BRINGING SITUATIONAL JUDGEMENT TESTS TO THE 21ST CENTURY: SCORING OF SITUATIONAL JUDGEMENT TESTS USING ITEM RESPONSE THEORY

Tom Haim Ron

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2019

Committee:

Michael J. Zickar, Advisor

Hyeyoung Bang
Graduate Faculty Representative

Joshua B. Grubbs

Samuel T. McAbee

ABSTRACT

Michael J. Zickar, Advisor

Situational judgement tests (SJTs) became popular selection instruments in the last three decades, due to their predictive validity, small subgroup differences, and high face validity. However, although SJTs have made a significant progress in the last century, there still remains a construct problem – it is not sure whether SJTs are a construct or a measurement method. In addition, almost in parallel to the advancement of SJTs, a new theory for scoring and testing has been developed – item response theory (IRT). IRT offers researchers and practitioners flexible models that fit various types of data and can be used to score tests and questionnaires and to learn about their psychometric qualities. In addition, some IRT models offer us a unique method to score multidimensional tests, which assess more than one construct. This study attempts to apply different IRT models to a leadership SJT in order to answer two main questions: one, is SJT a construct or a measurement method? And two, can IRT-based scoring benefit us in terms of validity and reducing subgroup differences over the classical scoring approaches? These questions were tested on three samples of Israeli soldiers who went through a selection process for officers' training school and had to take a leadership SJT as part of it.

The results of this study suggest that the picture is more complicated than it was originally thought. It appears that IRT has value over classical test theory (CTT) only for some samples, whereas CTT has more value in other samples. In regard to the construct vs. measurement method debate, it appears that multidimensional IRT models better fit the SJT that was used in this study, a testimony that sides with the SJT as a measurement method camp. Future research and limitations are discussed at the end of the manuscript.

I would like to dedicate this work to my family, and especially to my mother, father, and grandmother, who helped me, in many ways, to go through graduate school and that without them I would not be able to complete it.

I would also like to dedicate this work to my good friend, Andre, who recently passed away, that without him I would probably fly back to my country in the first two weeks after arriving in the United States for graduate school.

There are probably more people I should thank. I am very thankful for this experience.

ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Mike J. Zickar, who helped me decide on the idea behind this work and who guided me through the process of writing this manuscript.

I would also like to thank Dr. Sam McAbee, a member of my committee, who gave me meaningful comments on previous versions of this manuscript.

Third, I would like to thank Dr. Judy Goldenberg, and the Behavioral Sciences Department in the Israel Defense Forces for contributing the data for this project.

Finally, I would like to thank my colleague and friend, Christopher Gallagher, who was always there for me to consult about this work and helped me go through some of the technical and theoretical issues I had.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

INTRODUCTION

Situational judgement tests (SJTs) became popular as predictors in various selection systems over the last three decades and have been found to be valued predictors of on-the-job success (Weekley & Ployhart, 2006). The typical SJT includes a job-related scenario (or dilemma), based on critical incidents or job analysis, and then a few courses of action as response options (Motowidlo, Dunnette, & Carter, 1990). The respondent is asked to choose or rate the different courses of action according to the way they would behave if they were to encounter this situation. Some SJTs require the applicant to rate all courses of action, some require them to only choose the option that they *would* do, and others ask respondents to choose the option they *should* do, but the basic idea is the same across all SJTs: the respondent is asked to identify the best course of action in regard to a specified situation. Scoring is usually done by comparing applicant's choices to a key, which can be determined rationally or empirically.

***Table 1.* Sample SJT Item**

Your team manager has returned from a meeting where she has heard of an innovative technology development and she is very excited to implement this new technology in your department. She has placed you in charge of researching and developing a plan for implementation. However, as you begin your research, you realize that implementation of your manager's plan will be very expensive, with very little benefit. Your manager is very excited about this new technology and is really pushing to for its implementation.

What would you do?

A. You do not want to hurt your manager's feelings or second guess her judgment, so you come up with a plan, as she requested, and keep your personal feelings quiet.
B. You stop your research and simply explain to your manager that you do not feel the technology change is worth the risk.
C. Despite your feelings, you continue your research and come up with a plan. When you present to your plan to your manager, you express your concerns about the risks.
D. You set up a meeting with your manager's boss to see if he will help you change your manager's mind about the new technology.
E. You talk to a coworker about your concerns and convince him to approach your manager about your reservations for the new technology.

The goal of this study is to evaluate modern and advanced methods for scoring SJTs. Although SJTs have been popular for a while, a significant question remains about how best to score them. Using item response theory, a modern psychometric approach, different models will be evaluated with respect to their contribution to better construct, and predictive validity, as well as reducing between-groups differences. The main research question is whether the use of item response theory can improve these for situational judgement tests.

The development of SJTs varies across tests, but some core steps are common to most SJTs. The first step in developing an SJT is the collection of *critical incidents* (Flanagan, 1954) from incumbents or other subject matter experts (SMEs). These critical incidents could be from a wide variety of work situations or could be focused on job elements derived from job analysis. The next step would be identifying and deciding on constructs to be assessed using the SJT (Lievens & Motowidlo, 2016). Some scholars argue that any construct can be assessed using SJT, other claim that SJTs are their own unique construct – this will be discussed in a later section. The next step would be sorting and editing these critical incidents into the stems of the SJT items. Next, the reviewed and edited situations are assembled into a survey that will be administered to another group of SMEs that will be asked to identify possible responses to these situations. The responses are then sorted and revised to include multiple responses to each situation, and to span on a range of effectiveness (McDaniel & Nguyen, 2001).

The first widely used SJT, like the one described in the previous page, was the George Washington Social Intelligence Test (Moss, 1926). It was designed to measure certain factors of judgement, information, and memory related to dealing with people and carrying on social relationships. World War II saw continued attempts to utilize SJTs to measure judgement, and throughout the 1940s into the early 1960s, additional attempts were made to develop SJTs to assess supervisory and managerial potential (Weekley & Ployhart, 2006).

Research on SJTs has been on the rise since the late 1980s. Publications by Sternberg and colleagues (Sternberg, Wagner, & Okagaki, 1993; Wagner, 1987, Wagner & Sternberg, 1985) on "tacit knowledge" and by Motowidlo, Dunnette, and Carter (1990) on the "low fidelity simulation" stimulated renewed interest in SJTs. The increased popularity of SJTs is probably due to their demonstrated validity (McDaniel, Bruhn Finnegan, Morgeson, & Campion, 2001), reduced adverse impact (e.g., Motowidlo & Tippins, 1993; Pulakos & Schmitt, 1996; Weekley & Jones, 1999), and positive responses from applicants (Weekley & Ployhart, 2006; Chan & Schmitt, 1997). First, research has shown that SJTs have validity approaching that of cognitive ability tests. McDaniel and colleagues (2001), for example, estimated the corrected mean validity of SJTs to be .34. Furthermore, there have been several studies showing that SJTs have incremental validity above and beyond that of cognitive ability tests and personality assessments (e.g., Clevenger et al., 2001; Weekley & Ployhart 2005). These studies indicate that SJTs capture something different than traditional predictors of job performance. Other than that, researchers have consistently demonstrated that SJTs predict performance across a range of different organizational contexts (Christian, Edwards, & Bradley, 2010; McDaniel, Bruhn Finnegan, Morgeson, & Campion, 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007; Murphy & Shiarella, 1997; Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015).

Another advantage of SJTs is that the mean subgroup differences are typically small to moderate. Most importantly, SJTs show smaller racial subgroup differences than those observed for cognitive ability (e.g., Motowidlo & Tippins, 1993; Pulakos & Schmitt, 1996; Weekley & Jones, 1999). For example, Motowidlo and Tippins (1993) calculated *t*-tests for between groups differences between men and women and Black and Whites performance in an SJT. These *t*-tests, converted to effect sizes (Cohen's *d*) were *d* = .11 for men vs. women, and *d* = .36 for Whites vs. Blacks. Pulakos and Schmitt (1996) calculated the mean differences in SJT performance between Whites and Afro-Americans, and between Whites and Hispanics. They found effects sizes of *d* = .34 between Whites and Afro-Americans, and *d* = .05 between Whites and Hispanics. These effect sizes indicate small to moderate differences between groups.

Finally, SJTs have high face validity compared to other predictors. Although research has not fully examined this issue, it seems reasonable to expect SJTs to be readily accepted and explainable to applicants and may even offer the benefit of providing realistic preview of the job (Weekley & Ployhart, 2006). One study compared face validity perceptions of SJTs administered via paper-and-pencil versus video-based across different races. The results of this study showed that face validity for both methods was high, and between the methods, the video-based SJT was preferred (Chan & Schmitt, 1997).

The wide interest in SJTs that sparked research since the 1980s has also revealed several challenges with SJTs, specifically a construct-validity problem. Scholars have disagreed on whether SJTs are a construct or a measurement method. This debate is discussed in the next few pages and was part of what inspired the present study.

SJTs: Method or Construct?

Selection research puts an emphasis on the criterion-related validity of measurement methods. For example, we often speak of the predictive validity of interviews, assessment centers, work samples, SJTs and other measurement instruments (Schmitt & Chan, 2006). However, a question that might be considered of secondary importance or even ignored in practice, is what underlying constructs are being measured using these measurement methods.

Cronbach and Meehl (1955) defined *construct* as “some postulated attribute of people assumed to be reflected in test performance” (Cronbach & Meehl, 1955: p. 283), a definition consistent with our understanding of knowledge, skills, abilities, and other characteristics (KSAOs) in selection research. So, the question for this section would be whether scores on SJTs are indicative of some attribute that resides in individuals (e.g., a general “situational judgement” trait) or is it a method that organizations can use to assess multiple attributes? The distinction between constructs and methods is of high importance and discussed in Arthur and Villado’s (2008) work. They argue that comparisons between constructs and measurement methods (which is common in selection literature) are theoretically to conceptually uninterpretable and thus potentially misleading.

Cronbach and Meehl (1955) defined five criteria in the investigation of construct validity. First are *group differences* – if the definition of the construct implies that some difference should be expected between groups of different examinees, then we can directly assess whether those difference occur. In the case of an SJT, the probability of two groups of examinees that are hypothesized to have different levels of the construct, should score differently on an item or a group of items that represent this construct. A second criterion mentioned by Cronbach and Meehl (1955) was *homogeneity* of the items that were written to assess the same construct. Operationally, if items were written to assess the same construct, then they should be correlated. If there are groups of items in the measure such that the within-group correlations are higher than the between-group item correlations, it suggests additional construct(s). A third criterion for construct validity would be that the construct should reflect a *predictable and interpretable pattern of correlations* with other established measures. The target construct should correlate highly with similar construct(s) and should not correlate with unrelated constructs. Confirmatory and exploratory factor analyses are usually used in order to assess both homogeneity and an appropriate pattern of correlations. A fourth criterion, *stability of scores,* may or may not be indicative of construct validity depending on the theory that defines the construct. Mean change in situational judgement measures might be expected, as individuals gather more experience confronting with situations over time. If, however, situational judgement is a stable individual construct, mean scores may not change. The final criterion by Cronbach and Meehl (1955) would be the consideration of *process issues* in determining the nature of the construct underlying the measure. Such studies require a theory of how someone would come to achieve high or low score on some measure and generate empirical data that test the associated hypotheses. For example, one hypothesis is that SJT performance represents general cognitive ability (Schmidt & Hunter, 1993), whereas an alternative hypothesis is that it represents practical intelligence (Chan & Schmitt, 2002; Sternberg et al., 2000).

In the remainder of this section, I discuss the different points of view of SJTs as a construct or a measurement method, using Cronbach and Meehl's (1955) guidelines to assessing construct validity.

Schmitt and Chan (2006) concluded that SJTs should be considered a measurement method that is useful for very specific constructs. According to Schmitt and Chan (2006), SJTs have dominant constructs that are readily or almost inherently assessed in every SJT (e.g., adaptability, contextual knowledge, practical intelligence). Therefore, every SJT will have some major factor that measures these inherent traits, but the rest of the variance in responses to SJTs should be attributed to targeted constructs aimed by the writer of the test.

McDaniel and Nguyen (2001) also support Schmitt and Chan's (2006) opinion in support of SJTs as a measurement method. According to them, one can build an SJT where a specific construct (e.g., Conscientiousness) is a major determinant of individual differences in item responding. However, like Schmitt and Chan (2006), McDaniel and Nguyen (2001) also mention several correlates of SJTs in general. First, they mention general cognitive ability, which has a mean observed correlation with SJTs of .36 (with credibility interval of .17 to .75). Thus, it is unlikely that an SJT will not be correlated at all with general cognitive ability. Another variable that is weakly related to SJT scores is job experience (mean *r* = .07), as measured by tenure, especially for samples with inexperienced workers (given most of the job is learned during the first years on the job). Within the Big Five personality dimensions, Emotional Stability was found to have the highest correlation with SJTs (mean *r* = .31), and Agreeableness and Conscientiousness were also found to have positive correlations with SJTs (mean *r*s = .25 and .26, respectively). McDaniel and Nguyen (2001) argue that these findings explain the relationship that SJTs have with job performance (as all of the described above variables are also found to be related to job performance).

Other proponents of the SJTs as a measurement method approach are Patterson and her colleagues (2013). These authors argue that because SJTs are a measurement method, there is no single approach to designing them, and each SJT should be evaluated individually regarding issues of coaching, validity, and fairness. According to Patterson et al. (2013), SJTs measure understanding of effective behavior in a given situation. Relating to the finding about personality traits found in McDaniel and Nguyen (2001), Patterson and her colleagues argue that SJTs do not measure personality traits *per se*, but they measure implicit trait policies – beliefs about the costs and benefits of *expressing* certain traits, such as knowing that being agreeable is likely to be better in many situations.

Jackson and his colleagues (2017) challenge the SJT-as-measurement-method approach. They argue that the support of the SJT-as-methods approach is derived from SJTs' correlations with other constructs such as general mental ability or personality traits. However, they ask, what is it about SJTs that might lead to these relationships? In their study, they use Generalizability Theory to decompose multiple sources of variance in SJTs. They differentiate between three sources of reliable between-candidate variance in SJTs: (1) SJT-specific candidate main effects, which are analogous to a general judgement factor for SJTs (SJTs as construct approach); (2) candidate × dimension interaction, which are analogous to dimension-related effects (SJTs as a measurement method approach); and (3) candidate × situation (nested in dimension) interaction, which are analogous to situation-related effects. Using Generalizability Theory, their results support the SJTs-as-construct approach: the SJT-specific candidate main effects accounted for the largest source of variance, explaining between 47.67% and 67.35% of variance, which is 13 times larger than dimension-related effects and at least 19 times larger than situation-related effects. Jackson and colleagues (2017) try to explain what construct is measured, according to their results, in SJTs. They argue that regardless of specific situations, dimensions, or response items, some people consistently score higher than other on judging "appropriate" courses of action when faced with a situational dilemma. To put in other words, SJTs measure "judgmentability" or the ability to judge a variety of situations and react in an appropriate way. This construct has also been known as "tacit knowledge" or "practical intelligence" (Schmidt & Hunter, 1993).

