scoring Key	
SJT item #	SME	Respondent
1		
2		
3		
4		
5	X	
6	X	
7	X	X
8	X	
9		
10		
11	X	X
12	X	X
13		
14		
15	X	
16	X	
17	X	X
18		X
19	X	X
20		
Total	10	6

Note. X signifies that the item was incorrectly scored.

Table 8

Paired Sample t-test Using SME-Based Scoring Key

		Paired Differences							
					95% CI				
		M	*SD*	*SEM*	Lower	Upper	*t*	*df*	Sig (2-tailed)
Pair	PreScore-PostScore	-.82	1.82	.25	-1.31	-.33	-3.34	54	.002

Table 9

Paired Sample t-test Using Respondent-Based Scoring Key

		Paired Differences							
					95% CI				
		M	*SD*	*SEM*	Lower	Upper	*t*	*df*	Sig (2-tailed)
Pair	PreScore-PostScore	-1.93	2.20	.30	-2.52	-1.33	-6.49	54	.000

Table 10

Paired Sample t-test Using Hybrid Scoring Key

		Paired Differences							
					95% CI				
		M	*SD*	*SEM*	Lower	Upper	*t*	*df*	Sig (2-tailed)
Pair	PreScore-PostScore	-1.02	2.09	.28	-1.58	-.45	-3.62	54	.001

Table 11

Fall 2005 and Spring 2006 Response Option Endorsement and Effect Sizes (Odds Ratios)

		Fall 2005			Spring 2006		
		Endorsement			Endorsement		
SJT Item #	Response Options	Post	Pre	Odds Ratio	Post	Pre	Odds Ratio
1	1	2	5	0.37	5	6	0.83
	2	7	11	0.57	15	14	1.13
	3	37	30	2.19	34	35	0.97
2	1	43	27	10.08	46	26	6.41
	2	0	11	0.00	1	18	0.04
	3	3	8	0.33	7	11	0.60
3	1	0	3	0.00	6	1	6.75
	2	44	36	11.00	45	52	0.29
	3	1	6	0.15	3	2	1.56
4	1	20	19	1.09	24	29	0.69
	2	2	12	0.13	6	9	0.63
	3	24	15	2.25	25	17	1.86
5	1	39	24	5.11	35	31	1.38
	2	5	14	0.28	11	15	0.66
	3	2	8	0.22	7	7	1.00
6	1	7	11	0.57	16	19	0.78
	2	25	13	3.02	29	25	1.34
	3	14	22	0.48	10	11	0.89
7	1	14	8	2.08	9	9	1.00
	2	10	11	0.88	7	10	0.66
	3	22	27	0.65	39	36	1.29

		Fall 2005			Spring 2006		
		Endorsement			Endorsement		
SJT	Response			Odds			Odds
Item #	Options	Post	Pre	Ratio	Post	Pre	Ratio
8	1	12	17	0.60	11	16	0.61
	2	10	14	0.63	7	22	0.22
	3	24	15	2.25	37	17	4.59
9	1	1	5	0.18	2	2	1.00
	2	9	10	0.88	6	10	0.55
	3	36	31	1.74	47	43	1.64
10	1	1	2	0.49	2	1	2.00
	2	39	28	3.58	41	38	1.23
	3	6	16	0.28	12	15	0.73
11	1	7	11	0.57	7	4	1.79
	2	23	25	0.84	40	35	1.37
	3	16	10	1.92	8	14	0.47
12	1	28	20	1.94	35	30	1.46
	2	0	1	0.00	0	1	0.00
	3	18	24	0.56	20	24	0.74
13	1	13	13	1.00	19	23	0.75
	2	10	13	0.71	8	10	0.78
	3	23	20	1.30	26	21	1.51
14	1	31	30	1.10	26	38	0.42
	2	10	14	0.63	27	15	2.67
	3	5	2	2.68	1	2	0.50
15	1	13	12	1.12	20	22	0.88
	2	4	10	0.34	13	16	0.77
	3	29	24	1.56	21	17	1.42

