

SITUATIONAL JUDGMENT TESTS IN ASSESSING
SPECIFIC PERSONALITY CHARACTERISTICS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYDA ERİŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
THE DEPARTMENT OF PSYCHOLOGY

MAY 2013

Approval of the Graduate School of Social Sciences

Prof. Dr. Meliha Altunışık
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of
Master of Science

Prof. Dr. Tülin Gençöz
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully
adequate, in scope and quality, as a thesis for the degree of Master of
Science/Arts/Doctor of Philosophy.

Prof. Dr. H. Canan Sümer
Supervisor

Examining Committee Members

Prof. Dr. Canan Ergin (Özyeğin Üni, PSY)

Prof. Dr. H. Canan Sümer (METU, PSY)

Ass. Prof. Dr. Yonca Toker (METU, PSY)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Ayda Eriş

Signature :

ABSTRACT

SITUATIONAL JUDGMENT TESTS IN ASSESSING SPECIFIC PERSONALITY CHARACTERISTICS

Eriş, Ayda

MS., Department of Psychology

Supervisor: Prof. Dr. H. Canan Sümer

May 2013, 86 pages

Situational Judgment Tests are tools that are being utilized more and more for personnel selection purposes. Findings show that situational judgment tests have a number of advantages in personnel selection over some other tests, techniques, and methods. Among these advantages are considerable predictive validity, being less prone to biases observed in traditional self-report measures, and less adverse impact (McDaniel, Hartman, Whetzel, & Lee Grubb III, 2007; O'Connell, Hartman, McDaniel, Lee Grubb III, & Lawrence, 2007). The main purpose of the present study was to develop a situational judgment test that aimed to assess the Big Five personality factors. First, a situational judgment test tapping into the Big Five factors was developed for a large organization that functions in the manufacturing sector. Participants of the study were 304 white-collar employees of the organization. Reliability, construct validity and criterion validity of the developed test was examined as well as its relationship with performance outcomes. Results indicated that internal consistency of the developed measure was below the expected levels while test re-test reliability was satisfactory for some factors. Convergent and divergent validity, assessed through two different methods, were at acceptable levels. Finally the magnitude of the relationship between personality scores and performance outcomes was low to moderate. Results are discussed in addition to potential contributions and practical implications.

Keywords: Personnel selection, situational judgment tests, personality testing

ÖZ

KİŞİLİK FAKTÖRLERİNİ ÖLÇMEDE DURUMSAL MUHAKEME TESTLERİNİN KULLANILMASI

Eriş, Ayda

Yüksek Lisans, Psikoloji Bölümü

Tez Danışmanı: Prof. Dr. H. Canan Sümer

Mayıs 2013, 86 sayfa

Durumsal Muhakeme Testleri kullanımları giderek yaygınlaşan personel seçme araçlarıdır. Araştırmalar, bu testlerin personel seçmede hali hazırda kullanılan diğer test, teknik ve yöntemlere göre bazı avantajları olduğunu göstermektedir. Görece yüksek yordayıcı geçerlik, geleneksel öz beyan testlerinin maruz olduğu yanlılıklara çok açık olmaması ve ayrımcılık etkisinin düşük olması bu avantajlar arasında sıralanmaktadır (McDaniel, Hartman, Whetzel, & Lee Grubb III, 2007; O'Connell, Hartman, McDaniel, Lee Grubb III, & Lawrence 2007). Bu çalışmanın amacı beş temel kişilik faktörünü ölçmek üzere tasarlanmış bir durumsal muhakeme testi geliştirmektir.

Öncelikle, üretim sektöründe faaliyet gösteren büyük bir kurum için personel seçme amaçlı kullanılmak üzere beş temel kişilik faktörünü ölçecek bir durumsal muhakeme testi geliştirilmiştir. Çalışmanın katılımcıları bu kurumdan 304 beyaz yaka çalışandır. Geliştirilen testin güvenilirliği, yapı geçerliği, ölçüt bağımlı geçerliği ve performans çıktıları ile ilişkisi incelenmiştir. Sonuçlar, testin iç tutarlık katsayısının tatmin edici seviyenin altında olduğunu göstermiştir. Testin uyuşum ("convergent") ve uzaksak ("divergent") geçerliğinin ise kabul edilebilir seviyede olduğu gösterilmiştir. Son olarak, yordayıcı geçerliğe yönelik olarak, bu test ile ölçülen kişilik puanları ve performans çıktıları puanlarının düşük ile orta seviyede ilişkili oldukları bulunmuştur. Bu çalışmanın sonuçları, ilgili yazına olası katkıları ve pratik doğurguları tartışılmıştır.

Anahtar Kelimeler: Personel seçme, durumsal muhakeme testleri, kişilik testleri

To my father's daughter

ACKNOWLEDGMENTS

The author wishes to express her gratitude to her supervisor Prof. Dr. Canan Sümer for her guidance, advice, criticism, and insight throughout the research.

The author would also like to thank Assoc. Prof. Dr. Yonca Toker and Prof. Dr. Canan Ergin for their valuable suggestions and comments.

The test that lies in the center of the current thesis is part of a comprehensive research and development project conducted with the support of The Scientific and Technological Research Council of Turkey. The author would like to thank both academic and technical research teams that have effort in the project (TÜBİTAK TEYDEB Project No: 3110664)

Efforts of clinical psychology graduate students of METU were significant to finalize the current test. The author would like to thank Gözde Ikizer, Canan Büyükaşık-Çolak, Ferhat Yarar, Fatih Cemil Kavcıoğlu, Gaye Zeynep Çenesiz, and Öznur Öncül for their voluntary contribution.

The author also would like to emphasize the appreciation of her chance of working closely with Dr. Ayça Özen and support provided by Dr. Gizem Ateş during the study.

TABLE OF CONTENTS

PLAGIARISIM.....	iii
ABSTRACT.....	iv
ÖZ.....	v
DEDICATION.....	vi
ACKNOWLEDGEMENT.....	vii
LIST OF TABLES.....	xi
CHAPTER	
1.INTRODUCTION.....	1
1.1 Overview.....	1
1.2 Situational Judgment Tests.....	3
1.2.1 An Overview of the History of SJTs and Test Development Procedures Employed.....	3
1.2.2 Instruction Types.....	7
1.2.3 Format of SJTs.....	8
1.3 Reliability of SJTs.....	9
1.4 Validity of SJTs.....	10
1.4.1 Fairness and Adverse Impact.....	13
1.5 Moderators of Validity.....	13
1.6 What Do SJTs Measure?.....	15
1.7 Personality Assessment via SJT Methodology.....	19
1.7.1 Personality and Work Performance.....	19
1.7.2 Personality and SJTs.....	21
1.8 Current Study and Hypotheses.....	23

2. METHOD	26
2.1 Participants.....	26
2.2 Measures	26
2.2.1 Development of the SJT of Personality	26
2.2.2 NEO-Personality Inventory Revised (NEO-PI-R)	30
2.2.3 Big Five Inventory (BFI)	30
2.2.4 Personnel Multiple Reasoning Test (PMRT).....	31
2.2.5 Job Performance Measure	31
2.3 Procedure.....	31
3.RESULTS.....	32
3.1 Overview	32
3.2 Correlations among Variables, Reliabilities and Descriptive Statistics	32
3.3. Reliability: Internal Consistency and Test-Retest Estimates	36
3.4 Test Modifications	36
3.5 Construct Validity: Multitrait-Multimethod Matrix Approach.....	37
3.5.1 Convergent Validity.....	38
3.5.2 Divergent Validity	39
3.6 Divergent Validity Evidence in terms of Relationship with a Nonverbal Reasoning Test	41
3.7 Multitrait-Multimethod Matrix with Selected Items	41
3.8 Construct Validity: Confirmatory Factor Analysis Approach	44
3.9 CFAs: Testing for the Trait and Method Effects.....	48
3.10 Predictive Validity Analyses: Correlation of the SJT of Personality with Job Performance	50
4. DISCUSCION.....	52
4.1 Overview	53

4.2 Discussion of the Results Concerning Reliability and Validity	53
4.3 Strengths and Contributions of the Study	59
4.4 Practical Implications	61
4.5 Limitations of the Study and Suggestions for Future Research	61
REFERENCES	64
APPENDICES	70
APPENDIX A: Critical Incident Questionnaire	70
APPENDIX B: Example Item of SJT	75
APPENDIX C: Screenshot of computerized version of SJT	76
APPENDIX D: Items of Big Five Inventory (BFI)	77
APPENDIX E: Example Item of Personnel Multiple Reasoning Test (Screenshot).79	
APPENDIX F: Figure of Model 1: Null Model	80
APPENDIX G: Figure of Model 2: Trait Model.....	81
APPENDIX H: Figure of Model 3: Method Model	82
APPENDIX I: Figure of Model 4: General Trait Model	83
APPENDIX J: Figure of Model 5: Orthogonal Methods Model.....	84
APPENDIX K: Figure of Model 6: Correlated Methods Model.....	85
APPENDIX L: Tez Fotokopisi İzin Formu.....	856

LIST OF TABLES

TABLES

Table 3.1 Correlations among Study Variables, Reliabilities and Descriptive Statistics	46
Table 3.2 Test Re-test Reliability and Correlations among Factors of SJT	48
Table 3.3 Multitrait-Multimethod Martix for Personality Factors Assessed via SJT, BDI and NEO-PI-R.....	53
Table 3.4 Correlations among the SJT Factors and the Personnel Multiple Reasoning Test	54
Table 3.5 Multitrait-Multimethod MArtix with Selected Items	56
Table 3.6 Moldels and Charaacteristics of Models	57
Table 3.7 Correlatins among Manifest Variables	58
Table 3.8 Selected Fit Statistics and χ^2 Values for Models Tested	63
Table 3.9 χ^2 Difference Test Between Hiearchically Nested Models.....	64
Table 3.10 Correlations Between Performance Dimensions and SJT Factors	66

CHAPTER 1

INTRODUCTION

1.1 Overview

Situational Judgment Tests (SJT), as personnel selection tools, have attracted considerable research attention over the last 20 years. In a typical SJT item, a work related scenario, usually a problematic one, in which the required course of action is not obvious, is presented to participants. Participants choose their response from the response alternatives provided, all of which look reasonable. Although empirical evidence is much less established than the other constructs assessed in personnel selection (e.g., cognitive ability), studies about SJT show promising results in terms of reliability and validity. In addition, in terms of applicant reactions, it has potential to be rated high on job relatedness compared with more abstract types of selection tests although systematic research in this area is scarce (Bauer & Tuxillo, 2006).