More support in the SJT-as-constructs approach comes from Lievens and Motowidlo's (2016) article. According to them, SJTs measure *general domain knowledge* which is "knowledge about the utility or importance of traits such as these for effectiveness in a job that actually requires expressions of these traits for effective performance" (p. 4). They continue and make several arguments about SJTs: (1) SJTs are related to job performance because they measure procedural knowledge about how to behave effectively in different work situations; (2) general domain knowledge is one component of that procedural knowledge; (3) general domain knowledge is not acquired from specific job experience, rather it is learned through fundamental socialization processes and personal dispositions; (4) this type of knowledge can predict performance in work situations; and (5) SJTs should be developed to measure this type of knowledge deliberately and systematically. Further support for this theory comes from research by Krumm and his colleagues (2015). In their research, they showed that the item stems (the situational component of SJTs) are not necessary in order to answer SJTs correctly. In their first study, they had two conditions: In the first condition, a traditional SJT was used, whereas in the second condition, the situation description (item stem) was removed from each of the items. Results showed that the provision of context in the form of inclusion of situational stems had less impact than typically assumed. It did not matter for 71% of the items whether situation descriptions were included in terms of the number of correct solutions per item. In terms of the total score on this SJT, there was a difference of about 3 points (out of 30) between the two conditions. Lievens and Motowidlo (2016) use Krumm et al.'s (2015) results to emphasize how strongly SJT performance is related to general domain knowledge.

In conclusion, there are disagreements between scholars about the construct validity of SJTs. It is unclear whether we can use SJTs to measure practically any construct (the supporters of this approach point us to the correlations between SJTs designed to measure specific constructs to other measures of the same constructs) or whether SJTs are themselves a construct representing some sort of tacit knowledge, or "general domain knowledge" (Lievens &

Motowidlo, 2016). The scholars who support the SJT-as-measurement-method or the ones that are against it, both use different criteria from Cronbach and Meehl's (1955) guidelines to assessing construct validity. Mainly, each side uses the pattern of correlations between SJTs and other constructs as a support for their arguments. The SJT that will be used in the current research was designed to be multidimensional and measure (at least) three distinct constructs. Through the use of advanced psychometric analyses, I will try to find support for either of the points of view – whether we can actually distinguish between the three constructs that this SJT was designed to measure, or is it unidimensional, measuring general domain knowledge.

How Do People Respond to SJTs?

One of the most common explanations for why SJTs predict job performance is that SJTs require the respondents to make judgements based on their intentions, and that those intentions predict actual behavior. This is especially true when SJT items are phrased with "what *would* you do?" questions – in this case respondents have to consider how they would behave – they have to predict their own behavior (Brooks & Highhouse, 2006). However, the predictions that people make to answer SJTs are not always accurate. Research have shown that people tend to have biased judgements on time and affect-related situations (Brooks & Highhouse, 2006). For example, when asking someone to estimate when they will they start working on a project that is due a month from now, most people will give an optimistic prediction, whereas in fact they do not take into account barriers and distractions on the way to achieving this goal.

Another aspect that should be taken into account when considering how people respond to SJTs is the ambiguity of the context. Even in the case of well-written and detailed SJTs there is some ambiguity in regard to the context. Decision making research has suggested that most people are not bothered by this ambiguity (Ross, 1987; Ross & Nisbett, 1991). Instead of considering all the possible contexts, respondents tend to imagine the most plausible context and make predictions based on that (e.g., Arkes et al., 1988; Shaklee & Fischhoff, 1982). The most worrying aspect of these findings is that there might be difference between what the respondents perceive as "the most plausible context" and what SMEs who decide on the answer key perceive it to be. Respondents may be assessed poorly on an SJT item if they made different assumptions on the ambiguous context, compared to the experts who scored the SJT (Brooks & Highhouse, 2006).

Methods of Scoring SJTs

After writing an SJT, whether it is aimed to measure specific constructs, or it is designated as being a construct itself, measuring general domain knowledge or tacit knowledge, an important issue arises as to how to score participants answers on SJTs. Weekley, Ployhart, and Holtz (2006) differentiated between two groups of methods of scoring SJTs: multiple-choice methods and continuous or Likert-type-scale methods. In the simplest form of multiple-choice methods, one answer is designated as "correct" whereas the others are wrong. This enables the SJT to be scored as an ability test would be scored, with each item answered correct receiving 1 point, and each item answered incorrect receiving no points. In this approach, respondents are asked to "pick the best" out of the response options presented to the situation in the stem. However, Hanson, Borman, Mogilka, and Manning (1999) recognized that not all "wrong" answers are equally poor and decided to score their SJT by assigning each response option its mean effectiveness rating (according to SMEs). This method gives partial credit for choosing answers other than the best, that were still better than the worst answer. Other multiple-choice methods have been developed over the years, such as determining what are the best and worse responses (Motowidlo et al., 1990), or rank-ordering the response options according to their effectiveness (Weekley et al., 2004). In the latter case, the rank-ordering of the responses by the respondent would be compared to SMEs rank-ordering using Spearman's rank-order correlations. In fact, rank-ordering the responses provided better validity than did the "pick best" method or the "pick best/pick worst" methods (Weekley, Ployhart, & Holtz, 2006).

Another common group of methods of scoring SJTs has been to have the respondents use Likert-type scales. For example, Chan and Schmitt (2002) had respondents rate each response on a 6-point effectiveness scale. These ratings were then compared to SME ratings and assigned scores of 1, 2, or 3, depending on the percentage of agreement between the respondent and the SMEs. The use of Likert-type scales in SJTs is less common than the multiple-choice methods, but according to McDaniel and Nguyen (2001), this approach proposes several potential advantages. First, because responses to each item are independent, there is no ipsativity in the resulting scores. Second, the Likert-type methods produce more scores (because respondents often rate each response option of every item), and more data points offer potential benefits in terms of reliability and validity. Third, because such an approach allows the accumulation of response option ratings across situations, it might enable the SJT developer to measure more homogeneous constructs within a single SJT. Finally, the Likert-type approach might reduce the cognitive load of SJTs (compared to the multiple-choice approach; Weekley, Ployhart, & Holtz, 2006). Arthur, Glaze, Jarrett, White, Schurig, and Taylor (2014) also compared Likert-type SJT scales ("rate" scales) with other scales ("rank" scales and "worst/best" scales) and found it to be superior in terms of reliability, reducing subgroup racial differences, and showing lower correlation with general mental ability.

Up to this point I have described scoring methods for SJTs that have the underlying assumption that there is a "correct" answer to an SJT item (or that each response option has an absolute degree of effectiveness, if we are dealing with Likert-type scoring). The correctness of the answers, under this assumption, is determined by SMEs. However, unlike cognitive ability tests, SJT items usually do not have objectively correct answers and many of the response options are plausible (Bergman et al., 2006). Therefore, scholars have studied and identified at least five groups of methods of determining the scoring keys for SJTs. I will discuss these methods next.

The first group of methods is *empirical scoring*. In this method, item options are scored according to their relationships with a criterion measure (Hogan, 1994). Methods for empirical scoring of SJTs usually include the following process: choosing a criterion, developing decision rules, weighting items, and cross-validating results (Bergman et al., 2006). Empirical keys usually have high validity coefficients (Hogan, 1994; Mumford & Owens, 1987), but they are also dependent on criterion quality (Campbell, 1990), they have stability and generalizability issues (Mumford & Owens, 1987) and they often capitalize on chance (Cureton, 1950).

The second group of methods is *theoretical scoring*. Similar to biodata's rational method (Hough & Paullin, 1994), theory can be used to identify the best or worst options in an SJT. Options reflecting the theory are scored as +1, whereas options contradicting the theory are scored as –1. Neutral options are scored as zero. Theoretical approaches address one major criticism of empirical methods such as being atheoretical, and theoretical keys are more likely to generalize (Bergman et al., 2006). However, theoretical keys are more transparent and therefore are more susceptible to faking. Furthermore, the theory used to score the SJT might be flawed or incorrect (Hough & Paullin, 1994).

The third group of methods is *hybridized scoring*. In this group of methods, different keys that were created using different approaches are combined together. Two keys could be added at the option level, for example, allowing a positive score on one key to cancel out a negative score on the other. Another hybridization approach is substitution for zeroes. In this approach, the main key is used to score the correct answer (+1), the incorrect answer (–1) and the neutral answers (0), and then a second key is used just to score the neutral answers. Keys can also be differentially weighted, such that one key is used with the full scores and the other is fractionally weighted (Bergman et al., 2006). Hybridizing an empirical key with a theoretical key resolves some of the challenges of each of these keys because it both recognizes theory and relies less on pure empiricism.

The fourth group of methods is the most common one – *expert-based scoring* or *rational scoring*. This group of methods utilizes SMEs consensus to determine the answer key for an SJT. The most common way to use SMEs is to ask a group of them for their opinion on the different response options, which are then scored (according to agreement reached between SMEs) as correct (+1), incorrect (–1), or neutral (0). There are other methods of calculating scores using SMEs, among them are raw consensus, standardized consensus, dichotomous consensus, mode consensus, and proportion consensus (for reviews see: McDaniel et al., 2011; Weng et al., 2018). Another expert scoring approach contrasts the opinions of novices versus experts. Experts and novices complete the assessment, choosing the best option for each SJT item. If a response option was chosen as best by the group of experts (regardless of what the novices chose), then it is scored as the correct answer. If an option was scored as best by the group of novices, but not by the group of experts, then it is scored as incorrect (Bergman et al., 2006).

The fifth group of methods is *factorial scoring*. Factorial approaches are used when there are no *a priori* construct-based scales specified and items are not assumed to measure particular constructs. In this group of methods, items are scored based on factor analysis and item correlations. This approach is useful when theory does not define the relevant constructs, but the item pool could still produce meaningful dimensions. This approach can also be used to remove items from item pool if they do not load on an identifiable factor in the factor analysis (Bergman et al., 2006).

When designing an SJT, it is important to know which method for creating the scoring key is the most useful. Bergman and her colleagues (2006) have defined four standards that should be used in evaluating the different groups of methods: First, high criterion-related validity is required from an SJT. Second, high incremental validity is required (above and beyond personality and cognitive ability). Third, the method should minimize subgroup differences in order to reduce adverse impact. Finally, the key should produce construct-valid measures – the scores should correlate with other measures that the SJT should correlate with (i.e., convergent validity) and should not correlate with measures that the SJT should not correlate with (i.e., discriminant validity).

Using these standards, Bergman et al. (2006) assessed a leadership SJT that was scored in 11 different ways, using the different scoring approaches described above. They found that the empirical, SME and one of the hybrid keys all predicted the leadership criterion. These keys also provided significant incremental validity over cognitive ability and personality measures. None of the keys showed subgroup differences by sex, and all of the keys showed discriminant validity. They conclude their recommendations by recommending users of their SJT to use either the empirical, SME, or hybrid approaches for scoring the SJT.

In the next sections I will describe item response theory in detail. This study pertains to use item response theory as an innovative method of scoring SJTs, in addition to scoring them using the traditional methods that were described previously. Using item response theory is hypothesized to improve the validity and usability of SJTs.

Item Response Theory

Item Response Theory (IRT) is a system of models that defines a way of establishing the correspondence between latent variables and their manifestations (de Ayala, 2009). IRT is known to be a modern psychometrics theory, because it made significant advances over Classical Test Theory (CTT) in several substantive research areas (Morizot, Ainsworth, & Reise, 2006). IRT has also been referred to as "the most important statistical method about which researchers know nothing about" (Kenny, 2009). This lack of awareness is reflected in the lack of research of how to use IRT models to score SJTs. The brief history of research in this area will be covered in a later section. However, IRT has been used in Industrial-Organizational (I-O) psychology research for issues such as personality assessment, performance rating, intelligence assessment, vocational interests, and employee opinions (Foster, Min, & Zickar, 2017).

IRT models use person and item characteristics to estimate the relation between a person's underlying trait level (called *theta* – θ; e.g., Conscientiousness, GMA) and the probability of endorsing an item (LaHuis, Clark, & O'Brien, 2011). Different models of IRT have different levels of complexity and they estimate different parameters. The most basic model of IRT, the one-parameter model (1PL), assess an item's difficulty, or location, as represented by *b*. *Location* is defined as how much of latent trait θ would be needed to have a 0.50 probability of endorsing a correct item. Item location parameters typically range in value from –2.5 (item endorsed by nearly everyone) to +2.5 (item endorsed only by those very high in the trait). A more advanced model, the two-parameter model (2PL) includes a second parameter, which is the *a* parameter, which represents an item's ability to differentiate between individuals on the latent trait. IRT models may include a third parameter, the pseudo-guessing parameter which is related to the threshold for guessing on an item (Zu & Kyllonen, 2018).

In the simplest model, the 1PL model, a person's item responses are modeled from a single difference between their ability and the item's location in the context of a logistic model. The 1PL model is given as follows:

$$P\left(X_{ij} = 1 \middle| \theta_j\right) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}}$$

where $P\left(X_{ij} = 1 \middle| \theta_j\right)$ is the probability that person j with ability of θ_j answers item *i* correctly. b_j is the difficulty for item *i*. This estimation enables us to generate Item Response Function (IRF; also known as item characteristic curve – ICC) which graphically illustrates the relationship between the latent trait and a specific item. The IRFs can visually represent how items differ in terms of *a* and *b*. As seen in Figure 1, item location (*b*) is determined by the intersection of the IRF and a vertical line from probability of 0.50, and item discrimination (*a*) is determined from the slope of the IRF at the item's location. Items with high discrimination are represented by a steep IRF, and are better at differentiating individuals in the latent trait range around the point of inflection (Morizot et al., 2006).

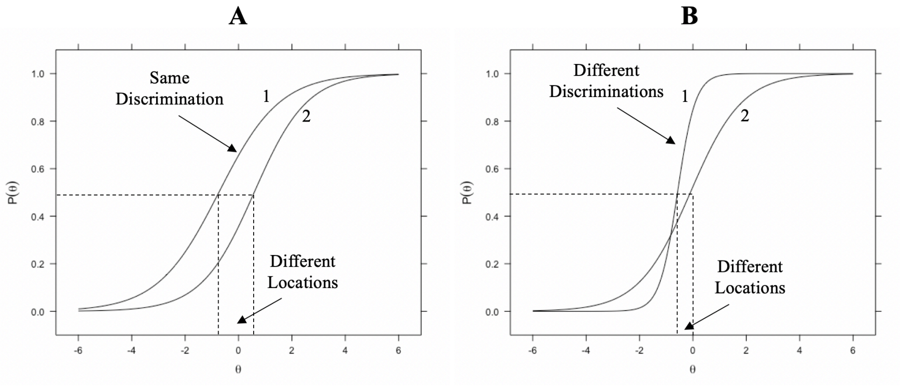

Figure 1. Item Response Functions

In traditional CTT, the precision of test scores is often indexed by a simple specific reliability coefficient. However, in IRT, the concept of reliability is replaced by that of item and scale information. Once the parameters (i.e., item location, item discrimination) have been estimated, each IRF can then be transformed to an *item information function* (IIF). This function indexes the degree to which an item is able to differentiate between individuals at different trait levels (Morizot et al., 2009).

In order to apply IRT models to a test, two basic assumptions must be satisfied: (1) the IRFs have a specified form, and (2) local independence has been obtained. The *form* of an IRF describes how changes in the latent trait relate to changes in the probability of endorsing a specified response category. The form specified in the 1PL, 2PL, and 3PL models (which will be used in this project) is logistic, which gives the S-shaped curves shown on Figure 1 (Schmidt & Embretson, 2003). The main premise of this assumption is that as the latent trait increases, the probability that a respondent will endorse the correct item also increases. There are other models that have other shapes, such as the generalized graded unfolding model (GGUM), though in this project I will only use logistic models.