		Fall 2005			Spring 2006		
		Endorsement			Endorsement		
SJT Item #	Response Options	Post	Pre	Odds Ratio	Post	Pre	Odds Ratio
16	1	13	12	1.12	15	12	1.34
	2	8	14	0.48	11	14	0.73
	3	25	20	1.55	29	29	1.00
17	1	37	8	19.53	31	19	2.45
	2	6	28	0.10	22	27	0.69
	3	3	10	0.03	2	9	0.19
18	1	21	22	0.92	31	21	2.09
	2	3	1	3.14	4	7	0.54
	3	22	23	0.92	20	27	0.59
19	1	25	22	1.30	37	21	3.33
	2	16	19	0.76	17	23	0.62
	3	5	5	1.00	1	11	0.07
20	1	1	5	0.18	1	1	1.00
	2	26	21	1.55	36	35	1.08
	3	19	20	0.91	18	19	0.92

	SME consensus	**No SME consensus**
Effect	Category A	Category B
No Effect	Category C	Category D

Figure 1. SJT item categories for hybrid scoring key. Items are placed into Category A when there is both an effect and SME consensus. Items are placed into Category B when there is an effect but there is not SME consensus. Items are placed into Category C when there is no effect but there is SME consensus. Items are placed into Category D when there is no effect and no SME consensus.

Appendix A

Common EO-SJT Items

Scenario 1
You receive an e-mail in which two female officers claim to have found threatening notes against female officers. One of the notes supposedly contains a list of female officers. Some of the names on the list had been crossed out, and beside these names were the words “going going gone” and a smiley face. The crossed-out names were female officers who had transferred to other installations. The female officers state that they have asked around, and that they have the impression the female officers left in large part due to intimidation by male officers. The female officers indicate that they are both on the list.
Your first course of action would be:

□ 1 Send a reply to the e-mail asking for more details--Where was the note found? What specific events support the intimidation claim?

□ 2 Inform the commander that there may be a sexual harassment problem at the instillation.

□ 3 Ask the two female officers to come see you in person as soon as possible, and to bring the note with them.

Scenario 2
A female reports to you that she has been receiving harassing text messages on her cell phone from a male. She explains that she exchanged cell phone numbers with the male because he lives close to her off-base, and that she asked him for a ride to the base one day. The female has saved the text messages on her phone and she shows them to you, indicating that all the messages come from the male’s cell phone number. Although there are many messages, none are explicitly sexual in nature, but some messages could be interpreted sexually. The female claims she asked the male not to text message her any more. This is supported by the fact that the number of messages decreased after the day she claims to have told him to stop. The female is calm and not too worried about the behavior, she is just more annoyed that the male has not completely stopped sending messages.
Your first course of action would be:

□ 1 Ask the female what she would like you to do for her.

□ 2 Document the dates, time, and content of all the text messages sent by the male.

□ 3 Inform the female that you will speak with the male and request that he stop all text messaging to the female.

Scenario 3
A Black male has come to you seeking to file a formal complaint against a White male in his unit. The Black male claims that the White male threw a dart close to his head while he was removing his darts from the board. The Black male further claims that when he asked the White male why he threw the dart, the White male said that he was just messing around and wasn’t trying to hurt the Black male. The Black male doesn’t believe the White male. The Black male believes that the White male was trying to intimidate him, and the Black male also says that it is rumored in the unit that this particular White male is a bigot.
Of the following, the best course of action is to:

□ 1 Inform the unit in charge of criminal investigations of a possible case of assault

□ 2 Explain the options of pursuing the incident as either a formal complaint or an informal complaint

□ 3 Informally speak with other unit members to form an opinion about the White male's level of bigotry

Scenario 4
An investigating officer is making an inquiry into a formal complaint from a White male officer that his Black male superior officer treats the minority officers better than he treats his white officers. The White male officer has claimed that the Black male officer shouts more at his White officers, gives out harsher punishments to the White officers, and allows Black officers a greater say in their assignments. The investigating officer has requested your assistance.
Of the following, the assistance you would volunteer to provide the investigating officer first is:

□ 1 Examine unit records and reports to see if the evidence supports the complaint

□ 2 Offer to conduct Cultural Experience exercises for the officers in the unit

□ 3 Provide the investigating officer with a list of suggested questions to ask of both the accused and the other officers in the unit

Scenario 5
A female reports that while seated in the dining facility next to a table of four males from her unit, the men began to talk about women and their hair colors. She claims that one of the males made a comment that blonds were hot, and that another male responded that “Yeah, blonds are hot, but red heads will rock your world and drive you crazy in bed." The male then turned to her (because she had red hair) and said “Isn’t that right?” When she ignored him he then said, “Maybe you’ll give me a test drive and see if I go crazy." The other men started laughing, and out of embarrassment and anger the female left the mess hall and came directly to see you. You speak to the male who allegedly directed the comments to the female, and he claims that he and the other males were only joking around. He expresses regret for upsetting the female, and offers to apologize to her.
Of the following, the most important course of action is:

□ 1 Ask the female if an apology from the male who directed the comments at her would resolve the issue to her satisfaction

□ 2 Train the entire unit about the consequences of creating a hostile work environment

□ 3 Ensure that the commander is aware that the issue is being resolved

Scenario 6
You are following up on a claim that a Black enlisted person has stated that people who are not fluent in English should not be allowed in the military. During a meeting the Black person states that someone not fluent in English puts everyone in the unit at risk.
Your next course of action is:

□ 1 Encourage the Black person to work at improving communication with personnel in the unit who speak English as a second language

□ 2 Ask the Black person to give a specific example of where unit performance was negatively impacted by poor communication among unit personnel

□ 3 After consulting with the chain of command, develop a training strategy to address the attitude of the Black person

Scenario 7
An Asian male complains to you that a White person recently came to him to fix a broken computer. According to the Asian male, when he said that he did not know anything about fixing computers, the White person responded “I thought you Asians knew everything about computers." The Asian male reports being upset by the comment, but that he did not say anything to the White person. He mentions the incident to you in passing during a conversation about the menu for an Asian themed program you are planning.
Of the following, your next question to the Asian male would be:

□ 1 What specifically upset you about the White person's statement?

□ 2 Do you want me to speak to the White person?

□ 3 Do you want to come to my office to talk more about this incident?

Scenario 8
A male (labeled M1) officer has told you that another male (labeled M2) officer used a Military owned video camera to secretly film female personnel, focusing mainly on women’s busts and buttocks. Furthermore, M1 claims that M2 showed the videotape to several male officers. Later, another member of the unit comes to you in private and tells you that the incident did occur, but that M2 taped over the videotape when he found out the EOA was involved.
Your next course of action would be to:

□ 1 Ask both complainants if they will provide written statements about the incident

□ 2 Speak with both complainants to find out more specific details about the making of the videotape, and subsequent occurrences involving the videotape

□ 3 Check with the JAG officer to determine if there are any potential legal issues regarding the alleged behavior of the accused

Scenario 9
You have worked with two members of a unit (one Black, one White) to resolve a conflict that stemmed from the Black person calling the White person a “dumb ass cracker.” In a follow-up meeting, both indicate that everything is fine between them. However, after the White person leaves the meeting, the Black person tells you that there is some tension with other Whites in the unit because of the incident.
Of the following, the next question you ask the Black person would be:

□ 1 Is he sure that he is not overreacting to the situation?

□ 2 How would you treat a White unit member if he / she used a racial slur against a fellow Black unit member?

□ 3 What actions can you take to improve the situation with other Whites in the unit?