Use of SJT items started in the 1920s with George Washington Intelligence Test, as a part of assessment of general intelligence. After the 1940s, such tests were used in the organizational area (Weekley & Ployhart, 2006). Many resources point the study by Motowidlo, Dunette, and Carter (1990) as the reintroduction of SJTs to personnel selection area. This study provides a clear guidance about development and scoring processes. The increasing research attention on SJTs has resulted in different test development procedures proposed by different researchers (e.g., McDaniel & Nguyen, 2001; Motowidlo, Dunette, & Carter, 1990). Research about advantages and disadvantages as well as implications of different development procedures is in its infancy and requires more systematic research (Weekley & Ployhart, 2006). As an example, starting the development procedure with collecting critical incidents is a commonly employed technique. When development process starts with critical incidents, the construct that is assessed by the test is defined a posteriori in most cases. Thus, although the test has high job-relatedness and relationship with the criteria, construct related validity remains ambiguous.

However, construct related validity is an important criterion in evaluating the psychometric quality of any selection tool in addition to its criterion related validity. It is believed that defining a construct to assess and then starting the test development process may result in enhanced construct clarity. In addition to the development technique, test format has implications for the end result; correlates of SJT scores may vary according to the type of instruction (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; McDaniel et al., 2001; McDaniel, Hartman, Whetzel, Nguyen, & Lee Grubb III, 2007). Type of instruction, the question that follows the scenario, has been mainly investigated in two categories. In the first category, participants are asked what they would do in response to the situation presented in the item. In the second category, participants are asked what they should do in the presented situation. The former is referred as behavioral tendency instructions and found to be related with personality correlated constructs whereas the latter was referred as knowledge instructions and found to be correlated with cognitive ability related constructs (McDaniel & Nguyen, 2001).

Two meta-analyses examined the ability of SJT in predicting job performance. These studies presented somewhat different results concerning the magnitude of the relationship between SJT scores and job performance (.34, .24). However, in both studies 90th credibility value was found positive (McDaniel, Nguyen, 2001; McDaniel, Hartman, Whetzel, & Lee Grubb, 2007), suggesting the generalizability of the findings.

Majority of the SJTs reported in the literature are criterion oriented rather than being construct oriented. That is, in most cases SJTs are not developed to predict specific constructs, but they are developed to have high relationship with the criterion of overall job performance. Notwithstanding the presence of this criterion-orientation, SJTs have been found to be correlated with various important attributes, like conscientiousness, agreeableness, and cognitive ability (O'Connell et al., 2007; Weekley & Jones, 1999).

Despite the availability of the evidence for criterion related validity, construct related validity of SJT still remains questionable. Thus, there is a concern to understand whether SJT's represent a method of measurement which can be used to assess different constructs or an indicator of an identifiable and meaningful new

construct itself as “situational judgment” or something else. Some researchers use SJTs to assess the construct tacit knowledge, which is suggested to be related to practical intelligence (Sternberg, Wagner, Williams, & Horvath, 1995). This approach assumes there is a construct and SJTs aim to capture this construct. Other researchers, who disagree with this approach, propose that SJT is a measurement method, with its limitations in terms of the constructs that could be assessed with it.

The present study, acknowledges SJT as a method to measure different constructs, and the purpose is to develop SJTs to tap into the construct domain of the Big Five personality factors. In Section 1.2 through Section 1.6 a brief review of the SJT literature, including an historical overview, reliability and predictive validity evidence, fairness issues, and approaches to construct validation, is presented. This review is followed by a section on personality assessment via SJT methodology. The final section of this chapter presents the hypotheses of the current study.

1.2 Situational Judgment Tests

In this section SJTs are examined in detail in terms of its history, test development, reliability and validity. In addition, construct related validity and arguments about the nature of constructs being assessed are discussed.

1.2.1 An Overview of the History of SJTs and Test Development Procedures Employed

Although different than contemporary SJTs used in personnel selection, history of assessing situational judgment dates back to 1920s. In earlier times, the motivation was to measure human judgment, thus, formats similar to SJT were used within intelligence scales like Binet Scale (1905). George Washington Intelligence Test (1926) employed a form of SJT that had most resemblance to today’s format (cited in Weekley & Ployhart, 2006).

Contemporary situational judgment tests are considered to have similarities with two widely used personnel selection tools; situational interview and work samples. In situational interviews, job-related situations are presented to applicants in an interview format, and the applicant is asked what he or she would do in response to the presented situation (Latham, Saari, Pursell, & Campion, 1980). The

differences between situational judgment test and situational interview are in their presentation, response, and scoring format. First, an SJT is presented in paper and-pencil format while a situational interview is presented in interview format. Second, in SJT response options are presented in multiple-choice-like format whereas in situational interview, interviewees generate their own response. Finally, concerning scoring, SJTs have an objective scoring key while in situational interviews despite presence of a general scoring key, interviewers' judgment can still play a role (Weekley & Ployhart, 2006). Motowidlo, Dunette, and Carter (1990), who can be considered to be the pioneers of the contemporary SJTs, state that their SJT development method is largely guided by the principles of situational interview described by Latham and colleagues (Latham, Saari, Pursell, & Campion, 1980).

Work sample tests, another valid method used in personnel selection, has also some commonalities with SJTs. In a typical work sample test, participants are presented with a miniature replica of the job and are asked to engage in job-tasks (Roth, Bobko, & McFarland, 2005). It is noted that, in terms of putting participant in a simulation-like situation, work sample tests and SJTs are similar. Nevertheless, the response formats of two methods are very different from each other (Weekley & Ployhart, 2006).

Motowidlo, Dunnette, and Carter's (1990) study, which describes SJTs as "low fidelity simulations" has been accepted as the point which simulated the popularity of situational judgment tests is refreshed as selection tools in the literature (Whetzel & McDaniel, 2009). Motowidlo and colleagues describe the simulation fidelity as a continuum. At one extreme there are high-fidelity simulations in which veridical representations of the task stimulus are presented and actual responses to perform a job are elicited. At the other extreme, written or spoken descriptions of the task stimulus are presented and written or spoken descriptions of responses are elicited. Although being more representative of the actual work settings, high fidelity simulations, like work sample tests, have two major disadvantages over low fidelity simulations. First, they are expensive to develop and implement and they require a certain level of experience from test takers. The situational interviews were described as low fidelity simulations by Motowidlo and colleagues.

The SJT development procedure described by Motowidlo, Dunnette, and Carter (1990) consists of a three-step process. In the first step, problematic situations or critical incidents, which will be turned into item stems, are collected. In the second step behavioral response alternatives are gathered from experienced employers. In the third step, response alternatives are evaluated in terms of their effectiveness and used to create the scoring key (see Motowidlo, Dunnette, & Carter, 1990; Motowidlo & Tippins, 1993). This three step approach to the development of contemporary situational judgment tests was adopted by or guided many researchers (e.g., McDaniel & Whetzel, 2007; Ployhart & Ehrhart, 2003). However, as expected, with the increasing amount of research in the area, different methods of item development, scoring, and instructions have also emerged (Weekley, Ployhart, & Holtz, 2006).

Item stems can be derived either by subject matter experts or by the researchers (Weekley, Ployhart, & Holtz, 2006). In the most common approach critical incidents that are collected from subject matter experts serve as the basis for item stem development (Motowidlo, Dunnette, & Carter, 1990; McDaniel & Nguyen, 2001). Test developers review the critical incidents for the purpose of developing item stems and sort them according to the constructs emphasized. At this stage, critical incidents need editing by the test developer considering several characteristics. For example, the test developer can select representing scenarios and omit recurrent ones. Length and format of critical incidents are kept similar. The items have to represent a wider range of jobs; therefore, if it is necessary, very specific jargon can be replaced with more generic ones. Furthermore, the situations should not include content that may raise legal concerns or inappropriate issues like workplace violence.

After collecting critical incidents, another group of subject matter experts (SMEs) provide alternative ways to behave in the situations that are described in the critical incidents. By this way, response options for the item stems are generated. Similar to the item stems, response options are edited by the researcher or the test developer. Finalized responses should be comprehensible and appropriate to use in a selection test. Furthermore, there should be a wide range of possibilities presented in the options. Very similar options should be omitted (McDaniel & Nguyen, 2001).