Local independence is obtained when the relationships among items (or persons) are adequately reproduced by the IRT model. That is, the principle of local independence states that no further relationships remain between items when the model parameters are controlled. Achieving local independence also implies that the number of different person variables (traits) in the model is sufficient to reproduce the data. Thus, if a model with only one person parameter is sufficient, then the data must be unidimensional if the local independence assumption is to be met (Schmidt & Embretson, 2003). The most common way to assure local independence is to calculate the correlations between the test items after controlling for θ. If correlations exist, then the dataset is locally dependent and does not meet the assumptions of the IRT model. Violations of local independence can result in biased parameters, model misfit, and an overestimation of model validity (Reckase, 2009).

IRT models can be used to score dichotomous or polytomous data (Ostini & Nering, 2006). The data type that one uses dictates which IRT model one should choose to analyze their test. Dichotomous data can be coded into two response options (e.g., correct and incorrect), and polytomous data can be coded into multiple categories. Polytomous data can be further divided between ordered and unordered response formats (Ostini & Nering, 2006). Ordered response format includes Likert-type scales, while unordered response formats include scales that cannot be *a priori* ordered in terms of their correctness.

IRT has five noticeable advantages compared to CTT. First, IRT scales are parameter invariant (de Ayala, 2009). That is, that the different item and person parameters estimated using IRT models are invariant in regard to different samples or different measurement conditions. Parameter invariance is of high importance if one wants to assess the degree of inferential generalizability across different samples or populations or measurement conditions for a given modeling context and thus constitutes a fundamental property of measurement for latent variable models (Rupp & Zumbo, 2006).

Second, IRT does not assume that different items are equally difficult or equally discriminating, and it incorporates other parameters of scale items into the respondents' score (Reckase, 2009). This is in contrast to CTT, in which the sum of the raw item scores is the total score (Warne, McKyer, & Smith, 2012). Because IRT examines the pattern of responses within the item to assess multiple parameters, it can provide a better estimate of individual's latent trait.

The third and fourth advantages relate to the characteristics of the response options. The third advantage of IRT relative to CTT is that some IRT models cater well to polytomous response options without an objectively correct response (Kenny, 2009), whereas with CTT items must have correct or incorrect answers. Fourth, IRT response formats may be better suited to ambiguous, difficult-to-interpret responses (Zu & Kyllonen, 2018). Both of these advantages will potentially be important in helping improve the scoring of SJTs.

The final advantage of IRT is that the scale information functions (SIF) allow researchers and practitioners to better understand whom they are differentiating between. Specifically, they can determine whether their selection instrument differentiates equally across varying levels of a significant latent construct, or only between specific levels of a latent construct. Consequently, depending on one's goals, items with particular characteristics can be added or removed to find new employees with specific level of some latent trait.

Some of these advantages are especially valuable when scoring SJTs. IRT can examine unordered polytomous data with ambiguously correct items, which is a characteristic of SJT items (Bergman et al., 2006). Second, some IRT models are particularly well suited for data in which identifying the best response is difficult. Finally, pattern scoring (which is used in IRT) are particularly important on SJTs because although individuals mean scores may be similar, the responses they select may be very different.

In the following section, different models of IRT relevant to scoring SJTs will be described. Each model has its own advantages and disadvantages. One of the most important models relevant to IRT scoring is the nominal response model, which enables scoring of nominal items (such as SJT items) without a specific correct answer. This model and others will be described in detail in the next section.

Scoring SJTs Using IRT Models

Scholars have come up with a large number of IRT models that aim to handle different types of response formats with different theoretical underpinnings (Ostini & Nering, 2006). Most IRT models can use the same types of keys used by most traditional SJT scoring methods as their raw input (e.g., consensus SME scoring). Seven models are relevant to SJT scoring and multidimensional IRT (MIRT): the nominal response model (NRM), the generalized partial credit model (GPCM), the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), the three-parameter logistic model (3PL), and two MIRT models, the M2PL and M3PL.

The Nominal Response Model

Bock (1972) designed the NRM to score unordered polytomous response formats (Ostini & Nering, 2006). The model assumes a continuous latent variable accounts for all the covariance among unordered items. The most valuable characteristic of the NRM is that it does not require a scoring key. Although one response is clearly correct relative to the other multiple-choice options, the model estimates the correct response based on the items' relations with θ. As such, the purpose of the NRM is to find implicit ordering in unordered categorical data, such as data from SJTs (Samejima, 1972). The NRM model is expressed as:

$$P(X_i = k|\theta) = \frac{e^{a_{ik}\theta + c_{ik}}}{\sum_{j=1}^{m_i} e^{a_{ij} + c_{ij}}}$$

where X_i is the response on the ith item and $P(X_i = k|\theta)$ is the conditional probability of choosing response category k (k =1, …, m_i) for item i; a_{ik} and c_{ik} are, respectively, the category slope and category intercept parameters for the kth category of item i. In this model, the probability that a person with trait-level θ selects option k on item i is given by the expression on the right – the ratio of selecting one category over the sum of all the other categories, as a function of ability (θ) with a varying category slope parameter (a_{ik}; or item discrimination), and a varying category intercepts parameter (c_{ik}; or endorsement likelihood) for each of the m_i response categories. By fitting the NRM to the response data, item parameters (a_{ik} and c_{ik}) can be estimated. The estimated category slopes within an item provide the empirical response option orders (Zu & Kyllonen, 2018).

The original NRM has also been expended recently to cases of multidimensional nominal response data, such as multidimensional SJTs (Revuelta, 2014). In the multidimensional case of the NRM, several item discrimination parameters are estimated – same as the number of dimensions the data is specified to have. Each item loads to a different degree on the different dimensions (factors) and the item discriminations resemble the different loadings of the items on the different factors (Chalmers, 2015). In this study, both the unidimensional NRM and the multidimensional NRM will be used, in order to assess the usefulness of using a multidimensional model above using a unidimensional one.

The Generalized Partial Credit Model

The second IRT model relevant to scoring SJTs is the GPCM (Muraki, 1992). The GPCM is similar to the NRM, with the added constraint of the slope parameter measured by a, which represents the item's ability to discriminate between respondents (Thissen & Steinberg, 1986). The GPCM requires that a given set of response options to a specific item stem have an explicit order in terms of their appropriateness. Thus, in contrast to the NRM, the GPCM requires a detailed key that orders each item within a stem from best response to worst response (Zu & Kyllonen, 2018). The GPCM is expressed as:

$$P\left(X_{jk}\middle|\theta,\alpha_j,\delta_{jk}\right)=\frac{e^{\sum_{h=1}^{k_j}\alpha_j(\theta-\delta_{jh})}}{\sum_{c=1}^{m_i}e^{\alpha_j(\theta-\delta_{jh})}}$$

where X_{jk} is the response for item *j*'s *k*th category, θ is the latent trait, α_j is the item discrimination, δ_{jk} is the transition location parameter between the *h*th category and the *h -1* category, m_j is the number of categories and $k = \{1, \dots, m_j\}$ (de Ayala, 2009).

Since the SJT data that will be used in this study does not have a pre-specified order between the response options, the GPCM will not be relevant and will not be used.

The One-, Two-, and Three-Parameter Logistic Models

The one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL), which have been used to score SJTs, are based on the logistic function. These IRT models can accommodate dichotomous response formats or data that can be coded into dichotomous response format. The 1PL is the least complex model, and it estimates only the item's location, *b*. The 2PL model estimates both item location (*b*) and item discrimination (*a*). Finally, the 3PL model estimates *b* and *a*, and also a parameter that measures the respondent guessing the best response option (*g*). The 3PL model is expressed as:

$$P\left(X_j = 1 \middle| \theta, \alpha_j, \delta_j, \chi_j\right) = \chi_j + (1 - \chi_j)\frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}}$$

where X_j is the response for the *j*th item, θ is the latent trait, α_j is the item discrimination, δ_j is the item location, and χ_j is the item's pseudo-guessing parameter (de Ayala, 2009).

Multidimensional IRT Models

Multidimensional item response theory (MIRT) models are referred to as either compensatory or non-compensatory. MIRT models that assume between-item multidimensionality are non-compensatory (e.g., an instrument aimed at evaluating healthy eating self-efficacy that includes affective and cognitive dimensions – lower location on the affective dimension will not be compensated by a higher location on the cognitive dimension). In contrast, within-item multidimensionality models are compensatory because a high score on one dimension can compensate for a low score in another dimension (e.g., an instrument aimed at evaluating capability of solving mathematical word problems might have a verbal dimension and a math dimension – higher location on the verbal dimension might compensate on lower location on the math dimension; Reckase, 2009; de Ayala, 2009). Multidimensional models are especially appropriate for SJTs where individual items can tap multiple latent constructs. Most attention in the literature have been given to compensatory models of MIRT (de Ayala, 2009), and this group of models might fit better for the SJT case, such as the one that will be analyzed in this study, in which location on one capability can compensate on a lower location on another capability.

In general, MIRT captures the same item information as IRT, though it does it in a multidimensional space. As such, θ in MIRT is a vector that measures multiple elements (de Ayala, 2009). In addition, MIRT captures *d*, which measures item difficulty (or location) in multidimensional space. Unlike *b*, *d* could include multiple locations for the same level of difficulty. Compensatory models estimate the *a* parameter for each latent trait that the item is assumed to measure. For example, an item measuring two dimensions will have associated a_1 and a_2.

Non-compensatory MIRT models are assumed to have items that measure only a single underlying θ, which is very unlikely with SJTs (McDaniel & Whetzel, 2005). On the other hand, the M2PL and M3PL are compensatory models that may be appropriate for scoring SJTs. They vary in the parameters that are estimated, but are both used to score dichotomous responses. The M2PL model estimates *d* and *a*. The M3PL also estimates a guessing parameter, *g*, to account for the observation that respondent may correctly answer a question that should require higher levels of θ (Reckase, 2009; Lord, 1980). The parameter *g* is estimated for each dimension on each item, similar to the *a* parameter. The M3PL model is expressed as:

$$P\left(X_{ij} = 1\middle|\theta_i, \alpha_j, \gamma_j, \chi_j\right) = \chi_j + (1 - \chi_j)\frac{e^{\alpha_j\theta_i+\gamma_j}}{1 + e^{\alpha_j\theta_i+\gamma_j}}$$

where X_{ij} is the response for item *j*, θ_i is the latent trait *j*, α_j is the discrimination parameter, γ_j is the location parameter, and χ_j is the pseudo-guessing parameter.

Previous Research of Scoring SJTs Using IRT

Research into scoring SJTs using IRT is quite limited. Zu and Kyllonen (2018) compared several CTT approaches and IRT models to score the Situational Test of Emotional Management for Youths (STEM-Y; study 1) and an SJT aimed to measure teamwork and collaboration (study 2). They used eight scoring methods: number correct, weighted sum, proportion-consensus, 1PL, 2PL, partial credit model (PCM), GPCM, and NRM. In study 1, the NRM score was shown to be superior to other scores in that it had higher reliability and its correlations with external variables were high. In study 2, however, all scores performed equally. This led the authors to inconclusive results. They conducted further analyses to understand why they had different results for the different SJTs in the two samples, and they concluded that in cases where item ambiguity is high, NRM may be the more appropriate method.

Another study that employed IRT models to score SJTs is Wright (2013). Wright examined the dimensionality of SJTs using MIRT and factor analysis. Wright first used exploratory factor analysis and confirmatory factor analysis (CFA) to form four factors of the SJT, and then used MIRT to come up with three more factors, for a total of seven factors. She attempted to predict job performance, controlling for personality and a measure of GMA. Her results showed that the overall SJT score increased the *R*-squared value from 0.14 to 0.23, the addition of the four CFA factors increased the *R*-squared value to 0.27, and the addition of the three MIRT factors increased the *R*-squared value to 0.41. These results showed that each of the SJT scores added incremental validity to the prediction of job performance.

Both Zu and Kyllonen (2018) and Wright (2013) found higher levels of validity using their respective IRT and MIRT models than with other methods for scoring SJTs. As such, IRT and MIRT show promise for scoring SJTs with higher levels of validity and provide practitioners with an efficient method to increase the validity of their selection instruments.

The Present Study

The present study focuses on determining the usefulness of using IRT and MIRT in scoring an SJT. I will compare the traditional CTT method of scoring an SJT, specifically, choose the best option and the second-best option, to other forms of scoring SJT, mainly the NRM (both unidimensional and multidimensional), MIRT and other models of IRT that have been mentioned in the previous section. The main research question that this study intends to deal with is whether the use of IRT can improve SJTs construct validity, predictive validity, and minimize between-groups differences.

Previous studies of employing IRT models to scoring SJT have provided inconsistent results. The main purpose of this study is to provide more conclusive results about the usefulness of using IRT to score SJTs. In addition, the construct of the SJT used for this study will be assessed, mainly, whether it should be treated as a unidimensional test or a multidimensional test. I will compare methods of scoring that assume different levels of multidimensionality and compare model fit indices together with different forms of predictive validity to make judgement on the usefulness of the different methods.

To determine the usefulness of different scoring methods, I will use the standards suggested by Bergman et al. (2006) which have been discussed before: proven criterion-related validity, incremental validity above and beyond personality measures and cognitive ability, minimization of subgroup differences, and the scoring method should produce construct-valid measures. More about the means to assess these standards will be detailed in the Method section.

Because previous research has not reached conclusive results about the usefulness of different IRT and MIRT methods in scoring SJTs, I will not form directional hypotheses, but will treat this research as an exploratory attempt to identify the most efficient way of scoring SJTs.

The first research question that will be analyzed regards a possible improvement in the criterion-related validity of the SJT using the IRT/MIRT models, compared to the traditional scoring method.

Research Question 1: Does using IRT methods increase the criterion-related validity compared to other methods? If so, are certain IRT models better than others?

The second research question that will be analyzed regards incremental validity of the SJT above and beyond measures of personality and general mental ability.

Research Question 2: Does the IRT-scored SJT provide incremental validity in predicting performance above and beyond personality measures and general mental ability?

The third research question that will be analyzed is about subgroup differences between men and women in performance in the SJT.

Research Question 3: Does the IRT-scored SJT show reduced subgroup differences between men and women?

This study intends to contribute to the literature about SJTs and IRT in several ways. First, as mentioned previously, there are only few studies who attempted to apply IRT to SJTs. This study intends to be another attempt to apply IRT to SJT in order to find out whether using IRT models can contribute to different aspects of SJT usefulness. Second, a practical contribution of this study would be in suggesting practitioners on new ways to score SJTs in order to increase different types of validity, mainly predictive validity of performance on-the-job. Third, a contribution to the IRT literature will be in applying unique IRT models such as the NRM to different scientific fields, such as I-O psychology. The NRM was used until now mainly in educational settings. In addition, this study will apply the multidimensional NRM model to the SJT data, something that have never been done before.

METHOD

For the purpose of this study, an archival dataset from the Israeli army (Israel Defense Forces; IDF) will be used. This archival dataset includes information from a selection process to officers' training school and thus includes variety of selection methods: personality tests (Big Five, thematic apperception test [TAT], sentence completion test), an interview with a psychologist, data analysis test, integrity test, biographical questionnaire, and leadership situational judgement test. The dataset includes only people who were selected to attend officers' training school.