Scenario 10
A Muslim officer calls you and asks to file a complaint against his superior Christian officer. The Muslim officer claims that while he was praying, his superior officer shouted orders at him, and that when he continued praying, the superior officer got upset and shouted the order a second time telling him to “stop that garbage and obey my orders.” The Muslim officer said he tried to explain the importance of prayers to the Christian officer, but he claims the Christian officer ignored him, and subsequently punished him for disobeying orders.
Of the following, the most important advice to give the Muslim officer is to:

□ 1 Consider filing a report with the unit in charge of criminal investigations

○ 2 Explain the differences between pursuing the incident as an informal complaint versus a formal complaint

○ 3 Recommend that the Muslim officer pursue the incident through the chain of command

Scenario 11
A female comes to you to seek your advice. For her birthday a number of the men in her unit pooled together and got her a t-shirt that read “Woman by nature. Bitch by choice.” She knows the men meant no harm by the shirt, but she was still insulted. She has not said anything to the men, but she wants them to know the gift upset her. The female wants to handle the problem herself, but she is struggling with how to approach the men. She likes the men--she doesn’t want them to get into trouble, and she doesn’t want to lose their friendship.
Your next course of action would be to:

○ 1 Offer to go the unit commander and suggest gender discrimination training for the unit

○ 2 Brainstorm with the female about possible ways that the female can approach the men in the unit about the incident

○ 3 Volunteer to arrange a meeting with the female and the men involved, with you serving as the meeting facilitator

Scenario 12
You receive a phone call from a Black male who tells you that he has been receiving anonymous letters that threaten his life. He explains to you that he started receiving these letters after he began dating a White female. He claims the most recent letter states that “Martin Luther King Jr. was all about Blacks and Whites working together and he ended up dead. If you think you're Martin Luther King Jr., then you will end up like King.” The Black male suspects that his girlfriend’s former boyfriend, who is also in the military, is sending the letters, but he has no evidence to prove this claim.
Of the following, the best course of action is:

○ 1 For you to immediately report the alleged threats to the unit that handles criminal investigations at your installation

○ 2 Speak with the ex-boyfriend about the situation, figuring that once the ex-boyfriend knows EOA is involved, he will stop sending the letters

○ 3 Advise the Black male to immediately inform chain of command of the alleged threats

Scenario 13
A US enlisted man of East Indian descent comes to you saying that although he is a motor pool driver he is not being allowed to drive. He claims that when he asked his White supervisor why he isn’t being allowed to drive, the White supervisor told him that he has seen how people in India drive, and he will not let an Indian drive a government vehicle. The Indian soldier is angry and he insists that he wants to file a formal complaint.
Of the following, the best course of action would be to:

○ 1 Start filling out the paperwork for a formal complaint

○ 2 Try to convince the complainant to pursue the issue as an informal complaint

○ 3 Explain to the complainant the role of the EOA, and your responsibilities

Scenario 14

Two female officers have reported to you that their male superior officer refers to them as “sweetheart” and “honey”, but only when there are no witnesses. According to the female officers, they approached the male superior officer and asked him to address them in an appropriate manner. Supposedly, he responded: “Yes Dears, I will do better.” The female officers further report that the behavior has not changed. The female officers are angry, and they want you to intervene and force the male officer to stop.

Your next course of action would be to:

□ 1 Ask the female officers to list, as best they can, the dates and times when the incidents occurred

□ 2 Contact the installation commander and inform him / her of the potential for a sexual harassment case against the male officer

□ 3 Instruct the female officers to contact the installation commander regarding the complaint

Scenario 15

An Asian female has asked for your help. She states that a Black enlisted person in the unit called her by the name of another Asian female. The Asian female claims that when she explained to the Black person that she was of Korean descent and the other female was of Japanese descent, the Black person responded “Welcome to my world of y’all look the same to me.”

Your next course of action would be to:

□ 1 Ask the Asian female if she has discussed the incident with the chain of command

□ 2 Advise the Asian female to handle the situation herself by sitting down and talking more with the Black person regarding the incident

□ 3 Use an open-ended questioning technique to find out more about the incident

Scenario 16

A White male (labeled WM1) comes to you concerning another White male (labeled WM2) in his unit. The WM1 who has stopped by tells you that while entering the office of WM2 he noticed what appeared to be a White supremacy web-site on the computer screen. The complaining WM1 is not certain about the nature of the web-site because WM2 quickly closed his web browser when WM1 entered the office.