For each situation all response alternatives should represent more or less reasonable/plausible courses of actions; however, only one alternative should be the exact one.

Another approach to derive item stems and response alternatives is theory-based methods. According to Weekly, Ployhart, and Holtz (2006), it is possible to write SJT items based on job analysis, a model or a theory that guides the underlying competencies of effective performance. In the literature, there are examples of SJTs developed based on theory to assess personality (Trippe, 2002) and integrity (Becker, 2005). In the construct driven approach where the SJT is to be developed to assess an a priori construct, a hybrid approach can be used. For example, in a recent study aiming to develop an SJT of integrity, Meijer, Born, Zielst, and van der Molen (2010), worked with a group of experienced employees in the critical incident collection stage. In that study participants were specifically instructed to generate incidents about integrity only. In the development process of the SJT for three personality factors (*agreeableness*, *conscientiousness*, and *extraversion*) Motowidlo, Hooper, and Jackson (2006a) wrote stems to tap these factors, and then used SME's to generate response alternatives.

Both ways of developing item stems have their advantages and disadvantages. As Weekly and Ployhart (2006) stated further research is needed to establish the relative effectiveness of relying on experts or theory in item and response option development.

As the last step, after collecting item stems and response options, the researcher should determine the effectiveness of each response option (i.e., scoring of the test). There are different methods of scoring reported in the literature, such as, expert based scoring, empirical scoring, and theoretical scoring. The most common method to decide on the effectiveness of alternatives is to use SMEs (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). A large group of experts or a selected group may serve for this purpose. The test developer can prepare a questionnaire form with the items and response options and make a large group of experienced incumbents to rate each alternative. Mean, median, mode and standard deviation of response option ratings of effectiveness serve as input to decide which options is the best (Motowidlo & Tippins, 1993). The scoring key may also be developed with a group of subject matter experts and experienced workers with a

predetermined level of agreement or consensus (McDaniel & Nguyen, 2001; Phillips, 1992, 1993). Another way to use experts for scoring is comparing novices and experts. After experts and novices complete the questionnaire, the results are compared to decide on the final scoring (Bergman et al., 2006). At this step of test development, similar to the first two steps, theory based approaches are possible as well as empirical approaches. In theoretical scoring, response options are scored to reflect a theory, that is, best, worst and neutral options are prescribed by a theory. In empirical scoring; however, response options are scored in relation to their correspondence with a predetermined criterion measure. In their study to develop SJT for retail workers, Weekley and Jones (1999) adopted a method based on empirical scoring. In this method, firstly, job performance means are computed for the sample. Then, the best response for a given SJT item was decided by identifying the response option with the highest correlation with job performance. In theoretical and empirical method, it is vital to base the development procedure to a well established theory or a correctly chosen and measured criterion (Bergman et al., 2006).

Regardless of the development technique, it is argued that SJT items ideally possess several characteristics. McDaniel and Nguyen (2001) examined item stems and response options in terms of five characteristics. These characteristics are item fidelity, item format, item length, item complexity, and item comprehensibility. Item fidelity is related to convergence of the test format with the actual job setting. As far as the format is considered, items can be presented in written (paper-and-pencil or computerized) or in short video format. In terms of length, items can vary from very detailed to very short descriptions of a situation. Complexity of the items may also range from very simple to very complex which is a characteristic related to the previous one. In general the longer the item, the higher the probability it will be more complex. Comprehensibility is related to cognitive load of the item which is affected by length and complexity.

1.2.2 Instruction Types

As mentioned earlier, the growing body of SJT research has given rise to varieties in types of instruction and scoring as well. Unlike item and response option

development strategies, there are considerable amount of research about the consequences of different instruction types and scoring of SJT items.

There are several types of instructions that can be used in SJT items. Basically, after presenting the dilemma, an applicant can be asked to choose how he/she would behave in that situation, or the applicant can be asked to evaluate the effectiveness of response alternative(s). The former type of instruction is called “behavioral tendency instruction” and the latter is called “knowledge instruction” (McDaniel & Whetzel, 2007). In addition to instructions that require single answers, multiple answers or rating of each response option may be required by preserving the behavioral tendency and knowledge format. A participant can be asked to choose what he or she will most likely and least likely do, or to choose the most effective and least effective responses. Finally, participants can be instructed to rank all the answers from most likely he or she will perform to least likely he or she will perform (behavioral tendency instruction) or from most effective response to least effective response (knowledge instruction) (Ployhart & Ehrhart, 2003).

Studies in general show that behavioral tendency instructions show higher correlations with personality measures, while knowledge instructions show higher correlations with cognitive ability scores (McDaniel & Whetzel, 2005). The knowledge and behavioral tendency instructions distinction was suggested to be relevant to maximal performance versus typical performance; abilities give information about one’s maximal performance and personality and interests give information about one’s typical performance (McDaniel, Hartman, Whetzel, & Lee Grubb III, 2007). Researchers who want to assess cognitive ability and related constructs are advised to use knowledge type of instructions whereas researchers who want to assess personality and related constructs prefer behavioral tendency type of instructions (Whetzel & McDaniel, 2009). Further implications of different response instruction on validity are discussed below. In the following section, different SJT format options are presented briefly.

1.2.3 Format of SJTs

The most commonly used format of presenting the test to the participants is written format, as a paper-pencil-test. A relatively new technique in the literature as

an alternative to classical written format is video-based format. In video-based SJT, as the name implies, the questions and response alternatives are presented visually with support of narration. It is suggested that visual presentation of a situation enriches the details thus increases fidelity. Considering that the SJT technique is indeed a simulation technique (Motowidlo, 1990), presenting a work related scenario via video is expected to increase fidelity. Relying on watching and listening, the format partly eases the cognitive burden of reading, which is suggested to be suitable for jobs where reading ability is not critical. On the other hand, an obvious disadvantage of video based tests is the increased cost of development. In addition, it is discussed that a video input may add irrelevant information to the item that may lead to error (Weekley & Jones, 1997). Also, although a paper-and-pencil format permits slow processing, in video-based formats if an applicant does not have the chance to replay the video, he/she should be very attentive to catch all the visual and auditory information presented (Kanning, Grewe, Hollenberg, & Hadouch, 2006).

1.3 Reliability of SJTs

Internal consistency reliability, test retest reliability, and parallel form reliability estimates are among the reliability estimates frequently used and reported in the selection research (Cook, 2009). The first estimate, although the most readily available estimate; is not seen as the most appropriate one for SJTs. Nevertheless, due to data collection and test development problems the latter two forms of estimates are not reported in the literature frequently (McDaniel & Whetzel, 2007).

SJT items are heterogeneous in nature. That is, they have multiple correlations with multiple constructs. For example, a sample item's correlational analysis presented in the review published by McDaniel and Whetzel (2005) show that a response option can be significantly and positively correlated with general mental ability, while being negatively correlated with agreeableness. Thus, the items do not load on a single factor when examined with factor analysis. Such a case makes Cronbach's alpha reliability index inappropriate for SJT's (Whetzel & McDaniel, 2009). The important issue here is whether the test is developed to tap a predetermined construct and if so what the nature of this construct is. In a recent

study by Meijer et al. (2010) the reported Cronbach's alpha level was .69 for a 14-item test which had been developed to tap integrity. In another study alpha was .72 for a 10-item test that assessed working effectively with others (O'Connell, 2007). Both of these reported internal consistency estimates are at barely acceptable levels. It is also important to note that when the test is developed to measure predetermined constructs, there must be enough items for each construct as alpha is a partial function of the number of items in the test. For example, Oswald et al. (2004) reported fairly low alpha levels, ranging between .22 and .55 for 12 dimensions of student performance with 3 to 6 items for each. Another 40-item SJT developed to assess the single factor named practical situational demands of managers (defined as resolving interpersonal conflict, multitasking, and handling emergencies) yielded an internal consistency reliability of .73, which was suggested to be below an acceptable level for a 40 item scale (Chan & Schmitt, 2002).

Test retest reliability is seen as a more appropriate estimate of reliability for the reasons mentioned above. Nevertheless, this type of estimate is not frequently reported in the literature (Whetzel & McDaniel, 2009).

Parallel form reliability is also an appropriate estimate of reliability for SJT. However, to be able to assess parallel form reliability, two versions of the same SJT, meeting all the stringent statistical requirements of parallel forms, need to be developed which does not seem practical. In the literature the study conducted by Clause and colleagues (1998) aimed to produce parallel forms of an SJT that assessed handling conflict, interaction with peers, and authority figures. The correlations among original and three alternative forms ranged between .72 and .77. Although there was an effort to develop parallel items and to follow precisely the same procedures in developing the parallel forms, it is noted in this study that even minor changes in the wording of an item could lead to responses different from the ones produced by its parallel item (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998).

1.4 Validity of SJTs

The growing popularity of SJTs in the literature is in fact related to their relationship with criteria (McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007), higher levels of user acceptance (Bauer & Tuxillo, 2006), and lower

levels of adverse impact (Weekley & Jones, 1999; O'Connell et al., 2007). There are quite a few of individual (Clevenger, Pereira, Wiechmann, Schmitt & Schmidt Harvey, 2001; Oswald, Schmitt, Kim, Ramsay & Gillespie, 2004; Becker, 2005) and meta-analytic studies examining the face validity, criterion related validity, incremental validity, and construct validity of SJTs (e.g., McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007; Weekley & Jones, 1999).