Participants

The dataset includes information about 5,536 soldiers, of which 59.8% were male. There was no information about the age of the participants, but usually soldiers participate in this selection process between the ages of 18-21; some exceptions are soldiers who studied for their bachelor's degrees before their military service; in these cases, the ages expected to go through this selection process are 21-23. 327 participants did not have complete SJT data (skipped items). In order to avoid completing missing data for the IRT analyses, these participants were not included in the analyses. Therefore, the final sample size was 5,209 participants.

As for minorities participating in the selection process, one of the minority groups that the IDF tracks their performance in this selection process are Ethiopian descendants Israelis. Only 1% (*N* = 58) of the sample were Ethiopian descendants Israelis.

In addition, participants are divided according to the type of officers' training that they are designated for. There are three types of trainings: training for non-combat soldiers (e.g., clerks, intelligence, human resources, IT), training for semi-combat positions (e.g., boot camp commanders, logistics), and training for combat positions (e.g., antiaircraft, infantry, artillery). In total, 51.4% of the sample were designated for non-combat training, 22.1% were designated for semi-combat training, and 26.5% were designated for combat training.

Measures

Leadership Situational Judgement Test: This test, also named the Media test, is a video computer-based SJT that includes 25 items. Each item presents a situation that an officer is likely to encounter on his job and presents five response options as to how the candidate would react in that situation. A transcript of one of the video-based items is presented below (translated from Hebrew):

> You are a company commander in boot camp. You have just been notified that one of your platoon leaders is not behaving in a matter that is appropriate for an officer in the IDF. The platoon leader is giving instructions to his platoon and then leaves to sit in the cafeteria, knowing that his class commander will lead the job. How will you react?
>
> ___ a. I will scold him in the staff meeting.
> ___ b. I will invite him to a personal meeting and clarify to him that he does not behave in a way expected from an officer in the IDF.
> ___ c. I will imply to him in the staff meeting that I know of his behaviors, using guiding questions such as "how do you pass training? Who is your most dominant soldier?" etc.
> ___ d. I will invite him to a personal meeting in which I will clarify that from now on he is under my watch, and if he will not improve his behavior, I will suggest dismissing him.
> ___ e. I will conduct an unannounced inspection and catch him "on the act".

The candidates are asked to rank the five response options from the best course of action (1) to the worse course of action (5). In practice, only the best course of action and the second-best course of action are recorded and saved. The current scoring method used for this SJT is rational scoring (Bergman et al., 2006). The respondent answers are compared to those scored by a group of experts (high ranked officers in the IDF). This group of experts chose the best course of action for each of the items. If the respondent has chosen the same course of action as the group of SMEs as their best course of action, they receive 2 points. If the respondent has chosen the best course of action according to SMEs as their second-best course of action, they receive 1 point. If the respondent did not choose the SMEs best course of action as either their best or second-best courses of action, they receive 0 points. This scoring key produces a score on the scale of 0-50 for the entire test. For the IRT and MIRT analyses, only the response option that was chosen as the best course of action will be used, because these models do not handle to valid responses to one item.

This SJT was built as a multidimensional test, assessing at least 3 dimensions of leadership problems: Professional performance, dealing with problematic soldiers, and dealing with unfavorable opinions of the leader. This study will apply both unidimensional models and multidimensional models to the SJT in order to assess whether measuring multiple constructs are actually possible with one SJT.

Personality: The Officer Personality Inventory (OPI) was used to measure the Big Five personality traits. The OPI is administered as part of the selection process and is modeled after the Big Five scales of emotional stability, conscientiousness, agreeableness, extraversion, and openness to experience. It has a total of 240 items, with 48 items per scale. The inventory includes behavioral statements that are self-rated on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree). Each of the OPI's scales is scored on a stanine scale, with a mean of 5 and a standard deviation of 2 (Fine, Goldenberg, & Noam, 2016).

General Mental Ability: General mental ability was measured using the IDF's Initial Psychotechnical Rating (IPR). The IPR is a test administered to all newly inducted soldiers as part of their initial assessment and placement process. It is a composite of four subtests measuring verbal and non-verbal abilities. Verbal ability is measured using a modified Otis-type verbal test (test measuring understanding of simple to complicated instructions) and a multiple-choice verbal analogies test, whereas non-verbal ability is measured using an arithmetic test and spatial analogies test. The IPR is scored on a stanine scale from 10 to 90, with an approximate mean of 50, and standard deviation of 20 (Fine et al., 2016).

Officer Training Course Grade: As the criterion for the criterion-related validity estimates, the final grade from the officer training course will be used. This grade is based on various tests that are administered during officers' training – some are practical tests (like navigation), and some are paper-and-pencil tests. The purpose of this grade is to estimate the performance of the soldier as an officer in the Israeli army. These grades also have a holistic judgement component, based on officers' training course commanders who spend most of their time with the soldiers evaluated and have high familiarity with them. These grades are on a scale from 0 to 99, with a mean of 79.47, and a standard deviation of 6.06 for this sample.

Analyses

The Media SJT was scored using the traditional summed score approach, in addition to the IRT and MIRT scoring, including the following models: Unidimensional NRM (no scoring key specified), multidimensional NRM (no scoring key specified), 1PL, 2PL, 3PL, M2PL, and M3PL. For the MIRT models, three dimensions were assumed (as the test was planned to estimate three constructs). For the 1PL, 2PL, and 3PL models, the SMEs-based scoring key was used, and answers were coded as 1 if the participant chose the best answer according to SMEs, or as 0 if they did not.

Prior to all main analyses, a differential item functioning (DIF) analysis was conducted, in order to assess whether the type of training (combat vs. semi-combat vs. non-combat) influenced the results' patterns and parameters estimated in IRT. The DIF analysis was conducted according to the steps presented in Tay, Meade, and Cao's (2015) paper – a baseline model was estimated and the IRT parameters estimated using the NRM on the actual data were compared to this baseline model in order to find differences. The results confirmed that there was in fact differential item functioning based on training type (see the Results section), and therefore the rest of the analyses were conducted separately for the different training groups.

The first research question that was analyzed was regarding a possible improvement in the criterion-related validity of the SJT using the IRT/MIRT models, compared to the traditional scoring method. Criterion-related validity was assessed using the Pearson correlation coefficient between the scores (the summed score in the traditional approach, and the thetas in the IRT and MIRT approaches) on the SJT and the criterion (the final grade in officers' training course).

The second research question that was analyzed was regarding incremental validity of the SJT above and beyond measures of personality and general mental ability. Incremental validity of the different SJT scores was determined using hierarchical regression, with personality and GMA entered in the first step, and the SJT scores added in the final step. Significant changes in R^2 indicated a positive incremental validity.

The third research question that was analyzed was about subgroup differences between men and women in performance in the SJT. Subgroup differences were assessed between men and women soldiers in the Media SJT. Subgroup differences in the different SJT items were estimated using effect sizes (Cohen's d) for the gender variable.

The construct structure of the SJT was assessed using model fit indices for fitting the different IRT and MIRT models (unidimensional or multidimensional) to the data. Good model fit indicated a coherent structure of the SJT, while poor model fit indicated ambiguous structure of the SJT.

RESULTS

Differential Item Functioning (DIF) Between Types of Training

In order to establish whether there were different response patterns between the different types of trainings that the candidates in the dataset went through (training for combat soldiers / semi-combat soldiers / non-combat soldiers), a differential item functional (DIF) analysis has been conducted. For the purpose of this analysis, the nominal response model (NRM) has been used (because it fits the data best – see next section, *Table 3*), assuming three-dimensional space (because the SJT was designed to measure three constructs). The results of this analysis are presented in *Table 2*.

Table 2. **DIF Analysis Between Training Types**

Item	*Wald* χ^2	*df*	*p*
1	47.57	20	<.01
2	42.41	20	<.01
3	223.34	20	<.01
4	60.70	20	<.01
5	18.46	20	.56
6	49.07	20	<.01
7	68.10	20	<.01
8	83.00	20	<.01
9	46.09	20	<.01
10	43.60	20	<.01
11	31.98	20	.04
12	79.34	20	<.01
13	40.91	20	<.01
14	83.62	20	<.01
15	36.85	20	.01
16	43.39	20	.01
17	66.45	20	<.01
18	69.69	20	<.01
19	113.02	20	<.01
20	26.50	20	.15
21	31.98	20	.04
22	61.17	20	<.01
23	32.87	20	.03
24	42.15	18	<.01
25	35.54	16	<.01

As can be seen in *Table 2*, only two items did not display DIF – items 5 and 20. Therefore, this analysis suggests that the rest of the analyses for this project will be done separately for the three training types. The different parameters' estimates (discrimination, scoring coefficients, and endorsement likelihood) for the three groups, are listed in *Appendix A*.

From looking in *Appendix A*, it seems that different loading patterns (discrimination estimates) arise for the three groups. The loading patterns for the combat training group and the non-combat training group seem to be quite similar, whereas the loading pattern for the semi-training group seems different than the other two groups (some are higher, and some are lower, there is no consistent pattern). As for the scoring coefficients (which indicate the categories' ordering), the categories seem to be organized into the same pattern in the semi-combat and the non-combat training groups, whereas they are different in the combat training group. Finally, for the endorsement likelihood parameter (which indicates the likelihood of a category to be endorsed) no meaningful differences were found between the three groups.

Another interesting finding from looking at *Appendix A* is in regard to the suggested answers by the NRM for each item. These can be inferred from the endorsement likelihood values. For each item, the category with the highest endorsement likelihood value is the suggested answer to this item. From looking at the third table of *Appendix A* and comparing it to the SMEs suggested answers key, it seems like the NRM and the CTT answers are quite different. In fact, when looking at the content of the answers chosen by the NRM to be the best answers for some of the items – picking them does not always make sense since there are obviously, from an unprofessional eye, a better way to deal with the situation presented in the item. For example, in item 6 the situation describes a soldier that brings down the platoon's spirit; the respondent is asked how they would deal with it. The answer suggested by the SMEs is to try and understand what is behind the soldier's provocations; the problem may stem from frustration and not from a political stand. Whereas the suggested answer according to the NRM is to clarify to the soldier that his job as a soldier is to act according to a policy that was determined by higher ranks. However, that is not the case for all the items, and for some items the answers picked by the NRM do make sense.

Fitting Different IRT Models to the SJT Data

Different IRT models were estimated in order to calculate latent trait scores for the leadership SJT. Both unidimensional and 3-dimensional IRT models were estimated. Model fit indices are displayed in *Table 3*. The model fit indices that were estimated were M_2 goodness-of-fit (Maydeu-Olivares & Joe, 2006), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMSR), and comparative fit index (CFI). There are no clear guidelines for what values of these indices indicate good fit, but the most common cut-offs for good fit are suggested by Kline (2016): an insignificant M_2 $(p > .05)$, RMSEA < .08, SRMSR < .08, and CFI ≥ .90. In addition, information criteria are provided: Akaike information criterion (AIC), and Bayesian information criterion (BIC); these indices are used mainly to compare between models, and the expectation is that a model with lower values fits better to the data.

***Table 3.* IRT Models Goodness-of-Fit**

Model / Training Type	M_2 *(df)*	*RMSEA (90% C.I)*	*SRMSR*	*CFI*	*AIC*	*BIC*
Combat Training						
Unidimensional 1PL	582.35 (299)**	.026 (.023 - .029)	.038	.107	41583.73	41719.69
Unidimensional 2PL	422.92 (275)**	.020 (.016 - .023)	.031	.534	41499.09	41760.55
Unidimensional 3PL	374.14 (250)**	.019 (.015 - .023)	.031	.609	41524.67	41916.85
Unidimensional NRM	207.72 (125)**	.022 (.017 - .027)	--	.507	71933.68	72979.50
Multidimensional 2PL	281.74 (228)**	.013 (.007 - .018)	.025	.831	41459.48	41966.70
Multidimensional 3PL	234.57 (203)	.011 (.000 - .016)	.025	.900	41470.61	42108.56
Multidimensional NRM	99.99 (78)*	.014 (.002 - .022)	--	.869	71780.46	73072.05
Semi-Combat Training						
Unidimensional 1PL	609.65 (299)**	.030 (.027 - .033)	.042	.069	34266.27	34397.57
Unidimensional 2PL	458.38 (275)**	.024 (.020 - .028)	.036	.450	34182.07	34434.57
Unidimensional 3PL	417.27 (250)**	.024 (.020 - .028)	.036	.499	34220.50	34599.26
Unidimensional NRM	176.19 (127)**	.018 (.011 - .025)	--	.796	59310.08	60310.00
Multidimensional 2PL	305.13 (228)**	.017 (.012 - .022)	.028	.769	34132.03	34621.89
Multidimensional 3PL	274.44 (203)**	.017 (.012 - .023)	.030	.786	34149.96	34766.07
Multidimensional NRM	108.56 (80)*	.018 (.008 - .025)	--	.881	59129.11	60366.39
Non-Combat Training						
Unidimensional 1PL	948.02 (299)**	.029 (.026 - .031)	.035	.057	78541.19	78694.39
Unidimensional 2PL	655.37 (275)**	.023 (.020 - .025)	.028	.447	78337.99	78632.61
Unidimensional 3PL	591.08 (250)**	.023 (.020 - .025)	.028	.504	78382.27	78824.21
Unidimensional NRM	277.20 (125)**	.021 (.018 - .025)	--	.672	135401.7	136580.2
Multidimensional 2PL	341.19 (228)**	.014 (.011 - .017)	.020	.836	78148.28	78719.85
Multidimensional 3PL	305.04 (203)**	.014 (.010 - .017)	.020	.852	78195.03	78913.91
Multidimensional NRM	131.51 (78)**	.016 (.011 - .021)	--	.885	135006.4	136461.9

Note. * $p < .05$, ** $p < .01$. RMSEA = Root Mean Square Error of Approximation, SRMSR = Standardized Root Mean Square Residual, CFI = Comparative Fit Index, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion, 1PL = One Parameter Model, 2PL = Two Parameter Model, 3PL = Three Parameter Model, NRM = Nominal Response Model.

As can be seen in the table above, there are some disagreements on the best fitting model to the SJT data. On the one hand, according to the Chi-Square-based fit indices (*M_2*, RMSRA, SRMSR, and CFI), the best fitting IRT models, in all training groups, are the 3-dimensional 3PL model and the 3-dimensional NRM. The unidimensional models did not provide adequate fit, compared to the multidimensional ones. However, none of the models achieved good fit, as would be suggested by Kline's (2016) cut-off scores in all fit indices. On the other hand, according to the information criteria (AIC and BIC) the best fitting model across all groups is the 3-dimensional 2PL model. The difference between the best models according to the information criteria and the Chi-square-based fit indices could be due to the fact that the information criteria (AIC and BIC) penalize models with more parameters, and this is why more complex models like the NRM and the 3PL model did not achieve the best fit according to those indices.

The main conclusion from this analysis is that the multidimensional models *better fit* the data than the unidimensional ones. This has implications in regard to the question of whether this SJT should be treated as a measurement method or as a construct (this finding supports the SJT as measurement methods point of view). For further discussion about this question, see the discussion section.