Your next course of action would be to:

□ 1 Ask the complainant of any other incidents or behaviors by the accused White male that indicate discriminatory behaviors

□ 2 Call the supervisor and suggest that he / she has the computer technician review the internet activity of the accused White male

□ 3 Probe the complainant concerning what he specifically saw that led to the conclusion that the website was that of an extremist group

Scenario 17

A female E3 claims that her male E7 supervisor has threatened to damage her career if she broke off their on-going romantic relationship.

Of the following, the best course of action is:

□ 1 Ask the female E3 what she would like you to do

□ 2 Inform the chain of command of the allegation

□ 3 Contact the male E7 to get his side of the story

Scenario 18

A female reports to you that she was approached in the dinning area by a male who said: "Hey mama you wanna ride this roller coaster?" while pointing to his groin. The female further reports that she ignored the male and that as she began walk away the male held onto her arm and said, "Just one ride." Supposedly, the female lost her temper and said "if you don't let me go I will hit you so hard that your unborn children will feel it." The female states that the incident happened a few weeks ago, and the male has not bothered her any further. However, she is still angered by the incident, and believes that men at the installation do not respect female service members.

Of the following, the most important course of action is to:

□ 1 Inform the commander that there may be a negative climate toward women on the installation

□ 2 Try to find witnesses to the incident

□ 3 Schedule the unit for sexual harassment training

Scenario 19

A Jewish officer gives you a call and he explains that the unit commander, who is Christian, gave each officer in the command a dog tag engraved with a New Testament bible verse that he called the "Shield of faith". The Christian unit commander told the subordinate officers to keep the shield of faith near their hearts because it would protect them. In private, the Jewish officer claims to have told the Christian unit commander of his discomfort with the gift. The Christian unit commander supposedly responded that it didn't matter that he was Jewish because "We both believe in the same God, and a warrior needs the Lord's protection".

Your next course of action would be to:

□ 1 Discuss with the unit commander the problems that a Jewish officer would have with a verse from the New Testament

□ 2 Ask the Jewish officer to express exactly what it is about the unit commander's behavior that offends him

□ 3 Request that the Jewish officer allow a written statement to be taken, and that the Jewish officer give the dog tag to you.

Scenario 20

You have been assisting the processing of a formal complaint by a motor pool driver of East Indian descent. The East Indian driver claimed that his White supervisor did not allow him to drive because of the perception that people from India are poor drivers. The investigating officer has determined that the White supervisor has stated on multiple occasions that people from India can't drive well. However, the White supervisor also provided documentation that in the prior two years the East Indian driver has been involved in three minor car accidents, two of which were his fault. The White supervisor claims that the accidents are the reason he does not let the East Indian drive. The investigating officer has asked your advice on how to proceed.

Of the following, the first recommendation you would make to the IO is:

□ 1 Formally discipline the White supervisor because regulations were not followed to relieve the driver of his duties

□ 2 Interview the East Indian driver again to see if the supervisor ever told him that the accidents were the reason for not being allowed to drive

□ 3 Counsel the White supervisor to be more aware of the impact of stating stereotypical beliefs about people of other racial / ethnic groups.

Appendix B

Questionable Item Questionnaire

1.		
Are there any response options you would never do or that are clearly less appropriate than the other options?	Y/N	
Are the response options all possible choices?	Y/N	
Is there more than one best/viable response option presented?	Y/N	
Could any of the response options be combined into one option?	Y/N	
2.		
Is the scenario unclear (wording, presentation)?	Y/N	
Are the response options unclear (wording, presentation)?	Y/N	
Is the question prompt inappropriate (for the scenario or the response options presented)?	Y/N	
Is the scenario too subjective, dependent on service, or past experiences)?	Y/N	
Is this a particularly difficult/easy question?	Y/N	

3.		
Does this item require knowledge, skills, and abilities critical for EOAs in the field?	Y/N	
What portions of the training program covers this material?		
Are these KSAs learned more so after training or after experience in the field?	Y/N	
Any additional comments?		Scenario? Question prompt? Response options?