Studies about SJTs show that they have acceptable level of relationship with various performance criteria. They predict student performance (Oswald et. al, 2004), supervisory ratings of job performance (Clevenger et al., 2001; McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Lee Grubb III, 2007; Weekley & Jones, 1999), contextual performance (O'Connell et al., 2007), career potential, leadership, and in role performance (Becker, 2005), negotiation skills (Phillips, 1993), sales skills (Phillips, 1992), and teamwork skills (Stevens & Champion, 1999).

According to the results of a meta analysis that comprised 39 studies about the validity of SJTs conducted until the year 2000, the estimated population criterion validity of situational judgment tests for predicting job performance was .34 (after corrected for measurement error in criteria), and 90th percentile credibility value was positive (McDaniel et al., 2001). A more recent meta-analysis that included more recent literature presents a slightly lower criterion related validity estimate of .26, but still having a positive 90th percentile credibility value (McDaniel, Hartman, Whetzel, & Lee Grubb III, 2007). These results support the generalized validity of situational judgment tests in predicting job performance.

In general, empirical evidence suggests that SJTs are useful in predicting job related outcomes. Chan and Schmitt (2002) discuss the reasons for the predictive power of the SJTs. They argue that since the test items are directly developed or sampled from criterion behaviors, the nature of the test brings prediction. Since the content of the test is developed, in most cases, by/with subject matter experts, or is based on critical incidents, or is derived from job analytic data, the test generated in the end is likely to closely match the criterion. This is a methodological advantage that makes SJT superior to many traditional, on-the-shelf tests. According to these researchers, another possible reason for the success of SJT is related to its heterogeneous nature. Because job performance itself is a heterogeneous construct

in nature, a method that is related to a wide range of constructs can be expected to have decent predictive ability. Though practically valuable, the risk involved in this suggestion is again explained by the same researchers. That is, according to researchers, multidimensional measures produce less clearly interpretable results, which in turn lower the ability to fully understand the constructs that are measured whether as predictor or criterion. As mentioned in an earlier section, this multidimensional nature is also responsible for the unsatisfactory internal consistency estimates reported for SJTs.

In terms of incremental validity, a study conducted to see the predictive power of a cognitive ability test, a personality test (conscientiousness, agreeableness, attention to detail, locus of control, and positive affectivity), and an SJT assessing working effectively with others with 10 items showed that the SJT explained additional variance over the cognitive ability test (3 %), the personality test (4 %) and over a composite of cognitive ability and personality test (1 %) in predicting task performance. It can be argued that although the incremental validity over the composite of both tests is not impressive, the incremental validity of SJT over cognitive ability and personality test is worth paying attention when task performance considered. In terms of contextual performance, on the other hand, SJT had incremental validity over cognitive ability (4 %) test but no incremental value over personality test (O'Connell et al., 2007). In another study, SJT was found to add incremental variance over cognitive ability, conscientiousness, job experience, and job knowledge (2 %) in a sample of federal investigators. The incremental validity emerged as a trend but failed to reach significance in two other samples in the same study (Clevenger et al., 2001). In another study Chan and Schmitt (2002) compared the predictive abilities of cognitive ability, all five factors of Big Five and SJT. They found that SJT had incremental validity over the other two measures in predicting core technical proficiency (5 %), job dedication (8 %), interpersonal facilitation (3 %), and overall job performance (4 %). A recent study examined the relationship between job knowledge and SJT, in predicting performance dimensions of medical trainees. It was found that SJT had incremental validity over knowledge test, explaining an additional 5.9% variance over knowledge test (Lievens & Patterson, 2011).

1.4.1 Fairness and Adverse Impact

The available evidence indicates that SJTs result in less adverse impact than many other selection tests/techniques, although group differences exist (O'Connell et al., 2007; Weekley & Jones, 1999). In a recent study, the reported differences in terms of race and gender were .38 and -.27, favoring whites and females, respectively (O'Connell et al., 2007). Examining a wider range of racial groups in their meta-analysis, Whetzel, McDaniel, and Nguyen (2008) found the same mean difference for Black and White participants (.38), and slightly lower differences for Hispanic and White (.24) and Asian and White (.29) samples favoring white participants. Women, again, had higher mean scores than men (-.11). First, it is important to note that all these mentioned differences were small differences. In addition, these differences were also found to be moderated by cognitive ability, personality, and response instructions factors. That is, as the correlation between SJT and cognitive ability increased, the mean group differences also increased in favor of white participants, and as the correlation between SJT and conscientiousness and agreeableness increased, the mean group differences for males and females increased in favor of female participants. In addition, situational judgment tests with knowledge instructions had higher mean differences, favoring men than behavioral tendency instructions, an expected result considering previous findings that show higher correlation between knowledge instructions and cognitive ability (Whetzel, McDaniel, & Nguyen, 2008). Regarding the results provided it can be concluded that SJTs have the advantage of lower levels of discrimination although the level of mean differences may vary according to the test development methods.

1.5 Moderators of Validity

Studies show that the validity of SJTs are moderated by the procedures of test development process, such as, scoring method (Bergman et al, 2006) and characteristics of the test itself, such as the level of detail in questions (McDaniel, 2001) and response instructions (McDaniel, Hartman, Whetzel, & Grubb 2007).

Characteristics of the test development process may have an effect on results. Job analysis, a vital factor for job relatedness, was investigated as a moderator of the

relationship between situational judgment tests and job performance. Not surprisingly, tests that were based on a job analysis were found to have higher validity (.38) than tests that were not based on job analysis (.29). These findings highlight the importance of job analysis in the test development process (McDaniel et al., 2001). Throughout the development phase, the scoring method employed to decide the effectiveness of each response option has been shown to have an effect on validity. In a comprehensive study, 11 scoring methods (including versions of empirical, theoretical, expert based and hybrid methods) applied to the same SJT items were examined in terms of relative effectiveness. Empirical and subject matter expert based scoring methods were found to predict the criterion and had incremental validity over cognitive ability and personality measures. In addition, these two scoring methods did not result in gender differences (Bergman et al., 2006).

The characteristics of the test itself may also affect the results. Format of the test is one of the important features. Test presentation format seems to affect the degree of fidelity, which is critical for a simulation. For SJTs, video based formats are considered as higher in fidelity than paper-and-pencil formats (McDaniel, Whetzel, Hartman, Nguyen, & Lee Grubb III, 2006). Although not used widespread, studies conducted to compare format of presentation of test shows that SJT items presented via video were perceived as more realistic and useful by the participants (Kanning et al., 2006) than were written SJT items.

Response instructions have also been investigated as a moderator of the relationship between test scores and job performance. As mentioned above, response instructions are examined in two broad groups; knowledge and behavioral tendency. McDaniel, Hartman, Whetzel, and Lee Grubb III's (2007) meta-analysis reported no moderating effect of response instruction. On the other hand, a study conducted by Ployhart and Ehrhart (2003) that explored the differences in terms of reliability and validity with different instructions for the same questions via a within subject design yielded different results. That is, the test with knowledge type of instruction resulted in a non-normal distribution with a higher mean and lower standard deviation than the test with behavioral tendency type instruction. In terms of reliability, the test with behavioral tendency instruction had higher test-retest reliability than the test

with knowledge instruction. In addition, relationship with the criterion was observed only with behavioral tendency type of instructions.

1.6 What Do SJTs Measure?

Although it was demonstrated that it has many advantages such as desirable levels of predictive ability and low discrimination against gender and race, it is important to understand what an SJT measures. Understanding the constructs that are assessed would help to relate them to relevant theoretical models and enhance understanding of the criterion. As suggested by Chan and Schmitt (2005), investigation of construct related validity of SJT is crucial. Hence, in this part, different approaches and viewpoints concerning the construct validity of SJTs are summarized.

In the SJT literature, there is a great effort to understand what an SJT measures. There are theoretical as well as empirical explanations provided to this question. The method versus construct discussion goes hand in hand with this discussion. That is, there is an effort to understand whether situational judgment is a construct by itself or a situational judgment test is a method that can be employed in the assessment of different constructs. Theoretical debates about what an SJT measures implicitly assume that there is a construct which can be called situational judgment or a form of intelligence. Opponents of this idea argue that in fact an SJT can be designed to assess different constructs but there are inevitable constructs that every SJT measures (McDaniel & Nguyen, 2001; Chan & Schmitt, 2005). In addition, relatively new in the literature, there are efforts to develop specific construct based SJTs.

A theoretical approach to the construct of SJT was suggested by Sternberg and colleagues. Sternberg, Wagner, Williams, and Horvath (1995) distinguish between academic and practical kinds of intelligence. The term tacit knowledge, as described by the researchers, is knowledge that practically intelligent individuals acquire and use. It has three features. First, it is procedural in nature; that is, related to “knowing how.” Second, it is relevant to the attainment of valuable goals and is practically useful. Finally, it is acquired with little help of others. The measure of this construct consisted of a set of work related situations with 5 to 20 response

options for each. Participants were asked to indicate how they would solve the problem. Although not labeled as a “situational judgment test”, this method is very similar to situational judgment tests. A recent review that examines common themes in the definitions provided for situational judgment, argues that there is much overlap between the definition of situational judgment and common sense. Thus, it is argued that situational judgment tests provide the most viable method for assessment of the attribute “common sense” in employment settings (Salter & Highhouse, 2009). Hence, Sternberg and colleagues and Salter and Highhouse assume that situational judgment is a construct and not a method of measurement.