In *Figure 2*, below, a demonstration of the IRFs for the first item in the SJT for the combat training group under the various models. For the unidimensional models, the x-axis is the latent trait values and the y-axis is the probability of this latent trait. The multidimensional figures are more complex. In each sub-figure (the small figures that appear in the rows and columns) appears a 3D chart of the intersection between factor 1 and factor 2 (the two horizontal axes are the two latent trait values, and the vertical axis is the probability). The different sub-figures show this 3D chart from a different angle, depending on the value of factor 3.

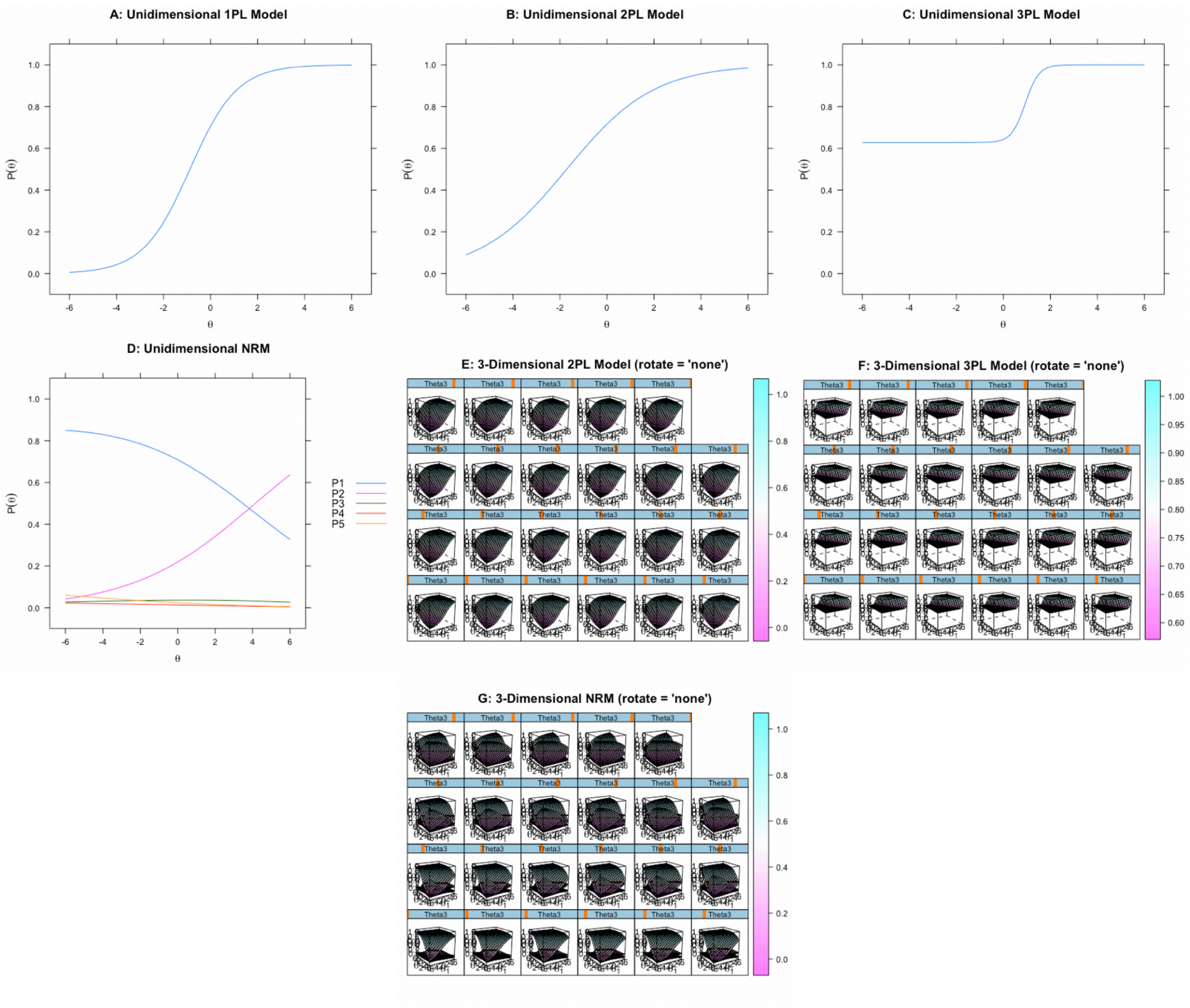

Figure 2. Item Response Functions for Item 1 in the Combat Training Group

In order to further familiarize the reader with the IRT models, and yet not to overwhelm with information, a short description of item parameters estimates is provided for each model.

- **Unidimensional 1PL Models:** Locations vary from -3 to 1.5, not covering the very high range of the latent trait.
- **Unidimensional 2PL Models:** These models are characterized with low discriminations that, for most items, do not surpass |0.5|. Locations look similar to the 1PL models.
- **Unidimensional 3PL Models:** Compared to the 2PL models, in these models more items have discriminations estimates that surpass |0.5|, some even reaching to |5|. As for the locations, it looks like there are not many differences than the unidimensional 1PL locations – covering the lower range of theta pretty well and covering only to the 1.5 point of the latent trait on the higher side. Guessing parameters are mostly low (close to zero) for most items, but for some items they surpass 0.20 (which is the expected guessing based on chance) and even reach levels of 0.6.
- **Unidimensional NRMs:** Discrimination estimates are quite low and do not surpass |0.5| in any of the models (across all training groups). The scoring coefficients vary between the different categories, with category 1 constant at 0 and category 5 constant at 4, they range between -10 and up to 19. The endorsement likelihood estimates range between -6 to 5.5 in the different categories between the different models.
- **Multidimensional 2PL Models:** Factor loadings (discrimination estimates) vary and approximately the same number of items load on each of the three factors. Some items' factor loadings do not surpass |0.5| and do not load on any factor. As for the locations, it seems consistent with previous models – locations vary between -2.5 and 1.5.

- **Multidimensional 3PL Models:** Factor loadings vary and approximately the same number of items load on each of the three factors. Compared to the multidimensional 2PL models, in the 3PL ones most item load on some factors and only a small number of items do not surpass |0.5| for any factor. Locations are consistent with the image portrayed in the previous models. As for the guessing parameters, most items do not surpass the 0.2 threshold, but about 10 items do have higher (and sometimes much higher) guessing parameter value.
- **Multidimensional NRMs:** Factor loadings are quite low and do not surpass |0.5| for most items in all models. The scoring coefficients and the endorsement likelihood parameter estimates look very similar to the unidimensional NRMs.

As for the factor structure, looking at the three multidimensional model across three training groups, I identified two trends. First, in the multidimensional NRM, as I have mentioned previously, none of the factor loadings surpass |0.5|, indicating that none of the items load substantially on any of the factors – making the factors meaningless. Second, in the multidimensional 2PL and 3PL, the three factors are indistinguishable: looking at their content, it looks like they deal with problems with subordinates or with peers (but these do not emerge as two separate factors). From these two trends, we can learn that these multidimensional models yield uninterpretable factor structures and the three factors represent indistinguishable aspects of leadership, and therefore, they will not be named in the following analyses.

Answering RQ 1: Is There Improvement to Criterion-Related Validity?

Before answering RQ1, correlation matrices were computed between the different types of scoring within the 3 training types. The results of these analyses are presented in the following table.

Table 4. ***Correlation Matrices Between Classical and IRT Scoring of Media SJT***

Combat Training

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Original scoring	1													
(2) 1-dim 1PL theta	.905**	1												
(3) 1-dim 2PL theta	.520**	.555**	1											
(4) 1-dim 3PL theta	.314**	.321**	.785**	1										
(5) 1-dim NRM theta	-.438**	-.445**	-.676**	-.538**	1									
(6) 3-dim 2PL F1 theta	.631**	.682**	.758**	.503**	-.658**	1								
(7) 3-dim 2PL F2 theta	.023	.018	.667**	.606**	-.334**	.081**	1							
(8) 3-dim 2PL F3 theta	.307**	.333**	.153**	.197**	.069*	-.005	-.008	1						
(9) 3-dim 3PL F1 theta	-.060*	-.058*	.107**	-.001	-.229**	.327**	.005	-.708**	1					
(10) 3-dim 3PL F2 theta	.330**	.360**	.804**	.771**	-.489**	.488**	.671**	.247**	-.065*	1				
(11) 3-dim 3PL F3 theta	.382**	.397**	.082**	-.002	-.167**	.485**	-.524**	.252**	-.139**	-.075**	1			
(12) 3-dim NRM F1 theta	.480**	.517**	.469**	.306**	-.599**	.677**	-.040	.020	.208**	.288**	.384**	1		
(13) 3-dim NRM F2 theta	.138**	.150**	.572**	.455**	-.198**	.175**	.662**	.167**	-.077**	.530**	-.273**	.015	1	
(14) 3-dim NRM F3 theta	.055*	.027	.122**	.157**	-.549**	.143**	.098**	-.234**	.177**	.084**	-.007	-.086**	-.244**	1

Note. * p < .05, ** p < .01. dim = dimensional, F1, F2, F3 = Factor 1, Factor 2, Factor 3.

Semi-Combat Training

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Original scoring	1													
(2) 1-dim 1PL theta	.908**	1												
(3) 1-dim 2PL theta	.549**	.602**	1											
(4) 1-dim 3PL theta	.522**	.567**	.875**	1										
(5) 1-dim NRM theta	-.425**	-.429**	-.473**	-.446**	1									
(6) 3-dim 2PL F1 theta	-.167**	-.193**	-.084**	.013	.354**	1								
(7) 3-dim 2PL F2 theta	-.541**	-.589**	-.979**	-.873**	.429**	-.017	1							
(8) 3-dim 2PL F3 theta	-.111**	-.098**	-.040	-.051	.330**	.032	-.024	1						
(9) 3-dim 3PL F1 theta	-.275**	-.310**	-.652**	-.663**	.251**	-.087**	.663**	-.032	1					
(10) 3-dim 3PL F2 theta	.296**	.335**	.417**	.321**	-.438**	-.617**	-.338**	-.087**	-.040	1				
(11) 3-dim 3PL F3 theta	-.045	-.024	-.044	-.090**	.216**	-.010	.006	.692**	.066*	-.050	1			
(12) 3-dim NRM F1 theta	.036	.050	.294**	.304**	.381**	.562**	-.367**	.376**	-.319**	-.289**	.248**	1		
(13) 3-dim NRM F2 theta	.153**	.190**	.365**	.304**	-.052	-.263**	-.321**	-.080**	-.136**	.288**	-.097**	-.084**	1	
(14) 3-dim NRM F3 theta	.466**	.480**	.634**	.588**	-.725**	-.012	-.634**	-.166**	-.472**	.254**	-.084**	.158**	-.164**	1

Note. * p < .05, ** p < .01. dim = dimensional, F1, F2, F3 = Factor 1, Factor 2, Factor 3.

Non-Combat Training

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Original scoring	1													
(2) 1-dim 1PL theta	.908**	1												
(3) 1-dim 2PL theta	.529**	.567**	1											
(4) 1-dim 3PL theta	.505**	.539**	.931**	1										
(5) 1-dim NRM theta	-.466**	-.468**	-.760**	-.712**	1									
(6) 3-dim 2PL F1 theta	.301**	.297**	.412**	.370**	-.494**	1								
(7) 3-dim 2PL F2 theta	-.408**	-.454**	-.793**	-.754**	.501**	.153**	1							
(8) 3-dim 2PL F3 theta	-.165**	-.193**	.170**	.148**	-.063**	-.003	-.097**	1						
(9) 3-dim 3PL F1 theta	.063**	.044*	.432**	.406**	-.345**	.397**	-.151**	.852**	1					
(10) 3-dim 3PL F2 theta	-.333**	-.376**	-.621**	-.635**	.364**	.303**	.872**	-.143**	-.174**	1				
(11) 3-dim 3PL F3 theta	-.414**	-.433**	-.508**	-.457**	.519**	-.772**	.143**	.429**	.041*	-.146**	1			
(12) 3-dim NRM F1 theta	-.522**	-.530**	-.701**	-.658**	.843**	-.473**	.510**	.249**	-.069**	.339**	.661**	1		
(13) 3-dim NRM F2 theta	-.007	.006	.388**	.350**	-.185**	-.039*	-.378**	.640**	.563**	-.366**	.201**	.065**	1	
(14) 3-dim NRM F3 theta	.173**	.198**	.217**	.224**	.023	-.363**	-.483**	-.147**	-.217**	-.455**	.067**	-.146**	.210**	1

Note. $^{*}p < .05$, $^{**}p < .01$. dim = dimensional, F1, F2, F3 = Factor 1, Factor 2, Factor 3.

From the matrices presented in the table above it seems that the score that is the most correlated with the original scoring, across all training types, is the unidimensional 1PL theta. As for the multidimensional models' scores, they seem to be negatively correlated or correlated to a low degree with the original scoring – indicating that they measure something distinct. As for the factor scores in the multidimensional methods, the different factors within each IRT model seem to be uncorrelated, even though an Oblimin rotation was used. As for the correlations between thetas from different IRT models – these seem to be moderate, indicating some similarities between the different scores.

Research Question 1 asked whether there will be an improvement using the different IRT models in criterion-related validity of the leadership SJT. In order to assess that, Pearson correlations were calculated between the traditional CTT scoring ("original" scoring) and the IRT-based theta scores and the criterion (officer training course grades). The results of this analysis are presented in *Table 5*. This table also presents the *z*-score of the difference between each IRT model scoring and the traditional score within each training type.

***Table 5.* Pearson Correlations Between Media Scores and Criterion Variable**

Predictor / Training Type	Correlation with Training Grade	*Z*-Score of the Difference from Original Scoring
Combat Training (*N*=1291)		
Original scoring	.174**	
Unidimensional 1PL theta	.131**	1.111
Unidimensional 2PL theta	.096**	2.009*
Unidimensional 3PL theta	.091**	2.137
Unidimensional NRM theta	-.097**	6.960**
3-dimensional 2PL factor 1 theta	.096**	2.009*
3-dimensional 2PL factor 2 theta	.004	4.359**
3-dimensional 2PL factor 3 theta	.074**	2.571*
3-dimensional 2PL weighted score[1]	.122**	1.342
3-dimensional 3PL factor 1 theta	-.056*	5.898**
3-dimensional 3PL factor 2 theta	.087**	2.239*
3-dimensional 3PL factor 3 theta	.104**	1.804
3-dimensional 3PL weighted score	.130**	1.137
3-dimensional NRM factor 1 theta	.087**	2.239*
3-dimensional NRM factor 2 theta	.062*	2.878**
3-dimensional NRM factor 3 theta	.000	4.461**
3-dimensional NRM weighted score	.107**	1.727
Semi-Combat Training (*N*=1070)		
Original scoring	.047	
Unidimensional 1PL theta	.039	.185
Unidimensional 2PL theta	.070*	-.533
Unidimensional 3PL theta	.097**	-1.16
Unidimensional NRM theta	-.078*	2.894**
3-dimensional 2PL factor 1 theta	.048	-.023
3-dimensional 2PL factor 2 theta	-.069*	2.685**
3-dimensional 2PL factor 3 theta	-.054	2.336*
3-dimensional 2PL weighted score	-.057	2.406*
3-dimensional 3PL factor 1 theta	-.058	2.429*
3-dimensional 3PL factor 2 theta	.015	.740
3-dimensional 3PL factor 3 theta	-.041	2.035*
3-dimensional 3PL weighted score	-.067*	2.638**
3-dimensional NRM factor 1 theta	.007	.925
3-dimensional NRM factor 2 theta	-.023	1.618
3-dimensional NRM factor 3 theta	.064*	-.394
3-dimensional NRM weighted score	.061*	-.324
Non-Combat Training (*N*=2560)		
Original scoring	.092**	
Unidimensional 1PL theta	.081**	.396
Unidimensional 2PL theta	.056**	1.293
Unidimensional 3PL theta	.061**	1.113
Unidimensional NRM theta	-.069**	5.779**

Predictor / Training Type	Correlation with Training Grade	*Z*-Score of the Difference from Original Scoring
3-dimensional 2PL factor 1 theta	.012	2.869**
3-dimensional 2PL factor 2 theta	-.066**	5.671**
3-dimensional 2PL factor 3 theta	-.103**	7.012**
3-dimensional 2PL weighted score	-.121**	7.668**
3-dimensional 3PL factor 1 theta	-.077**	6.068**
3-dimensional 3PL factor 2 theta	-.046*	4.950**
3-dimensional 3PL factor 3 theta	-.074**	5.960**
3-dimensional 3PL weighted score	-.125**	7.814**
3-dimensional NRM factor 1 theta	-.097**	6.793**
3-dimensional NRM factor 2 theta	-.081**	6.213**
3-dimensional NRM factor 3 theta	.034	2.081*
3-dimensional NRM weighted score	-.102**	6.975**

Note. * $p < .05$, ** $p < .01$.