Appendix C

SME-based, Respondent-based, and Hybrid Scoring Keys

	Scoring keys		
Item #	SME	Respondent	Hybrid
1	3	3	3
2	1	1	1
3	2	2	2
4	3	3	3
5	-	1	1
6	-	2	2
7	3	-	1 and 3
8	-	3	3
9	3	3	3
10	2	2	2
11	2 and 3	-	2
12	3	1	1 and 3
13	-	-	-
14	-	-	-
15	-	3	3
16	-	3	3
17	-	1	2
18	1 and 3	-	1 and 3
19	1	-	1 and 2
20	2	2	2

Rolanda Findlay

109 Williams Hall (0436)
Blacksburg, VA 24061
Email: rfindlay@vt.edu

Education:

Virginia Polytechnic and State University
Blacksburg, VA
Ph.D. in Industrial & Organizational Psychology expected May 2009

Temple University
Philadelphia, PA
B.A. in Psychology May 2004

Publications/Manuscripts

Findlay, R., Hauenstein, N. (2007). *The Development of a Hybrid Scoring Key for a Situational Judgment Test Designed for Training Evaluation.*

Hauenstein, N., Findlay, R., Kalanick, J., Esson, P. (2007). *Situational Judgment Tests and Training Evaluation.*

Findlay, R., Hauenstein, N. (2006) EOA Training Evaluation: The Development and Implementation of a Situational Judgment Test.

Hauenstein, N., Esson, P., Findlay, R., Kalanick, J. (2005). *Using SJTs to Evaluate Equal Opportunity and Diversity Training Programs.*

Hauenstein, N., Esson, P., Findlay, R., Kalanick, J. (2005). *Assessment of the Effectiveness of Equal Opportunity Advisor Training: A Re-examination of the McIntyre Study.*

Conference and Academic Presentations:

- "Situational Judgment Tests and Training Evaluation" Society of Industrial Organizational Psychologists, NYC, NY, April 2007
- " The Development of a Hybrid Scoring Key for a Situational Judgment Test Designed for Training Evaluation" Graduate Student Assembly, Blacksburg, VA, March 2007
- "Situational Judgment Tests and Training Evaluation" Graduate Student Assembly, Blacksburg, VA, March 2007
- "Evaluation of DEOMI's Equal Opportunity Advisor Training Program: The Transfer of Training Challenge" Defense Equal Opportunity Management Institute, Cocoa Beach, FL, January 2007
- "Using SJTs to Evaluate Equal Opportunity and Diversity Training Programs" Society of Industrial Organizational Psychologists, Dallas, TX, May 2006
- "Now or Later? Inter-temporal Choices and Individual Differences" Temple University McNair Symposium Philadelphia, PA, July 2003

- "Do We Still Need Daddy? How Father Involvement Influences Student Achievement" McNair Conference University of MD, College Park, March 2003; and McNair National Conference Delavan, Wisconsin, November 2002

Educational or Professional Honors or Awards:

- 2nd Place Poster Presentation GSA Research Symposium 2007
- ONR Summer Faculty Researcher 2006, 2007
- Dean's Assistantship by Virginia Tech 2004-present
- Perservantia Vincit Award by the Russell Conwell Center 2003
- W.W. Smith Scholarship Recipient by the Russell Conwell Center 2003
- W.W. Smith Senior Prize Finalist by the Russell Conwell Center 2003
- 1st Place Research Presentation by Temple University McNair Symposium 2002

Research Experience:

Virginia Tech and the Department of Defense-DEOMI Cocoa Beach, FL
Advisor: Dr. Neil Hauenstein
Summer 2005-present
"Using SJTs to Evaluate Equal Opportunity and Diversity Training Programs"
A situational judgment test (SJT) was created to measure the effectiveness of a military diversity training program. SJTs are a low-fidelity simulation that will provide accurate measures of training transfer from the classroom to the field. The SJT offers many benefits, without the challenges presented by conventional testing techniques. This SJT is in the process of being validated. The major question being targeted is whether this test can successfully detect and measure a difference in students prior to and after training. The initial results are promising and show that the SJT is effective in detecting training effects. This research is ongoing and is being expanded to include the evaluation of other performance areas.