Researchers who oppose the above argument, on the other hand, argue that this item type that Sternberg and colleagues (1995) used is a situational judgment test that is widely used in the personnel selection field and there is no empirical evidence that they do measure a single general factor, tacit knowledge or practical intelligence. Rather it is a method of measurement that may assess several constructs (McDaniel & Whetzel, 2005). According to Chan and Schmitt (2005), SJT is not a construct itself. A situational judgment test can be constructed as a method of testing just like an interview. However, this method still has some constraints on the range of constructs that can be assessed. SJTs have dominant constructs which are associated with the core characteristics of the content of a typical test. These core characteristics are; practical situational demands (i.e., realistic demands found in practical or everyday situations are described), multidimensionality of the situational response (i.e., good judgment is a function of multiple trait and abilities), and criterion-correspondent sampling of situations and response options (i.e., test developers adopt a domain sampling approach). It is suggested that consistent with these core characteristics, three constructs are primarily dominant in SJTs; which are adaptability, contextual knowledge, and practical intelligence. These constructs are almost inherently assessed in every SJT (Chan & Schmitt, 2005; Schmitt & Chan, 2006).

Studies that present correlations with well established constructs like cognitive ability and personality show empirical evidence for the constructs assessed with situational judgment tests. Job knowledge and work experience were also investigated as a factor. McDaniel and Nguyen (2001) argued that SJTs were

measurement methods that can be constructed to assess different constructs. Nevertheless, consistent with Chan and Schmitt's argument (Chan & Schmitt, 2005; Schmitt & Chan, 2006), these authors also stated that there are limits for the constructs that can be assessed with SJT. Also, they asserted that there are constructs that are measured in any SJT. These constructs are cognitive ability, since any measure of judgment is expected to have correlation with cognitive ability, and job experience, especially in inexperienced samples.

Many studies examined the correlations between SJT scores and cognitive ability and personality variables. For example, in McDaniel and colleagues' (2001) meta-analysis, the correlation between SJT and cognitive ability was reported to be .46. However, there were also studies in this meta-analysis reporting no correlation between cognitive ability and SJT scores. In a recent study by Chan and Schmitt (2005), an SJT was developed to assess practical situation demands like resolving interpersonal conflicts, handling emergencies and so on. While it was correlated with Conscientiousness, Extraversion, Agreeableness, Neuroticism, and Openness to Experience (.23, .24, .29, -.20, .19, respectively), this measure was not correlated with cognitive ability scores ($r = -.02$, ns). Clevenger and colleagues (2001) reported results of three different samples and three different SJTs in relation with cognitive ability and personality scores. Three samples differed in terms of the correlations between SJT and cognitive ability (.01, .17, and .53) and the correlations between SJT and conscientiousness (.16, .21, and .00). In a relatively recent study O'Connell and colleagues (2007) conducted regression analysis to predict SJT scores. It was concluded that SJT scores were a function of cognitive ability, conscientiousness, agreeableness, and positive affect ($R = .49$).

Another factor that is suggested to have correlation with SJT scores is job experience. Similar to cognitive ability and personality, there are studies reporting significant correlations and studies reporting no correlations between SJT scores and job knowledge and work experience. Meta-analytical results show a small positive correlation (.05) between SJT and experience (McDaniel & Nguyen, 2001). A study examining the relationship between SJT scores and demographic factors reported that tenure was not related with SJT scores, but also not related with performance criteria either (Motowidlo & Tippins, 1993). Another study with two samples found

correlations of .16 and .26 between experience and SJT scores, which is higher than the correlation between performance and experience (Weekley & Jones, 1999). Clevenger et al. (2001) reported correlations between job knowledge and SJT scores as .13, .19 and -.13 for three different samples.

Based on the reviewed literature, it is important to note that construct related validity studies, especially the results of meta-analyses on the validity of SJTs, should be interpreted with caution since they treat all SJTs as the same, regardless of the possibility that different SJTs may tap different constructs (Whetzel & McDaniel, 2009). In such a case it is understandable that individual studies to find different results in terms of relationship between SJT scores and the other constructs.

McDaniel and Nguyen (2001) argued that SJT was a measurement method that can be constructed to assess different constructs; for example, in the form of an interpersonal factor measure that present items related to interpersonal situations. Consistently, more recently there are efforts in the literature that define a construct in the first place and use SJT as an alternative method of assessment for that constructs. This approach provides a clear answer for the discussion about what an SJT measures. For example, Becker (2005) developed an SJT of integrity. This test was found to be correlated with career potential ($r = .26$), leadership ($r = .18$), in-role performance ($r = .24$) and overall job performance ($r = .22$), which are suggested as integrity relevant outcomes. This finding can be interpreted as criterion-related validity evidence of specific construct oriented SJTs. Meijer and colleagues (2010), also developed an SJT of integrity, and construct related validity of the test was reported both as convergent and discriminant validity. The correlation between the SJT that assess integrity and the two other integrity tests were significant ($r = .23$ for Honesty-Humility Test, and $r = -.36$ for How I Think questionnaire) and the correlation between the integrity SJT and cognitive ability test was not significant ($r = .13$).

Trippe (2002) developed an SJT of personality including three factors of Big Five personality: *Conscientiousness*, *Agreeableness* and *Openness to Experience*. The results of this study did not show a clear convergent and divergent validity

pattern; however, it is concluded that this effort shows that it is possible to develop construct oriented SJTs with proper developmental rigor.

The current study treated SJT as a method of measurement. That is, it was believed that SJTs could be used to assess a priori determined, specific, job-related constructs. More specifically, the present study aimed 1) to develop SJTs tapping into the domain of Big Five personality factors and 2) to establish reliability and validity of the developed SJTs. Both criterion-related validity and construct validity were examined for a relatively large white-collar sample. The construct validity was evaluated by examining the pattern of correlations between a conventional measure of the Big Five and the SJTs developed to assess the same personality dimensions/factors. Criterion-related validity was evaluated by examining the relationships between the developed SJTs and supervisory ratings of job performance. The reasons for choosing the Big Five personality dimensions as the assessment target of SJTs is described in the following section.

1.7 Personality Assessment via SJT Methodology

In this section, first, the literature about personality work performance relationship is summarized. Then, assessment of personality via SJT method was discussed in the light of available literature.

1.7.1 Personality and Work Performance

Personality assessment is a widely used technique in the recruitment and selection processes in the USA and Turkey for both managerial and non managerial positions (Piotrowski & Armstrong, 2006; Sözer, 2004). Predictive ability of such tests has been the focus of many studies in the literature for long time. There are critical meta-analytic studies which show the relationship between personality and job performance variables (e.g., Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001; Salgado, 1997). These studies investigate the relationship of the Big Five dimensions separately for different occupational groups (managers, sales, professionals, etc.) and criterion type (i.e. subjective and objective criteria).

Results reported by Barrick and Mount (1991) showed that for professionals (engineers, architects, attorneys, accountants, teachers, doctors, and ministers) the

relationships of the Big Five dimensions with job proficiency (i.e. performance ratings) reported as corrected coefficients .05 for *Extraversion*, .04 for *Emotional Stability*, .02 for *Agreeableness*, .20 for *Conscientiousness*, and -.08 for *Openness to Experience*. According to the results of this comprehensive study, *Conscientiousness* was a valid predictor of job performance across occupational groups and across various criteria ($p = .20 - .23$). *Extraversion* was a predictor of performance dimensions of managers and sales workers, where interaction with other people constitutes an important part of the job ($p = .15 - .18$). *Emotional stability* had lower correlations with performance ($p = 0 - .12$). *Openness to Experience* was a significant predictor of training performance ($p = .25$).

Another meta-analysis by Salgado (1997) reported similar but higher correlations. Results of this meta-analysis showed that, in addition to conscientiousness, emotional stability was a valid predictor of three different criteria (i.e., rating, training, and personnel data) across occupations. Emotional stability had relationship with job performance varying from .12 to .27 and conscientiousness had relationship with job performance varying from .11 to .39. In this study, for combined occupational groups, reported correlations of supervisor-rated job performance was .18 for *Emotional Stability*, .12 for *Extraversion*, .02 for *Openness to Experience*, -.02 for *Agreeableness*, and .26 for *Conscientiousness*.

In another study, Barrick, Mount, and Judge (2001) analyzed the results of 15 meta-analyses conducted to test the relationship of personality and performance variables. According to the results of this mega meta-analysis, corrected correlations with supervisory ratings were reported to be .13 for *Extraversion*, .07 for *Emotional Stability*, .13 for *Agreeableness*, .31 for *Conscientiousness*, and .07 for *Openness to Experience*.

In the light of the reviewed literature, it seems fair to conclude that *Conscientiousness* is a valid predictor of job performance. *Emotional Stability* and *Agreeableness* have relatively lower relationship with performance criteria. *Extraversion* seems to be a significant predictor for occupations that involve social interaction. *Openness to Experience* is valid predictor of training performance.