[1] Weighted scores are linear-regression based weights with all 3 factors inserted as the independent variables.

Overall, it seems that the Media SJT had low criterion-related validity with this specific criterion. It is important to note that some *Z*-scores are quite large; when this was the case, it is because the comparison was between a positive correlation and a negative correlation – the change of sign causes a large gap between the correlation, even though their magnitudes (in absolute values) are quite similar. When trying to compare the different scoring methods, it seems that the results of this analysis differ based on the training type. Among participants who were targeted to go through combat training, the highest correlation between the SJT and the criterion was achieved when the SJT score was calculated using the traditional method. The IRT-based scores achieved lower correlations than the traditional scoring, or at least insignificantly different correlations than the traditional scoring. Among participants who were targeted to go through semi-combat training, the highest correlation between the SJT and the criterion was achieved when the SJT score was calculated using the unidimensional 3PL IRT model. However, this correlation is not significantly different than the correlation between the traditionally scored SJT and the criterion. Lastly, among participants who were targeted to go through non-combat training, the highest correlation between the SJT and the criterion was achieved when the SJT score was calculated using the multidimensional 3PL IRT model. This correlation was also found to be significantly different than the correlation between the traditionally scored SJT and the criterion. However, the correlation between the multidimensional 3PL average theta and the criterion is in the wrong direction, indicating that participants who achieved higher scores on the latent traits performed worse on the criterion.

The primary conclusion from these analyses is that the IRT methods of scoring do not improve on the traditional method of scoring in the case of criterion-related validity.

Answering RQ 2: Does the SJT Provide Incremental Validity?

Research Question 2 asked whether the SJT scores will provide incremental validity in predicting the criterion above and beyond personality measures and cognitive ability, and specifically, whether the IRT-scored SJT will provide that incremental validity. In order to answer this question, a series of hierarchical linear regressions were conducted. In the first step of the regression, the IPR (general mental ability score) and the 5 OPI scores (each for every personality trait of the Big Five) were entered. In the second step of the regression, the Media score was entered: in the case of the traditional scoring and the unidimensional IRT models, the score/theta was entered alone, and in the case of the multidimensional IRT models, the three thetas were entered in this step. In order to assess whether there is incremental validity, the second step of the regression analysis should show a significant increase in R^2. The results of the regression analyses are presented in *Table 6*.

Table 6. **Regression Analyses of Cognitive Ability, Personality, and SJT to Training Grade**

Predictor(s) / Change in R^2	*Combat Training*		*Semi-Combat Training*		*Non-Combat Training*	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
Cognitive ability (IPR)	.219**	.198**	.117**	.114**	.114**	.105**
OPI – agreeableness	.138**	.125**	.067	.066	-.009	-.015
OPI – conscientiousness	.046	.052	.038	.039	.011	.015
OPI – extraversion	-.026	-.016	-.019	-.019	.000	.002
OPI – neuroticism	.042	.055	.042	.043	-.056*	-.055*
OPI – openness to experience	.028	.025	-.007	-.008	.020	.016
Media traditional scoring		.146**		.032		.084**
Change in R^2	.062**	.021**	.017**	.001	.016**	.017**
Cognitive ability (IPR)	.219**	.206**	.117**	.115**	.114**	.108**
OPI – agreeableness	.138**	.131**	.067	.065	-.009	-.012
OPI – conscientiousness	.046	.048	.038	.039	.011	.013
OPI – extraversion	-.026	-.019	-.019	-.020	.000	.001
OPI – neuroticism	.042	.052	.042	.043	-.056*	-.055*
OPI – openness to experience	.028	.024	-.007	-.007	.020	.016
Media 1-dim 1PL theta		.103**		.022		.075**
Change in R^2	.062**	.010**	.017**	.000	.016**	.006**
Cognitive ability (IPR)	.219**	.215**	.117**	.111**	.114**	.111**
OPI – agreeableness	.138**	.131**	.067	.062	-.009	-.014
OPI – conscientiousness	.046	.045	.038	.037	.011	.014
OPI – extraversion	-.026	-.019	-.019	-.017	.000	-.001
OPI – neuroticism	.042	.048	.042	.042	-.056*	-.057*
OPI – openness to experience	.028	.025	-.007	-.011	.020	.017
Media 1-dim 2PL theta		.074**		.052		.051*
Change in R^2	.062**	.005**	.017**	.003	.016**	.003*
Cognitive ability (IPR)	.219**	.215**	.117**	.108**	.114**	.111**
OPI – agreeableness	.138**	.134**	.067	.062	-.009	-.016
OPI – conscientiousness	.046	.046	.038	.036	.011	.015
OPI – extraversion	-.026	-.020	-.019	-.016	.000	-.001
OPI – neuroticism	.042	.048	.042	.042	-.056*	-.057*
OPI – openness to experience	.028	.025	-.007	-.014	.020	.017

Predictor(s) / Change in R^2	*Combat Training*		*Semi-Combat Training*		*Non-Combat Training*	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
Media 1-dim 3PL theta		.070*		.084**		.057**
Change in R^2	.062**	.005*	.017**	.007**	.016**	.003**
Cognitive ability (IPR)	.219**	.212**	.117**	.109**	.114**	.107**
OPI – agreeableness	.138**	.131**	.067	.061	-.009	-.014
OPI – conscientiousness	.046	.048	.038	.046	.011	.019
OPI – extraversion	-.026	-.022	-.019	-.018	.000	-.001
OPI – neuroticism	.042	.044	.042	.039	-.056*	-.056*
OPI – openness to experience	.028	.023	-.007	-.015	.020	.014
Media 1-dim NRM theta		-.063*		-.067*		-.065**
Change in R^2	.062**	.004*	.017**	.004*	.016**	.004**
Cognitive ability (IPR)	.219**	.208**	.117**	.104**	.114**	.097**
OPI – agreeableness	.138**	.134**	.067	.064	-.009	-.011
OPI – conscientiousness	.046	.049	.038	.039	.011	.011
OPI – extraversion	-.026	-.017	-.019	-.015	.000	.011
OPI – neuroticism	.042	.053	.042	.042	-.056*	-.049
OPI – openness to experience	.028	.023	-.007	-.017	.020	.007
Media 3-dim 2PL factor 1 theta		.064*		.039		.021
Media 3-dim 2PL factor 2 theta		.004		-.054		-.072**
Media 3-dim 2PL factor 3 theta		.061*		-.044		-.101**
Change in R^2	.062**	.007*	.017**	.006	.016**	.013**
Cognitive ability (IPR)	.219**	.207**	.117**	.112**	.114**	.099**
OPI – agreeableness	.138**	.131**	.067	.062	-.009	-.010
OPI – conscientiousness	.046	.048	.038	.043	.011	.008
OPI – extraversion	-.026	-.017	-.019	-.020	.000	.010
OPI – neuroticism	.042	.052	.042	.040	-.056*	-.050
OPI – openness to experience	.028	.023	-.007	-.009	.020	.010
Media 3-dim 3PL factor 1 theta		-.044		-.028		-.080**
Media 3-dim 3PL factor 2 theta		.070*		.022		-.068**
Media 3-dim 3PL factor 3 theta		.079**		-.035		-.068**
Change in R^2	.062**	.013**	.017**	.003	.016**	.013**
Cognitive ability (IPR)	.219**	.215**	.117**	.109**	.114**	.091**

Predictor(s) / Change in R^2	*Combat Training*		*Semi-Combat Training*		*Non-Combat Training*	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
OPI – agreeableness	.138**	.131**	.067	.066	-.009	-.006
OPI – conscientiousness	.046	.044	.038	.043	.011	.020
OPI – extraversion	-.026	-.018	-.019	-.015	.000	.007
OPI – neuroticism	.042	.054	.042	.045	-.056*	-.049
OPI – openness to experience	.028	.025	-.007	-.014	.020	.004
Media 3-dim NRM factor 1 theta		.058*		-.007		-.083**
Media 3-dim NRM factor 2 theta		.065*		-.009		-.079**
Media 3-dim NRM factor 3 theta		-.002		.045		.031
Change in R^2	.062**	.008*	.017**	.002	.016**	.013**

Note. * $p < .05$, ** $p < .01$. IPR = Initial Psychotechnical Rating, OPI = Officer Personality Inventory, dim = Dimensional, 1PL = One Parameter Model, 2PL = Two Parameter Model, 3PL = Three Parameter Model, NRM = Nominal Response Model.

From *Table 6* it seems that the results are consistent across the different types of scoring – both traditional and IRT-based. For combat training and non-combat training, it appears that the SJT scores have incremental validity above and beyond cognitive ability and the Big Five personality traits. However, for participants designated to go through semi-combat training, the SJT scores did not provide incremental validity, apart from one regression, when the SJT scores were based on the latent trait scores that were computed using the unidimensional nominal response model – in that case, for this group of participants, the SJT scores had incremental validity.

As for the models that provide the highest change in R^2 (which indicates the highest incremental validity), it looks like the traditional scoring provides the highest incremental validity above and beyond cognitive ability and personality traits, whereas the different IRT models provide similar (and lower than the traditional scores) R^2 values.

Answering RQ 3: Does the SJT Show Reduced Subgroup Differences?

Research Question 3 asked whether the SJT scores will reduce subgroup differences between genders. *Table 7* presents the frequencies of men and women within each training group.

***Table 7.* Gender Frequencies Within Different Training Groups**

Training Type	Gender	*N*	Percent
Combat Training	Male	1344	97.5%
	Female	35	2.5%
Semi-Combat Training	Male	527	45.7%
	Female	626	54.3%
Non-Combat Training	Male	1230	45.9%
	Female	1446	54.0%

As can be seen in the table above, the number of women in the combat training group is too small in order to make meaningful comparisons between men and women. Therefore, the next set of analyses will be conducted only on the semi-combat training and the non-combat training groups.

In order to answer this research question, Cohen's *d* (Cohen, 1988) scores were calculated to estimate gender's effect size on the SJT scores. The results of this analysis are presented in *Table 8*.

As can be seen in Table 8, the direction of the gender effect is not consistent across models and across groups. Some effects are positive, indicating that men perform better on the SJT than women, but most *d*'s are negative, indicating that women perform better than men on the SJT. As for the effect sizes, the results differ between the two training groups. For the semi-combat training, the smallest effect sizes were achieved for the average theta of the multidimensional 2PL model, and for the second factor of the multidimensional 3PL model. For the same group, the largest effect sizes were demonstrated in the unidimensional NRM's theta and for factor 3 of the multidimensional NRM model. As for the non-combat training the smallest effect size was achieved for the average theta of the multidimensional NRM, while the largest effect size was achieved for the second factor of the multidimensional 3PL model.

There are two main conclusions that can be drawn from this analysis. First, most of the effects are in the opposite direction than expected, indicating that the minority group (women) perform better on this SJT than the majority group. Second, there is no one IRT model (or the traditional scoring) that shows consistently across groups the least subgroup differences. Therefore, one cannot recommend the best approach to calculating scores that will achieve the least subgroup differences between men and women.

Table 8. **Gender Differences on Different SJT Scores**

	Semi-Combat Training					Non-Combat Training				
	Men		Women			Male		Female		
Scoring Method	Mean	*S*	Mean	*S*	**Cohen's *d***	Mean	*S*	Mean	*S*	**Cohen's *d***
Original Scoring	30.962	4.508	31.278	3.675	**-.077**	31.294	4.137	31.578	3.854	**-.071**
1-dim 1PL theta	-.023	.548	.021	.447	**-.088**	-.018	.511	.019	.482	**-.075**
1-dim 2PL theta	.012	2.053	.554	2.098	**-.261**	.042	1.907	.430	1.997	**-.199**
1-dim 3PL theta	-.220	1.871	.285	1.724	**-.281**	-.026	1.844	.316	1.927	**-.181**
1-dim NRM theta	.057	1.489	-.382	1.506	**.293**	-.095	1.541	-.325	1.542	**.149**
3-dim 2PL factor 1 theta	-.687	3.568	-.337	3.506	**-.099**	-.162	4.071	-1.063	4.167	**.219**
3-dim 2PL factor 2 theta	-.119	3.594	-1.191	3.750	**.292**	-.373	3.947	-1.753	4.374	**.331**
3-dim 2PL factor 3 theta	.028	4.121	1.015	4.025	**-.242**	-.210	4.024	.070	4.010	**-.070**
3-dim 2PL average theta	-.259	2.211	-.171	2.158	**-.040**	-.248	2.304	-.915	2.497	**.278**
3-dim 3PL factor 1 theta	.617	3.635	-.206	3.351	**.236**	-.048	3.876	.042	4.091	**-.022**
3-dim 3PL factor 2 theta	.213	3.828	.163	3.094	**.014**	-.520	4.303	-2.105	4.818	**.347**
3-dim 3PL factor 3 theta	.391	3.539	.716	3.296	**-.095**	-.111	3.784	.358	3.575	**-.127**
3-dim 3PL average theta	.407	2.210	.224	1.759	**.091**	-.226	2.052	-.568	2.184	**.161**
3-dim NRM factor 1 theta	-.226	3.068	.342	2.960	**-.188**	-.447	2.815	-.831	2.858	**.135**
3-dim NRM factor 2 theta	.574	2.923	.091	2.992	**.163**	.016	3.076	.157	2.811	**-.048**
3-dim NRM factor 3 theta	.165	3.031	1.212	3.263	**-.333**	.364	3.330	.642	3.269	**-.084**
3-dim NRM average theta	.171	1.678	.549	1.737	**-.221**	-.023	1.908	-.011	1.752	**-.006**

Note. S = Standard Deviation, dim = Dimensional 1PL = One Parameter Model, 2PL = Two Parameter Model, 3PL = Three Parameter Model, NRM = Nominal Response Model.