Virginia Tech and the Department of Defense-DEOMI Blacksburg, VA
Advisor: Dr. Neil Hauenstein
Fall 2004-Spring 2005
"Assessment of the Effectiveness of Equal Opportunity Advisor Training: A Re-examination of the McIntyre Study"
This research study was a re-analysis of a research study conducted by Dr. McIntyre for DEOMI. The original study was designed to examine the effectiveness of a military diversity program using a video based, pre-post examination. A few limitations were identified in the design and methodology of the initial McIntyre study. Our study attempted to address the limitations of the original McIntyre study by recoding the McIntyre data using a new rubric that 1) was more manageable and 2) objectively based and 3) the coders were blind to the pre-post design condition. The results show that a more effective method for measuring diversity training may be needed.

Virginia Tech, Center for Organizational Research Blacksburg, VA
Advisors: Dr. Neil Hauenstein
"Compensation Study of Comparable Law Enforcement Agencies"
Fall 2004-Spring 2005
Roanoke City Police Officers Association contracted the Center for Organizational Research at Virginia Tech to conduct a survey of compensation of law enforcement agencies. A survey investigating compensation and applicable statistics was first developed. A representative group of Virginia and

surrounding area law enforcement agencies were then chosen and solicited for participation in this project. Findings of the compensation study were provided directly to the Roanoke City Police Officers Association and to the other participating municipalities in this project.

Temple University, Ronald McNair Program **Philadelphia, PA**
Advisors: Dr. Donald Hantula and Carter Smith
"Now or Later? Inter-temporal Choices and Individual Differences"
Summer 2003
This research examined whether people's individual differences had a correlation or relationship with their monetary inter-temporal choices. The specific question targeted was whether certain groups of people discount money more or less over time. The results and conclusions followed suit with the previous research, showing no consistent or significant patterns of correlation. The conclusion is that monetary inter-temporal choices are made on the basis of experience and are a product of the environment, not one's age, gender, or race.

Temple University, Ronald McNair Program **Philadelphia, PA**
Advisor: Dr. Daniel Tompkins
"Do We Still Need Daddy? How Father Involvement Influences Student Achievement"
Summer 2002
The purpose of this research was to show that fatherly involvement in the home learning process will benefit and give an added advantage to the student. Data was obtained through an interview and a convenience survey. The results illustrated a strong correlation between high father involvement in the home-learning process and high academic achievement, and low father involvement and low academic achievement.

Work Experience:

ONR Summer Faculty Research Program, DEOMI **Cocoa Beach, FL**
Summer Faculty Researcher
May 2006-August 2006; May 2007-August 2007
Evaluate the effectiveness of the EOA training program using a situational judgment test (EO-SJT). Specifically, during the 10 week program, revised the current version of the EO-SJT. Created new scoring keys for each version of the EO-SJT using an innovative hybrid scoring technique. Re-evaluated SJT data from current and previous classes of EOA trainees using new scoring key. Results provide initial evidence of a positive effect due to EOA training in the realm of complaint processing. In addition, created 50 additional SJT items for future use in parallel forms of the EO SJT. Presented findings in DEOMI end of summer research forum and in a submitted research paper.

Virginia Tech, VT-PREP **Blacksburg, VA**
Program Evaluation, Graduate Research Assistant
January 2005-present
Evaluate the performance of the Virginia Tech Post Baccalaureate Research and Preparatory Program (VT-PREP). Ensure that the program is offering and completing all of the responsibilities and activities promised and outlined in their grant to the National Institute of Health. Duties include database and online survey maintenance, creation of semester progress reports, and yearly progress reports to the National Institute of Health.