Hence, in general, it can be stated that, despite the variations in the predictive abilities of different personality dimensions, personality, in general, seems to be a

relevant construct in the prediction of job performance. In the current study specific personality factors as assessed by the SJT method is believed to show not only acceptable levels of both reliability and validity but also relatively better predictive ability as compared to the traditional assessment of personality.

1.7.2 Personality and SJTs

Meta analytic findings in the literature have shown that SJT scores have relationship with personality variables (e.g., Chan & Schmitt, 2005; Clevenger et al., 2001). This finding implies that regardless of the specific construct targeted in the development of an SJT, overall SJT scores are likely to have relationship with personality related constructs. Yet, it is important to clarify that the observed significant correlations between SJTs and personality factor in general do not necessarily mean that any SJT can well be a personality assessment tool. However, it is also important to point that an SJT, indeed can be developed to assess personality. Also, from an implicit trait policy theoretical framework, SJT methodology can be a very fruitful avenue for personality assessment with a number of practical advantages.

According to Motowidlo, Hooper, and Jackson (2006b), the concept of *implicit trait policy* (ITP), is a term used to describe “implicit beliefs about the causal effect of traits expressed by various actions on the effectiveness of those actions” (p.63). They state that individuals have implicit beliefs about the importance of personality traits that they use in determining behavioral effectiveness. These implicit beliefs are stable differences between individuals and are causally affected by personality traits. For example, when judging the effectiveness of an action, some people rely on ITPs that are about the level of agreeableness more than other people (since personality traits have causal effects on ITPs, these individuals are agreeable individuals). When such a person is asked to judge the effectiveness of two SJT response options, one reflecting high level of agreeableness and the other reflecting low levels of agreeableness, this person is likely to judge the high level agreeableness option as an effective one. Individuals that place less importance on agreeableness in judging the effectiveness of an action (i.e., disagreeable individuals), on the other hand, are likely to rate the agreeable

response option, probably, as only slightly more effective than the disagreeable option.

The other source of variance together with personality traits in ITPs is experience/learning. Individuals learn costs and benefits of expressing different kinds of traits. This learning may be either a general learning, learning the general principles, or it can be more direct learning in the form of procedural learning based on experience. Theory posits that as individuals get exposed to various work related situations, they develop an understanding of consequences of their actions in more specific situations. It is suggested that in SJTs, the scoring key is determined by/with SMEs measures procedural knowledge. On the other hand, SJTs that are scored with expert judgments about the level of trait that a response option presents measure ITPs. There are empirical studies conducted to test the ITP hypothesis using SJT. In an example study, the SJT was designed to tap extraversion, agreeableness and conscientiousness. For the purpose of ITP assessment, response options of the questions were scored according to the level of expression of the intended trait by independent researchers. In the study design, participants filled SJT by rating the effectiveness of each option in addition to a well known personality test; NEO FFI. ITP for each trait is calculated by correlating participants' effectiveness rating and the option's intended level. ITPs of Agreeableness and Extraversion were significantly related with corresponding NEO FFI scores on these traits but not for Conscientiousness. Thus, it is suggested that SJT format carry information about individuals standing on the traits that are intended to be measured (Motowidlo, Hooper, & Jackson, 2006a). Researchers concluded that SJTs were valid predictors of job performance, and they might have the advantage of revealing information about personality traits in the form of implicit measures.

In addition to the above presented theoretical justification for the use of SJT in assessing specific personality measures, SJT based personality assessment seems to present an additional advantage compared to the conventional assessment of personality. Response distortion or fakability has been an issue of concern in relation to personality assessment using conventional techniques (Ones, Viswesvaran, & Reiss, 1996; Ones & Viswesvaran, 1998). Response distortion, "situation-specific intentional distortion of responses," is also referred by the terms

such as (lack of) frankness, social desirability, exaggerating personal strengths, self enhancement, and faking (Sackett, 2011, p. 380). In the literature there are different methods to estimate faking in personality scales. In the most widely used methods, participants are instructed to present themselves as good or bad (i.e. they are instructed to fake good or bad). In addition, social desirability scales are used in combination with personality scales. A comprehensive meta-analysis conducted to examine the effects of faking on personality test scores showed that, first of all, all factors of the Big Five were equally susceptible to faking. Participants, when instructed to present themselves as good, were able to manipulate their scores up to half of a standard deviation. Finally, social desirability scales were found as the most faked scales (Viswesvaran & Ones, 1999).

The literature of SJTs, on the other hand, does not contain enough studies to conduct a meta-analysis on response distortion. However, SJTs are expected to have some advantages over traditional methods of personality assessment. That is, as the items in a typical SJT are not too transparent concerning what is being measured, they are likely to be less subject to faking or social desirability effects. Furthermore, SJTs which have a forced choice response format are likely to have an advantage over the traditional personality assessment methods that employ Likert type response formats. Using forced choice response format was suggested as a way of preventing potential response distortion (Cook, 2009; Paulhus & Vazire, 2007).

1.8 Current Study and Hypotheses

The results of a recent meta-analysis show that most of the studies use SJTs as a measure of a specific construct (Christian, Edwards, & Bradley, 2010). However, there are also considerable amount of studies that employ SJT and fail to identify the construct(s) measured by the test. According to Christian, Edwards, and Bradley among, 161 manuscripts and articles published between 2005 and 2008, 66% of the studies used SJTs to measure leadership, interpersonal skills, basic personality tendencies, and teamwork skills while 33 % did not report the constructs measured or did not provide enough information to determine the constructs measured.

This study approached SJT as a method to use in assessing specific constructs (i.e., the Big Five dimensions). In other words, in the present study, the SJT method was used to assess specific personality constructs. As stated above, the SJT methodology was expected to provide a useful alternative to conventional personality assessment.

The main goal of the study is to develop SJTs aiming to assess the Big Five personality dimensions. Since the SJT is developed to assess a predetermined construct, reliability estimates of the test in terms of inter item reliability and test retest reliability, are expected to be at acceptable levels.

Participants are also administered a more conventional and widely used personality measure (i.e., NEO-PI-R, McCrea & Costa, 1992). Convergent and divergent validity of the SJTs are examined in terms of their relationship with the NEO-PI-R dimensions. More specifically, the NEO-PI-R dimensions are used in establishing convergent and divergent validity of the developed SJTs.

Criterion related validity of the SJT is examined via its relationship with supervisory ratings of job performance. It is expected that personality factors would predict job performance. However, as summarized above, some factors are expected to have higher relationship (i.e., *Conscientiousness* and *Emotional Stability*) than others. Furthermore, a test of nonverbal reasoning, as a measure of general cognitive ability, is also used on an exploratory basis for validation purposes. The relationship between cognitive ability scores and SJT scores are expected to be low and insignificant. This expectation may sound counterintuitive in the light of the studies reporting positive correlations between SJT and cognitive ability. However, in the present study the SJT is developed to assess specific personality dimensions and hence were less likely to tap into general cognitive ability or practical intelligence. In the light of the reviewed literature, the following hypotheses are tested:

Hypothesis 1: Reliability estimates of SJTs measuring the Big Five dimensions of personality have acceptable reliability estimate in terms of internal consistency reliability and test re-test reliability.

Hypothesis 2: The same personality factors assessed with NEO-PI-R and with the SJT methodology have higher correlations than the correlations between different personality factors assessed with NEO-PI-R and with the SJT methodology.

Hypothesis 3: The pattern of the relationship between personality factors assessed with SJTs and job performance are parallel to those reported between the conventional assessment of the Big Five dimensions and job performance.

In addition to the above hypotheses, the correlations between the SJT scores and a cognitive ability test score are examined to provide supportive evidence for divergent validity.

CHAPTER 2

METHOD

2.1 Participants

Participants of the current study were 304 white collar employees from a company operating mainly in the manufacturing sector. Various occupations and positions are represented from engineering, finance, and information technologies departments. In terms of gender distribution 25% of the participants ($N = 78$) were women and 75% of the participants ($N = 226$) were men. Age of the participants ranged between 20 and 58 years ($M = 34.53$ years, $SD = 7.78$ years). Distribution of education level of the participants was as follows: 23 had a graduate degree, 171 had a bachelors degree, 39 were vocational college graduates, 68 had a high school degree (43 vocational high school, 25 regular high school), and three were primary school graduates. Total work experience of participants ranged from one to 468 months ($M = 127.67$ months, $SD = 100.01$ months) while tenure ranged from one to 372 months ($M = 77.90$ months, $SD = 82.13$ months).

2.2 Measures

2.2.1 Development of the SJT of Personality

A situational judgment test aiming to tap into the Big Five dimensions of personality as theorized by McCrea and Costa (1992) was developed. The development of the SJT basically followed the three-step approach prescribed by Motowidlo, Dunette, and Carter (1990) along with the implicit trait policies (ITP) theory framework provided by Motowidlo, Hooper and Jackson (2006). The three steps were critical incident collection, item stem and response option development, and focus groups with SME's and scoring key development; which are explained in detail below.

Critical incident collection:

The first step of the test development was critical incident collection from employees via a questionnaire (see Appendix A for Critical Incidents Questionnaire). In the instruction part of the questionnaire in addition to the information about the questionnaire, employees were provided with brief descriptions of each personality factor. Each questionnaire consisted of two parts; the first part was about a positive incident and the second part was about a negative incident. Employees were instructed to think about an incident, situation or event in which a personality factor of the actor played a major role. They were then asked to describe the situation, the observed outcome/consequence, and what course of action could have been better (or worse for the second part) on the part of the actor in that situation. In the final question, they were asked to indicate what personality factor of the actor played a role in the described situation.