DISCUSSION

Although SJTs have been popular for at least 3 decades, a significant question remains about how best to score them. The goal of this study was to evaluate the usefulness of applying an advanced scoring method, item response theory, to situational judgement tests. The usefulness of IRT in scoring SJTs was evaluated using the standards established by Bergman et al. (2006): by establishing criterion-related validity (RQ1), establishing incremental validity of the SJT scores above and beyond personality and cognitive measures (RQ2), and by assessing whether the SJT scores minimize subgroup differences (RQ3). A secondary question that was evaluated (which was also related to Bergman et al.'s standards) was whether SJTs should be treated as a construct or as a measurement method. In this section, a summary of the project's results will be presented, together with limitations, suggestions for further research and final conclusions.

This study's first research question was whether the use of IRT will improve the SJTs criterion-related validity. Overall, the results suggest low criterion-related validity with officer training course final grades. This could be due to several reasons, but probably mainly due to the poor quality of the criterion (see the limitations section below). In regard to improvements due to the use of IRT models, different results were observed dependent on the type of training. For the combat training group, none of the IRT models provided an improved validity compared to the original scoring. The opposite has occurred – some of the IRT-based scores presented a significant *decrease* in validity. For the semi-combat training group, most IRT models (the unidimensional 2PL model, the unidimensional 3PL model, the unidimensional NRM, the multidimensional 3PL model, and the multidimensional NRM) provided better validity compared to the original traditional scoring. Finally, for the non-combat training group, only the multidimensional IRT models-based scores provided improved validity, whereas the unidimensional models showed a significant decrease in validity compared to the original scoring.

As for the answer for this research question, it seems that this research has not changed what was already known in the literature (Wright, 2013; Zu & Kyllonen, 2018) – there is uncertainty whether IRT models are more useful than the CTT-based scoring. However, one finding that is worth mentioning is the fact that the multidimensional IRT models performed better (validity-wise) in some cases. Therefore, the main conclusion is that when using a multidimensional SJT, it is better to score it using multidimensional IRT models than to use CTT-based scoring or unidimensional IRT models. Further support for this conclusion can be found in the fit indices' values of the multidimensional models – these fit the data better than the unidimensional ones in all three groups. This model fit finding makes sense, considering that this specific SJT was designed to measure three constructs. However, one interesting finding was that in the multidimensional NRMs factor loadings (discriminations) on all three factors did not surpass the threshold of |0.5|, thus indicating that in the nominal model, the factors were not meaningful. Also, in the other multidimensional models (the 2PL and the 3PL), the factors were indistinguishable, with items measuring different aspects of leadership loading on the same factors. Another finding that supports the lack of usability of the nominal model for the SJT assessed in this study is the answers suggested as the "correct answers" by the model. Some of these answers did not make sense based on their content. These findings suggest that while it is advised to use multidimensional models for multidimensional SJTs, it is important to test these models' fit and also their factor structures. If the factor structure is uninterpretable (like in this case), or the fit indices are not satisfactory, then using the traditional method of scoring might be preferred.

The second research question that this study raised was whether the IRT-based scores of the SJT will provide incremental validity above and beyond that of personality measures and cognitive ability. The results suggested that for non-combat and combat training groups, the SJT scores – whether they were IRT-based or CTT-based – provided some incremental validity. However, in the semi-combat training group, only the unidimensional NRM IRT-based scores provided incremental validity above and beyond personality traits and cognitive ability. This finding provides further support to previous findings (e.g., Clevenger et al., 2001; Weekley & Ployhart, 2005) that SJTs capture something different than traditional predictors of job performance. As for differences in the amount of incremental validity added due to the use of IRT models, it turns out that the traditional scoring method actually provides more incremental validity than the IRT-based scoring – both unidimensional and multidimensional. This could be due to the fact that the IRT models did not reach optimal fit to the data and therefore the loss of fit might also be related to the loss in incremental validity.

The third research question in this study was whether SJT scores will reduce subgroup differences between genders. This question was not assessed on the combat training group due to small number of women. However, in the two other groups the results seemed to be inconsistent. Some effects were positive (indicating that men perform better than women) and some were negative (indicating that women perform better than men). As for the effect sizes, there was no one IRT model that provided smaller effect sizes for both groups. However, the multidimensional IRT models provided smaller effect sizes than the unidimensional ones.

Looking at the three research questions, it seems that several themes come up from this study's results. First, when using an SJT that is *designed* to measure several constructs – using a multidimensional model provides better validity, to some extent. Second, overall, the IRT models did not show superiority over the CTT-based scores in this SJT. In fact, in some cases (like criterion-related validity in some of the groups), the IRT-based scores showed worse results. There could be many reasons to why this pattern was demonstrated in this study. Two plausible explanations are the suboptimal fit of the IRT models to the data and the low discrimination (factor loading) that many items showed in most IRT models. This indicates that although the IRT method has received vast support, it might not be useful for purely nominal data like the one that is produced in SJTs. Third, the fact that the multidimensional IRT models (specifically, the multidimensional 3PL model and the multidimensional NRM) performed better and fit better to the data than the unidimensional helps us get a better direction to an answer to a question that was raised a long time ago – whether SJTs are a construct or a measurement method (Arthurt & Villado, 2008). The proponents of the SJT-as-constructs approach suggest that SJTs measures "tacit knowledge" (Schmidt & Hunter, 1993), and therefore SJT-derived data should be treated as unidimensional. On the other side, the proponents of the SJT-as-measurement-method suggest that SJTs can measure any construct (Schmitt & Chan, 2006; McDaniel & Nguyen, 2001) and therefore, if an SJT is designed to measure several constructs, then its data should be treated as multidimensional and no one single factor will explain enough variance to support a unidimensional point of view. *Table 3* clearly shows the superiority of the multidimensional IRT models over their unidimensional counter parts. Thus, this study provides some support to the SJTs-as-measurement-method point of view, however, one should note that good fit of the multidimensional models to the data is required but not sufficient to support the SJT-as-measurement-method point of view. One also needs interpretable factors that are differentiated and make sense – something that was not achieved in this study.

Implications

According to my review of peer-reviewed articles and published dissertations, not much research has been done on applying IRT to SJTs. This could be due to two reasons. One, is that IRT is a rather new practice and it is yet to be tested on SJTs, which is unlikely, because both have existed for three decades. Two, which is more plausible in my opinion, is that research on the topic has been done, but it reached inconclusive results that made it not acceptable for publication in peer-reviewed journals. This study was another attempt at applying IRT models to scoring and interpretation of SJT data. The main contribution over other studies, was by using multidimensional IRT models, and specifically, the multidimensional nominal response model. The use of multidimensional models broadens our knowledge on how SJTs work and what do they measure. The first implication of this study is that SJTs should be used as measurement methods, at least according to some support to that point of view that was achieved in this project. This means, that in practice, I-O psychologists who design an SJT should first think of the constructs they intend to measure. This could be theory-driven or based on job analysis. For scholars, this implication means that SJTs should not be treated as one selection instrument. They should be researched in the context of the constructs that they intend to measure.

A second implication that can be taken from this study is that the use of IRT models to score and interpret SJTs should be limited to cases in which the SJT is multidimensional and that it is proven that multidimensional IRT models improve predictive validity. As can be learned from this study, IRT models did not always yield an increase in predictive validity, and that sometimes the traditional scoring performed better in producing predictive validity and incremental validity over other common measures.

Finally, it was found that the SJT used for this study showed low to medium (depending on the scoring method) subgroup differences between men and women, favoring the women. This finding is in line with previous findings showing no difference or a slight difference (favoring women) in performance in SJTs (Ployhart & Holtz, 2008). This could be significant when designing a selection battery and considering adding an SJT as one of the selection instruments. Adding an SJT to a selection battery helps in reducing gender subgroup differences and adverse impact against women. This is a major advantage in some cases, especially when the job requires one to use more discriminating selection instruments (like physical ability exams).

Limitations

This study is not without its limitations. First, the SJT that was used in this project was designed specifically to assess the candidates' ability to choose effective leadership behaviors in a military setting. As such, it is not certain that the results achieved for this SJT will be similar to results achieved with other SJTs, maybe that are used in other settings relevant to the common workplace. However, it is worth noting that in the two studies that applied IRT to SJTs (Wright, 2013; Zu & Kyllonen, 2018), similar results were found, thus indicating that maybe this SJT, despite of its unique content, is no different than other SJTs used in more traditional selection systems.

Second limitation to this study is in comparing between the traditional scoring method (CTT) and the IRT-based scoring. In particular, the traditional scoring method requires the candidates to choose the best option out of the five and the *second-best* option, and thus enables partial credit if the answer chosen by the group of SMEs as the most effective one is chosen as second best by the candidate. Contrary to that, the IRT-based scoring did not enable partial credit. There are specific IRT models that can accommodate partial credit (like the generalized partial credit model), but they require specific ordering in all answer options – which is unavailable in our case. This creates an unfair comparison, in which the traditional scoring has more information than the IRT-based scoring, so the poorer results achieved for the IRT models in some cases can be justified by lacking the amount of information that is available to the traditional scoring. However, it is important to mention that the possibility of scoring the SJT in the traditional way, using only the best option (giving 1 point for choosing the same option as the group of SMEs, and no points for any other option, without considering the "second-best" option) was explored, and it was found that there was almost perfect correlation (r = .91 for all three training groups) between this method of scoring and the method that was used in the analyses presented in the Results section. This suggests that there will not be a major difference utilizing a CTT-based scoring using the same amount of information as the IRT models.

Another limitation to this study is in the use of a criterion of poor quality. In this study, the final grade of officer training course was used as criterion. This grade is a composite of several components, some of them are essentially unrelated to the purpose of the selection battery that the Media SJT is taken from. This selection battery's goal is to check for the fit of the personality of the candidate to be an officer in the Israeli army. However, the training school final grade includes components like navigation capabilities, knowledge of Israeli history, and others, that are defiantly unrelated to the personality of the subjects. Maybe, if another criterion was available, one that is aimed at measuring the same thing as the selection battery, and specifically, the SJT, we could have seen better relationship, and especially better criterion-related validity.

One final limitation in this study, that might also explain the differences between the three training groups is differential restriction of range, that could not be overcome in this study.

First restriction of range, that applies to all training groups, is the fact the candidates who participate in this selection process are pre-screened based on their cognitive abilities and performance in their initial service in the army. Second, some candidates (mainly ones that participate in the combat training) are also screened based on their performance in commanding courses (which are aimed at bettering their leadership skills) – this restriction of range does not apply to the non-combat training group, and to some of the semi-combat training subjects. In the ideal scenario, I would be able to correct the statistical estimates for restriction of range, but this requires data for the entire population of potential participants in this selection process – which is not available. The data in this study, as mentioned in the Method section, is only of participants who successfully passed the selection process (meaning – after the pre-screening and after the screening that resulted from this selection process).

Future Research

Suggestion for further research of the relevancy of IRT models to SJTs are derived from the limitations section. First, I would suggest fitting IRT models and comparing them to CTT-based scoring in a variety of SJTs that assess different constructs, to assess whether there are consistent results across different types of SJT. If there are consistent results – then we can conclude something important about SJTs as a measurement method. If there are no consistent results – then it might be worth exploring what constructs are better assessed using IRT and which are better assessed using CTT.

Future research can also utilize an SJT with more detailed answer key, specifically, an answer key in which all of the response options are ranked in their efficiency. This way, partial credit models can be used and might achieve better results than IRT models that use less information in their estimation. However, it is worth noting that the generalized partial credit model has been used before in this context (Zu & Kyllonen, 2018), and achieved similar results to the regular logistic models (1PL, 2PL, and 3PL). Therefore, it is not clear if more information about the scores can actually benefit in terms of preferring to use IRT-based scores over CTT-based scores.

Finally, it is worth working with multiple criteria that encompass variety of aspects of fit of the participants to the job we intend to measure. The criteria should be standardized, objective as much as possible, and related to the predictors. In addition, future research should seek to have the entire pool of applicants, so a correction to restriction of range, if present, can be applied.

Final Conclusions

One purpose of this study was to discover whether the use of IRT models to score SJTs can benefit us in terms of validity (construct and predictive) and reducing subgroup differences. The results suggest that there is no one definite answer. As was demonstrated in this study, results varied by the specific sample that was used (combat training, semi-combat training, and non-combat training samples), and in some samples the IRT-based scores achieved better results and in others it did not. The main conclusion from this research, is that classical test theory is not a “dead horse” (Zickar & Broadfoot, 2009) yet – in some cases it was found to be superior to IRT and future research should explore when it is better to use CTT to score tests and questionnaires, and when it is better to use IRT-based scoring.

A second goal of this study was to try and determine an answer in the long debate of whether SJTs are a construct or a measurement method. This study used multidimensional IRT to approach this question. The results of this study support the SJT-as-measurement-method approach, and thus suggest that SJTs should be compared based on the construct that they are measuring and should not be treated as one construct estimating “tacit knowledge”.

REFERENCES

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology, 73*, 305-307.

Arthur Jr, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, *99*(3), 535.

Arthur Jr, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435.

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Brooks, M. E., & Highhouse, S. (2006). Can good judgement be measured? In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement and Application* (pp. 39-55). Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, J. P. (1990) Modeling the performance prediction problem in industrial and organizational psychology. In M.D. Dunnette and L.M. Hough (Eds), *Handbook of industrial and organizational psychology*, Vol. 1 (pp. 687–732). Palo Alto: Consulting Psychologists Press.

Chalmers, P. (2015, June 24th). RE: nominal response model [Online discussion group]. Retrieved from: https://groups.google.com/forum/#!topic/mirt-package/0rg6hDqkRVU.

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

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

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83-117.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Crobnach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulleting, 52,* 281-302.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10*, 94–96.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

Fine, S., Goldenberg, J., & Noam, Y. (2016). Integrity testing and the prediction of counterproductive behaviours in the military. *Journal of Occupational and Organizational Psychology, 89,* 198-218.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327.

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: lessons learned and paths forward. *Organizational Research Methods*, *20*(3), 465-486.

Hanson, M. A., Borman, W. C., Mogilka, H. J., & Manning, C. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment*. Mahwah, NJ: Lawrence Earlbaum Associates.

Hogan, J. B. (1994) Empirical keying of background data measures. In G.S. Stokes, M.D. Mumford and W.A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69–107). Palo Alto: Consulting Psychologists Press.

Hough, L., & Paullin, C. (1994) Construct-oriented scale construction: The rational approach. In G.S. Stokes, M.D. Mumford and W.A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 109–145). Palo Alto: Consulting Psychologists Press.

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2017). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology, 90,* 1-27.

Kenny, D. A. (2009). Founding Series Editor Note in: de Ayala, R. J. (2009). The theory and practice of item response theory. In Little, T.D (Series Ed.) *Methodology in the Social Sciences*, New York, NY: The Guilford Press.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th edition). New York, NY: The Guilford Press.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgement in situational judgement tests? *Journal of Applied Psychology, 100,* 399-416.

LaHuis, D. M., Clark, P., & O'Brien E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods, 14*, 10-23.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgement tests: From measures of situational judgement to measures of general domain knowledge. *Industrial and Organizational Psychology, 9,* 3-22.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Maydeu-Olivares, A., & Joe. H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713-732.

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

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91.

McDaniel, M. A., Morgerson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgement tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 80,* 730-740.

McDaniel, M. A., Nguyen, N. T. (2001). Situational judgement tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

McDaniel, M. A., Psotka, J., Legree, P. J., Powell Yost, A., & Weekley, J. A. (2011). Toward an understanding of situational judgement item validity and group differences. *Journal of Applied Psychology, 96,* 327-336.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*, 515-525.