Temple University, Social and Organizational Psychology **Philadelphia, PA**
Decision Making Laboratory Supervisor
June 2003-June 2004
Assist in the management of a computer based decision making experiment. The current study is evaluating the rate at which people make decisions and discount money versus vacation days over varying periods of time. Duties include conducting debriefing interviews on participants, ensuring data quality, recruiting participants, record-keeping, literature searches, and other basic laboratory operations.

Temple University Math and Science Upward Bound Program **Philadelphia, PA**

Residential Tutor/Mentor
June 2002-August 2004
Support, encourage, and counsel students while acting as a positive role model, reflecting the appropriate social, academic, and interpersonal skills required for high school and college success. Supervise students during college tours, educational/cultural trips, and related program activities.
Live in the residential hall during the summer portion of the program. Tutor students and assist them with their research projects. Develop and maintain a positive flow of communication between staff and students.

Temple University Executive Office of the President **Philadelphia, PA**
Office Assistant
October 2002- July 2003
Duties included making copies, faxing, filing, and hand delivering sensitive and confidential materials; greeting visitors, typing, ordering supplies, preparing the conference rooms for Board of Trustee meetings, and opening and sorting office mail.

Teaching Experience:

Virginia Tech, Introduction to Psychology **Blacksburg, VA**
Graduate Teaching Assistant—3 sections of recitation Fall 2004
Designed teaching curriculum. Proctored exams, graded class assignments, and published grades on webpage. Enforced classroom rules, interacted with students, addressed student questions and complaints, provided instructional assistance, and acted as a liaison between the students, and the course professors.

Temple University, Computer and Information Sciences **Philadelphia, PA**
Computer Application Course—3 sections. Teaching and Laboratory Assistant Fall 2002
Provided assistance with the creation, design, and upkeep of student web pages, spreadsheets, databases and PowerPoint presentations during both class time and assigned office hours. Graded all assignments, midterms, and finals. Responsible for publishing class grades on web page, and dealing with any grade discrepancies.

Temple University, Intellectual Heritage/Core Curriculum **Philadelphia, PA**
Teaching Assistant Summer 2002
Assisted in grading papers, exams, and homework. Provided group and individual instruction and assistance. Facilitated group discussions, and presented a mini-lecture series on the Ten Commandments.
http://courses.temple.edu/pericles/Purpose10comm.htm

Listing of Relevant Graduate Coursework:

Statistics for Social Science Research, Research Methods, Measurement Theory, Advanced Statistics, Advanced Topics in Organizational Psychology, Advanced Topics in Industrial Psychology, Psychological Perspectives in Social Psychology, Personality Processes

Extracurricular Activities, Professional and Association Memberships:

Women’s Leadership Conference Publicity Chairperson 2007
Society of Industrial Organizational Psychology, member and TIP Ambassador, 2004-present
Virginia Tech Graduate Resident Advisor, 2006-present
Phi Beta Kappa, member, 2004-present
Temple University Honor’s Program, 2002-2004
Temple University Psychology Honor’s Program, 2002-2004
Psi Chi, Psychology honor society, member, 2002-2004
McNair Post-Baccalaureate Achievement Program, research and teaching fellow, 2002-2004
Temple University Budget Review Committee, member/student advisor, 2003

Speaking Engagements:

- Speaker at the DEOMI Summer Faculty Debriefing, 2006, *"EOA Training Evaluation: The Development and Implementation of a Situational Judgment Test"*.

- Speaker at the Temple University Russell Conwell Dinner 2002, *"Circle of Giving"*

Community Involvement/Service:

- GLC Clothing Drive Coordinator 2006-2007
- Each One Reach One Mentor Program, Mentor 2006
- Multicultural Academic Opportunities Program (MAOP) Graduate Group Leader 2005
- College of Liberal Arts Mentor, 2002
- Residence Hall Success Volunteer, 2002

Technical and Specialized Skills:

I am proficient in SPSS, Microsoft Word, Excel , PowerPoint , Access , Outlook , Front Page and Dreamweaver , and survey creation (www.survey.vt.edu). I am also functional in the Spanish language.