A total of 120 questionnaires were distributed to the white collar employees in various departments, resulting in the collection of 240 critical incidents. Complete and meaningful incidents were sorted and they were assigned to a personality factor based on content.

Item stem and response option development:

Incidents that were identified as proper for the test development were reviewed by the researcher to be converted into SJT items. In addition to the information obtained from critical incidents, in the item development stage, the Big Five theoretical framework was adopted. Though, the instructions in the critical incidents questionnaire specifically asked for personality related incidents, not all the incidents were suitable to use in personality related scenarios. Thus, personality factor related parts were added to the items and response options according to the theoretical framework.

The main incident provided in the questionnaire served for item stem development. The essence of an SJT question is to present ambiguous, unobvious situations to the participants along with alternative behavioral responses. Hence, in the present study, scenarios collected from job incumbents were edited so that each one represented a challenging situation that may be dealt in a number of different ways. Item stems were moderate in length ($M = 105$ words). The items were

developed in such a way that they would be applicable to white collar employees in a wide range of positions in manufacturing sector. Therefore attention was paid not to use specific jargon of a position or a department.

In developing the response options, information gathered via the critical incident questionnaire was utilized. While doing so, however, the theoretical framework set by the Big Five Model was abided by. Each item in the test was intended to tap into one of the five personality factors. As a result, the response options developed for an item were behavioral demonstrations of that factor in varying degrees. For example, for an item that aimed to assess conscientiousness, the response options represented different courses of action that people with different levels of conscientiousness would follow according to the theory. Thus, five response options resulted as reflections of five levels of conscientiousness from 1 = behavioral demonstration of very low conscientiousness level to 5 = behavioral demonstration of very high conscientiousness level. In total 36 items were developed, five questions for Openness to Experience factor, eight questions for Conscientiousness, six questions for Extraversion, nine questions for Agreeableness, and eight questions for Emotional Stability.

Another source used in item development was the study by Sümer et al. (in progress). In this study, an SJT was developed to assess various attitudinal and personality-related factors including the Big Five Personality factors for blue collar employees working for the same organization. Situational judgment test questions from this study which aim to assess personality were reviewed for the present study. Some of the scenarios were adapted for white collar employees, by preserving the main theme but changing the specific details (e.g., jargon, work details) originally developed for blue collar employees. Response options were also reviewed both for adapting them to the white collar context and for making them more in line with the theoretical foundation of the Big Five model. A total of 24 questions were adapted from the blue collar SJT set, which increased the number of questions to 10, 14, 10, 10, and 16 for Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability, respectively. All the questions were reviewed by the researcher for the purpose of both theoretical and grammatical corrections.

Focus groups with SME's and development of the scoring key:

Two sets of focus groups were conducted with two different groups of SME's. The first group consisted of experts from the organization that the test was developed for, the second groups was composed of experts from the clinical psychology field.

The aim of the first focus group session was to ensure that the content of the scenarios were appropriate for the target group. Four focus groups were conducted with five to eight, middle to high level managers as the SMEs. Each group evaluated 15 SJT items and group meetings lasted approximately two hours. SMEs in the focus groups were asked to read the items, and indicate whether the situation depicted in the item was realistic and plausible within their organization. If an item was found to be realistic by each and every participant, they were then asked to provide the best and worse course of action in response to the situation. This information was used to further revise/improve the response options developed by the researcher.

At the end of this study, following revisions were done for some of the items; jargon was adjusted to address all the employees; additional information was added to some items necessary for the solution of the problem presented; and new response options were added.

The aim of the second focus group was to finalize the scoring key. Originally, the response options were developed to reflect predefined levels of the personality factors. The scoring of the response options was finalized by the SME group consisting of doctoral level clinical psychology students. Five focus group sessions were conducted with three different experts in each. Each group worked on a scoring key for one personality factor. The SMEs were presented with the definitions of Big Five personality factors as defined by Costa and McCrea (1992). At the beginning of the session, participants read the questions and rated the response option from "1 = This behavior is very typical of low level of the factor in question" to "5 = This behavior is very typical of high level of the factor in question," individually. In the second part of the session, SMEs shared their individual opinions and discussed until a consensus was reached. Sessions lasted approximately two hours. At the end, the scoring key, in which all the response

options were assigned a rating from 1 to 5 on the intended factor, was developed. A decision was made to use the “would format” in the SJT items as it was shown to have higher correlations with personality (McDaniel & Nguyen, 2001).

The final version of the test consisted of 60 SJT items ordered randomly. Following the software development, the test became a computer-based one in terms of application and data recording.

2.2.2 NEO-Personality Inventory Revised (NEO-PI-R)

The NEO-PI-R, developed by McCrea and Costa (1992), is one of the most widely used Big Five personality inventories. Factors and the facets of factors in this inventory are as follows: Neuroticism (anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability), Extraversion (warmth, gregariousness, assertiveness, activity, excitement-seeking, positive emotions), Openness to Experience (fantasy, aesthetics, feelings, actions, ideas, values), Agreeableness (trust, straightforwardness, altruism, compliance, modesty, tender-mindedness), and Conscientiousness (competence, order, dutifulness, achievement striving, self-discipline, deliberation). Adaptation of this inventory to the Turkish culture was conducted by Gülgöz (2002). Alpha coefficients at the facet level were found to range between .44 - .78 for Neuroticism, .56 - .75 for Extraversion, .45 - .77 for Openness to Experience, .44 - .72 for Agreeableness, .69 - .84 for Conscientiousness in the Turkish sample. This inventory was used for validation purposes in the present study since it is a widely used Big Five personality assessment tool in many countries and cultures including Turkey.

2.2.3 Big Five Inventory (BFI)

The BFI was developed by Benet-Martinez and John (1998) to assess personality using the Big Five framework, and it was adapted to Turkish by Sümer and Sümer (2002). The inventory consisted of 44 adjectives/phrases for which participants are asked to consider the question “I consider myself as ...” and indicate their responses on a 5-point scale (1= I don’t agree at all; 5= Completely agree). The administration of the inventory takes 10 to 20 minutes. For the present study, a computer-administered version of the BFI was created. In the current study internal

consistency reliabilities of the BFI were as follows: .77 for Openness to Experience, .73 for Conscientiousness, .73 for Extraversion, .51 for Agreeableness, and .74 for Neuroticism.

2.2.4 Personnel Multiple Reasoning Test (PMRT)

The PMRT, developed by Sümer, Er, Sümer, Ayvaşık, Mısırlısoy, and Erol-Korkmaz (2011), is a nonverbal reasoning test developed for personnel selection purposes. The PMRT consists of five subtests; analogy, matching, sequencing, reconstruction, and mental rotation and each subtest contain 20 items. Alpha coefficients for the subtests were reported to be .85 for analogy, .78 for matching, .82 for sequencing, .76 reconstruction, and .87 for mental rotation. The test is completely computerized.

2.2.5 Job Performance Measure

The job performance measure used in the present study was composed of the relevant dimensions of the annual performance evaluations of the organization for white collar employees. Performance data were obtained from personnel files. The performance evaluation form is filled out by the immediate supervisor of each employee. An employee is assessed on the dimensions that are relevant to his/her job. The assessment is conducted on a five point scale. There were four performance dimensions used in the current study, which were *self development*, *innovation*, *teamwork*, and *team leadership*. These four dimensions were believed to be more related to the domain of personality. Detailed information about these dimensions are provided in the results and discussion sections.

2.3 Procedure

All participants completed the SJT followed by the BFI. A subset of the participants took NEO-PI-R ($N = 92$), and PMRT ($N = 145$). Job performance data were obtained from the Human Resource Department for 138 participants. Except for the SJT and BFI, which were embedded into the same software, the tests were administered in different sessions to minimize common method problems and demand characteristics.

CHAPTER 3

RESULTS

3.1 Overview

This section presents descriptive statistics as well as the analyses conducted to finalize the developed SJT and to test the hypotheses of the study along with some additional analyses. First, correlations among the study variables and descriptive statistics are presented. Second, analyses based on the revisions/modifications made in the SJTs are described. Then the analyses testing the hypotheses of the study are presented. The first hypothesis was about the reliability of the scale and it was assessed through inter item reliability and test re-test reliability. The second hypothesis was about the construct validity of the test. Construct validity, in terms of convergent and divergent validity was tested with two different methods, Multitrait-Multimethod Matrix and Confirmatory Factor Analysis procedures. In addition, one exploratory hypothesis related to divergent validity was tested. Finally, the third hypothesis was about criterion validity. The relationship between personality scores and job performance was examined via correlational analyses results.

3.2 Correlations among Variables, Reliabilities and Descriptive Statistics

The correlations among study variables, reliabilities, means and standard deviations are presented in Table 3.1. Among the demographic variables, age was only significantly correlated with one personality variable which is Neuroticism measured by BFI ($r = .14, p < .05$). Men were found to score higher on the Extraversion scale of NEO-PI-R. Level of education was found positively correlated with three factors of SJT, Openness to Experience, Extraversion and Emotional Stability and negatively correlated with one factor of NEO-PI-R, Extraversion.