Moss, F. A. (1926). Do you know how to get along with people? Why some people get ahead in the world while others do not. *Scientific American, 135,* 26-27.

Morizot, J., Ainsworth, A. T., Reise, S. P. (2009). Toward modern psychometrics: Application of item response theory models in personality research. In R.W. Robins, R. C., Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York, NY: Guilford, 2007.

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

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

Mumford, M. D., & Owens, W. A. (1987) Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1–31.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50*, 823-854.

Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage Publications.

Patterson, F., Ashworth, V., Kerrin, M., & O'Neill, P. (2013). Situational judgement tests represent a measurement method and can be designed to minimise coaching effects. *Medical education*, *47*(2), 220-221.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153-172.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241-258.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Spring.

Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, *38*(7), 549-562.

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464-480.

Ross, M. (1987). The problem of construal in social inference and social psychology. In N/ E/ Grunberg, R. E. Nisbett. J. Rodin, & J. E. Singer (Eds.), *A Distinctive Approach to Psychological Research: The Influence of Stanley Schachter* (pp. 118-130). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ross, M., & Nisbett, R. E. (1991). *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw-Hill.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63-84.

Samejima, F. (1972). A general model for free response data. (Psychometric Monograph No. 18) Richmond, VA: Psychometric Society.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general ability and job knowledge. *Current Directions in Psychological Science, 2,* 7-8.

Schmidt, K. M., & Embretson, S. E. (2003). Item response theory and measuring abilities. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of Psychology*, Vol. 2. (pp. 429-445). Hoboken, NJ: John Wiley & Sons.

Schmitt, N., & Chan, D. (2006). Situational judgement tests: Method or construct? In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement and Application* (pp. 135-155). Mahwah, NJ: Lawrence Erlbaum Associates.

Shaklee, H., & Fichhoff, B. (1982). Strategies of information search in causal analysis. *Memory and Cognition, 10,* 520-530.

Sternberg, R. A., Wagner, R. K., & Okagaki, L. (1993) Practical intelligence: The nature and role of tacit knowledge in work and at school. In H. Reese & J. Puckett (Eds.), *Advances in Lifespan Development* (pp. 205-227). Hillside, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J., Forsythe, G. N., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, A. A., & Grigorenko, E. L. (2000). *Practical Interlligence in Everyday Life*. Cambridge, UK: Cambridge University Press.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3-46.

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika, 51*, 567-578.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52,* 1236-1247.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436-458.

Warne, R.T., McKyer, E.L.J., Smith, M.L. (2012). An introduction to item response theory for health behavior researchers. *American Journal of Health Behavior, 36*, 31-43.

Weekley. J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 50,* 679-700.

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

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgement testing. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement and Application* (pp. 1-10). Mahwah, NJ: Lawrence Erlbaum Associates.

Weekley, J. A., Ployhart, R. E., & Harold, C. (2004). Personality and situational judgement tests across applicant and incumbent contexts: An examination of validity, measurement, and subgroup differences. *Human Performance, 17,* 433-461.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgement tests: Issues in item development, scaling, and scoring. In J. A. Weekley, & R.

E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement and Application* (pp. 157-182). Mahwah, NJ: Lawrence Erlbaum Associates.

Weng, Q., Yang, H, Lievens, F, McDaniel, M. A. (2018). Optimizing the validity of situational judgement tests: The importance of scoring methods. *Journal of Vocational Behavior, 104,* 199-209.

Wright, N. (2013). New strategy, old question: Using multidimensional item response theory to examine the construct validity of situational judgement tests. (Doctoral dissertation, North Carolina State University, 2013).

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. *Statistical and methodological myths and urban legends*, 37-61.

Zu, J., & Kyllonen, P. C. (2018). Nominal response model is useful for scoring multiple-choice situational judgement tests. *Organizational Research Methods,* published online.

APPENDIX A. PARAMETER ESTIMATES FOR MULTIDIMENSIONAL NRM

***Table 9.* Parameters Estimates for the 3-Dimensional Nominal Response Model Divided by Training Type**

Item Discrimination

Item	*Combat Training*			*Semi-Combat Training*			*Non-Combat Training*		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
1	-0.17	-0.28	0.21	-0.04	-0.27	0.25	-0.08	-0.33	0.10
2	0.15	0.00	0.15	0.02	-0.03	-0.05	0.10	0.02	0.09
3	0.16	-0.11	-0.09	-0.02	-0.06	0.09	-0.14	-0.05	0.02
4	-0.07	-0.07	0.20	-0.08	-0.17	-0.02	0.04	-0.17	-0.10
5	0.04	-0.08	-0.05	0.29	0.03	0.13	-0.04	-0.02	0.02
6	-0.11	-0.04	0.04	0.04	-0.10	-0.02	0.12	-0.07	-0.01
7	-0.01	-0.10	0.03	-0.12	0.00	0.22	-0.25	-0.16	-0.04
8	-0.03	-0.05	0.05	0.01	-0.04	-0.04	0.04	-0.02	-0.06
9	0.05	-0.02	-0.05	-0.02	0.02	0.03	-0.06	-0.01	0.05
10	-0.11	0.02	-0.03	0.16	-0.03	0.03	-0.10	0.01	0.01
11	0.04	-0.15	-0.05	-0.07	0.01	0.08	-0.05	-0.11	0.05
12	0.14	-0.28	-0.24	-0.01	-0.05	-0.05	0.02	0.00	-0.03
13	-0.13	0.02	-0.06	0.13	-0.02	-0.02	-0.14	0.00	0.06
14	0.19	-0.12	0.11	-0.07	-0.03	-0.02	-0.10	-0.13	-0.21
15	-0.24	-0.03	0.00	0.08	-0.08	-0.07	0.19	-0.05	-0.03
16	0.03	0.07	0.03	0.04	-0.07	0.17	-0.01	-0.11	0.12
17	0.04	0.07	-0.03	0.00	-0.03	0.03	-0.02	-0.06	0.00
18	-0.23	0.00	-0.15	0.08	-0.05	0.02	-0.08	0.01	-0.05
19	0.07	0.01	0.00	-0.08	0.02	-0.02	-0.05	0.01	-0.02
20	0.64	-0.19	-0.41	0.02	-0.07	-0.08	0.12	-0.05	-0.11
21	0.50	-0.33	-0.07	-0.05	-0.04	0.31	-0.17	-0.13	0.07
22	-0.11	0.01	-0.02	0.13	-0.05	-0.01	0.20	-0.06	0.06
23	0.06	0.00	-0.01	0.22	-0.12	-0.14	0.11	0.00	-0.01
24	-0.12	-0.10	0.00	0.03	-0.09	0.00	0.11	-0.14	0.00
25	0.09	0.00	0.00	-0.08	0.00	0.00	-0.06	0.00	0.00

Scoring Coefficients

Item	*Combat Training*					*Semi-Combat Training*					*Non-Combat Training*				
	ak_0	ak_1	ak_2	ak_3	ak_4	ak_0	ak_1	ak_2	ak_3	ak_4	ak_0	ak_1	ak_2	ak_3	ak_4
1	0.00	2.07	1.52	2.83	4.00	0.00	2.37	1.33	5.90	4.00	0.00	1.94	0.50	1.77	4.00
2	0.00	3.49	5.48	-0.09	4.00	0.00	-4.54	-0.59	8.78	4.00	0.00	2.54	-3.35	7.62	4.00
3	0.00	4.75	6.43	6.19	4.00	0.00	1.07	5.05	4.73	4.00	0.00	3.88	6.74	6.36	4.00
4	0.00	5.77	3.53	2.66	4.00	0.00	6.44	-1.63	8.54	4.00	0.00	6.39	4.00	4.84	4.00
5	0.00	-4.30	0.08	-0.65	4.00	0.00	1.79	3.72	1.50	4.00	0.00	-3.74	0.31	1.88	4.00
6	0.00	14.58	5.14	2.33	4.00	0.00	9.45	1.59	4.69	4.00	0.00	12.01	5.20	5.02	4.00
7	0.00	2.09	-4.92	3.86	4.00	0.00	-2.35	0.88	5.63	4.00	0.00	2.24	1.79	4.43	4.00
8	0.00	-4.11	-3.11	-3.17	4.00	0.00	-7.68	-2.48	-1.16	4.00	0.00	-1.82	-0.36	8.51	4.00
9	0.00	-5.48	-8.04	-2.20	4.00	0.00	4.19	-13.38	-3.81	4.00	0.00	-6.27	-9.34	-4.12	4.00
10	0.00	0.30	10.39	-3.42	4.00	0.00	-0.44	0.88	-1.97	4.00	0.00	2.12	-5.08	6.73	4.00
11	0.00	0.36	-3.55	3.52	4.00	0.00	7.44	1.82	7.84	4.00	0.00	0.31	3.70	4.25	4.00
12	0.00	3.20	3.54	3.45	4.00	0.00	-10.92	-4.74	-1.27	4.00	0.00	-7.88	-1.57	-5.67	4.00
13	0.00	-1.99	-2.10	2.50	4.00	0.00	-1.83	-2.83	-0.76	4.00	0.00	8.69	8.01	4.89	4.00
14	0.00	1.75	4.22	5.37	4.00	0.00	-3.73	2.56	2.78	4.00	0.00	-3.75	-1.11	0.50	4.00
15	0.00	0.53	-1.71	-2.97	4.00	0.00	3.78	-0.54	-3.27	4.00	0.00	-1.92	-3.03	-4.30	4.00
16	0.00	1.45	11.10	-1.69	4.00	0.00	4.20	2.22	6.69	4.00	0.00	1.75	-0.42	6.23	4.00
17	0.00	-1.07	7.37	-2.18	4.00	0.00	-4.19	-7.21	8.63	4.00	0.00	-12.25	-2.35	8.05	4.00
18	0.00	-2.30	3.27	-1.12	4.00	0.00	-7.47	4.16	-5.09	4.00	0.00	10.28	-10.95	2.10	4.00
19	0.00	0.87	-1.79	-7.10	4.00	0.00	6.51	-1.73	-12.26	4.00	0.00	4.40	-2.90	-7.01	4.00
20	0.00	0.95	1.74	2.56	4.00	0.00	4.41	-1.02	-6.81	4.00	0.00	4.99	-2.17	-4.87	4.00
21	0.00	2.99	2.16	3.06	4.00	0.00	3.90	1.30	3.09	4.00	0.00	3.43	-0.39	2.74	4.00
22	0.00	-5.85	-4.94	-2.05	4.00	0.00	-8.91	-0.95	-2.57	4.00	0.00	-4.67	-2.81	-5.33	4.00
23	0.00	8.16	-2.28	-0.19	4.00	0.00	-2.15	-0.20	3.00	4.00	0.00	-5.04	-1.88	2.88	4.00
24	0.00	5.46	8.01	0.85	4.00	0.00	5.74	8.68	4.80	4.00	0.00	4.98	7.96	3.15	4.00
25	0.00	-8.37	0.08	6.18	4.00	0.00	-13.94	-9.07	7.65	4.00	0.00	-12.27	1.15	7.78	4.00

Endorsement Likelihood

Item	*Combat Training*					*Semi-Combat Training*					*Non-Combat Training*				
	d_0	d_1	d_2	d_3	d_4	d_0	d_1	d_2	d_3	d_4	d_0	d_1	d_2	d_3	d_4
1	0.00	-1.26	-3.01	-4.28	-4.08	0.00	-0.53	-2.46	-4.94	-2.66	0.00	-0.58	-2.24	-3.79	-2.46
2	0.00	3.24	0.63	0.89	0.74	0.00	3.62	0.89	1.17	1.12	0.00	3.38	0.09	0.61	0.91
3	0.00	1.74	3.02	3.29	0.83	0.00	2.68	3.73	3.09	1.52	0.00	3.37	4.30	3.67	2.08
4	0.00	-3.15	-6.02	-1.74	-3.85	0.00	-2.64	-6.91	-2.41	-4.08	0.00	-3.17	-6.72	-2.20	-3.72
5	0.00	1.41	-3.26	2.53	-3.40	0.00	1.29	-3.98	2.21	-2.79	0.00	0.87	-3.75	1.95	-3.61
6	0.00	-3.68	-2.52	-1.58	-1.91	0.00	-2.87	-3.07	-1.74	-2.11	0.00	-3.13	-3.04	-1.87	-1.92
7	0.00	-1.44	1.82	1.34	1.74	0.00	-1.56	1.21	1.78	1.92	0.00	-0.59	1.32	2.19	2.04
8	0.00	2.87	1.22	1.62	0.41	0.00	3.51	1.70	2.32	1.17	0.00	3.42	1.51	1.67	0.94
9	0.00	-3.77	-0.72	-3.97	-0.03	0.00	-4.55	-0.76	-4.35	-0.09	0.00	-4.25	-1.07	-4.53	0.03
10	0.00	1.98	-1.71	2.93	0.68	0.00	2.38	-0.78	3.61	0.27	0.00	1.69	-1.52	3.34	0.27
11	0.00	1.41	-2.97	1.27	-2.90	0.00	1.83	-3.52	1.70	-2.73	0.00	1.70	-2.65	1.64	-2.86
12	0.00	2.98	0.16	0.74	1.69	0.00	2.15	-0.44	-0.30	0.61	0.00	2.54	0.25	0.32	1.01
13	0.00	2.67	2.77	1.51	1.41	0.00	2.70	2.81	1.47	0.62	0.00	3.50	3.51	1.88	1.42
14	0.00	4.01	3.24	2.12	-0.63	0.00	3.57	2.82	1.70	-0.62	0.00	4.25	3.47	2.43	-1.85
15	0.00	-1.70	-0.11	2.31	-3.80	0.00	-2.60	0.22	2.44	-4.13	0.00	-1.86	0.46	2.63	-5.53
16	0.00	1.55	1.22	2.64	2.13	0.00	1.73	1.59	2.80	2.19	0.00	2.22	1.84	3.01	2.64
17	0.00	-1.36	0.71	-0.51	0.15	0.00	-1.09	0.80	-0.01	-0.09	0.00	-1.48	0.66	0.10	-0.16
18	0.00	-0.62	-0.84	0.13	-2.86	0.00	-0.88	-1.47	0.12	-3.38	0.00	-0.78	-1.72	0.28	-2.78
19	0.00	-0.32	2.04	0.42	-0.13	0.00	-0.34	2.09	0.34	0.74	0.00	-0.36	2.06	0.66	1.18
20	0.00	0.82	5.28	6.50	1.12	0.00	-2.57	3.39	4.75	-0.92	0.00	-2.02	4.13	5.37	-0.98
21	0.00	6.56	5.26	6.66	3.81	0.00	4.97	3.47	5.17	2.34	0.00	5.26	3.55	5.28	2.69
22	0.00	1.80	-1.04	-3.00	-1.62	0.00	2.12	-0.99	-3.71	-1.90	0.00	2.60	-0.13	-2.36	-2.22
23	0.00	0.65	-3.02	-5.46	-3.71	0.00	0.78	-2.51	-4.48	*NA*	0.00	1.10	-2.23	-5.60	-3.70
24	0.00	-1.12	-2.61	-5.18	-3.14	0.00	-1.13	-2.62	-5.08	-3.41	0.00	-1.19	-2.74	-4.08	-3.34
25	0.00	-1.08	-1.80	0.80	1.18	0.00	-1.95	-2.80	0.66	0.88	0.00	-2.18	-2.83	0.78	0.69