Among the personality variables measured with the SJT, Openness to Experience was positively correlated with Extraversion and Emotional Stability. Conscientiousness was significantly correlated with Agreeableness and Emotional Stability. Extraversion was correlated negatively with Agreeableness and positively with Emotional Stability. The only significant intercorrelation among NEOPI-R factors was a positive correlation between the Neuroticism and Conscientiousness factors.

All personality variables assessed via BFI were significantly correlated with each other and all the correlations, except for Neuroticism, were in positive way while Neuroticism was negatively correlated with all the other personality factors.

The correlations among the corresponding factors of the SJT, BFI, and NEO-PI-R were mostly in the expected direction and magnitude mostly, which will be examined in detail in the following sections. Emotional Stability, Openness to Experience and Extraversion factors assessed with SJT had high correlations with their corresponding factors assessed with NEO-PI-R. Openness to Experience, Conscientiousness and Extraversion factors assessed with SJT, on the other hand, have high correlations with the same factors assessed with BFI.

Openness to Experience and Extraversion factors of SJT had high correlations with their conceptual counterparts in both the NEO-PI-R and BFI, whereas, Emotional Stability factor of SJT had a high correlation only with Emotional Stability factor of NEO-PI-R. Conscientiousness factor of SJT had a high correlation with Conscientiousness factor of BFI. Agreeableness factor of SJT did not correlate with the conceptual counterpart of the other two other measures.

Table 3.1 *Correlations among study variables, reliabilities and descriptive statistics*

Variables	1	2	3	4	5	6	7	8	9	10
1 Age	-									
2 Gender	-.19**	-								
3 Level of Education	-.05	.10	-							
4 Tenure	.76*	-.12**	-.21**	-						
5 Work Experience	.92*	-.17**	-.20**	.79**	-					
6 SJT Openness to Experience (10)	.01	-.03	.31**	-.07	-.08	.44				
7 SJT Conscientiousness (15)	.00	.02	.10	-.08	.01	.00	.25			
8 SJT Extraversion (10)	-.05	-.07	.26**	-.11	-.08	.40**	.09	.52		
9 SJT Agreeableness (9)	.02	-.08	.00	-.04	.04	-.11	.18**	-.20*	.22	
10 SJT Emotional Stability (16)	.05	-.03	.17**	-.05	.00	.21**	.21**	.27**	.06	.17
11 NEO-PI-R Openness to Experience	.18	-.17	.26*	.12	.13	.37**	.09	.20*	-.04	.12
12 NEO-PI-R Conscientiousness	-.05	-.02	-.07	.07	-.08	.00	.18	.08	-.06	.00
13 NEO-PI-R Extraversion	-.02	.26*	.04	-.06	.02	.08	.08	.36**	-.02	-.08
14 NEO-PI-R Agreeableness	.00	-.15	-.02	.09	.00	-.02	.12	.11	.07	.04
15 NEO-PI-R Neuroticism	.07	.19	-.26**	.30**	.17	-.14	-.04	-.21*	.05	-.39**
16 BFI Openness to Experience	.05	.02	-.02	-.02	.06	.21**	.03	.17**	-.02	.00
17 BFI Conscientiousness	-.06	.07	.00	-.06	-.03	.04	.12*	.10	.04	.10
18 BFI Extraversion	-.11	-.04	.07	-.14*	-.13*	.05	.09	.13*	-.07	.04
19 BFI Agreeableness	-.06	.06	-.07	-.08	-.04	.07	.09	.08	.08	.04
20 BFI Neuroticism	.14*	-.01	-.03	.13*	.15**	-.08	-.05	-.11	-.03	-.07
Mean	34.54	-	-	77.90	127.67	3.48	3.84	3.94	3.63	4.26
Standard Deviation	7.78	-	-	82.13	100.01	.42	.30	.42	.45	.24

Table 3.1 (continued)

Variables	11	12	13	14	15	16	17	18	19	20
11 Openness to Experience	-									
12 NEO-PI-R Conscientiousness	.16	-								
13 NEO-PI-R Extraversion	.07	.05	-							
14 NEO-PI-R Agreeableness	.09	.03	.00	-						
15 NEO-PI-R Neuroticism	-.05	.35**	.13	.08	-					
16 BFI Openness to Experience	.28**	.15	.25*	-.15	.07	.77				
17 BFI Conscientiousness	-.02	.45	.13	.15	.00	.40**	.73			
18 BFI Extraversion	-.08	.10	.19	-.23*	-.08	.39**	.27**	.73		
19 BFI Agreeableness	.02	.26*	.17	.16	.12	.25**	.40**	.13*	.51	
20 BFI Neuroticism	.08	-.30	.02	-.08	.20*	-.25**	-.45**	-.23**	-.41**	.74
Mean	52.14	53.46	52.03	50.7	41.83	4.03	4.32	3.60	4.17	2.19
Standard Deviation	8.36	8.28	7.80	8.59	9.78	.50	.49	.56	.40	0.59

Note. Gender 1 = Men, 2 = Women; Level of Education 1 = Primary and Secondary School, 2 = High School, 3 = Vocational High School, 4 = Vocational College, 5 = University Graduate, 6 = Master's Degree. Work Experience = Total work experience in months. Tenure = Work experience in the organization in months. SJT Variables assed on a five point scale. BFI variables are measured on 5 point-Likert Scale 1 = Disagree Strongly 5 = Strongly Agree. * $p < .05$, ** $p < .01$. $N_{\text{NEO-PI-R}} = 95$, $N_{\text{PGMT}} = 145$. Number of items for each SJT factor is indicated in parenthesis. Alpha's are bolded in diagonals

3.3. Reliability: Internal Consistency and Test-Retest Estimates

Internal Consistency reliability estimates of the factors of SJT were presented in Table 3.1. The values ranged between .17 and .52. The factor with highest reliability was Emotional Stability and the factor with lowest reliability was Agreeableness. In general, the values were not in satisfactory level.

Test-retest reliability study was conducted with a subset of participants (N = 59). According to the results, test-retest correlation of the factors of the SJT ranged between .75 and .22. As can be seen in Table 3.2, except for Agreeableness factor all the correlations were significant. In addition, pattern of the correlations among factors were parallel in test 1 and test 2 (see Tables 3.2 and 3.1). The results of analysis showed that Openness to Experience and Extraversion factors have satisfactory test-retest reliability ($r = .72, .75$, respectively) Conscientiousness and Emotional Stability factors have acceptable test-retest reliability ($r = .43, .48$, relatively) while Agreeableness factor have insignificant test-retest reliability ($r = .22$).

Table 3.2 *Test Retest Reliability and Correlations among Factors of SJT*

Test 1	Test 2				
	Openness to Experience	Conscientiousness	Extraversion	Agreeableness	Emotional Stability
Openness to Experience	.72**				
Conscientiousness	-.02	.43**			
Extraversion	.40**	-.08	.75**		
Agreeableness	.12	.17	.03	.22	
Emotional Stability	.22	.11	.23	.14	.48**

3.4 Test Modifications

The SJT developed in this study with the purpose of assessing the Big Five dimensions originally consisted of 60 items. Prior to validity analyses items were investigated individually to identify strong and weak ones. This analyses was conducted by examining the correlations between the individual SJT items of a personality factor with the mean score of the corresponding NEO-PI-R and BFI

factors. These analyses were conducted for each factor separately. Some items had significant correlations in the expected direction while some items had insignificant or zero correlations. As a rule of thumb, a decision was made to eliminate the items with correlations lower than .08 with the corresponding NEO-PI-R factor.

Accordingly, three items in Openness to Experience, eight items in Conscientiousness, three items in Extraversion, five items in Agreeableness and 8 items in Emotional Stability failed to meet the .08 criteria and thus were eliminated, resulting in 33 items; seven items for Openness to Experience, seven items for Conscientiousness, seven items for Extraversion, four items for Agreeableness, and eight items for Emotional Stability. While decreasing the number of items, this procedure increased the correlations among the corresponding SJT, BFI and NEO-PI-R factors, in turn contributing to the convergent validity of the developed test (See Table 3.3). Shortening the test was also advantageous for practical reasons since test completion time dropped significantly. All the analyses described below, unless otherwise stated, were conducted with shortened version of the test.

3.5 Construct Validity: Multitrait-Multimethod Matrix Approach

Table 3.3 presents Multitrait-Multimethod Matrix (MTMM) which is constructed according to Campbell and Fiske (1959) using correlation coefficients between the three techniques (i.e., NEO-PI-R, BFI, and SJT) five constructs (i.e., Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability). Reliabilities, which can also be defined as monotrait-monomethod values, are presented at the diagonal in parenthesis. Italicized correlations show heterotrait-monomethod correlations. These correlations are results of analysis of different traits which are measured by the same method. Together with the reliability diagonal, italicized correlations show the monomethod block. Underlined values show heterotrait-heteromethod correlations. These values are results of correlations of different traits measured by different methods. Bolded correlations show the validity diagonal. These correlations are results of analysis of the same traits which are measured by different methods. Combination of heterotrait-heteromethod values and the validity diagonal values compose the heteromethod block (Campbell & Fiske, 1959).

According to Campbell and Fiske (1959), multitrait-multimethod matrices have four requirements in the establishment of the construct validity of a measure. The first requirement is that values in the table should be higher than zero. This requirement is related to convergent validity. The second requirement is that values in the validity diagonal should be higher than values in its column and row in the heterotrait-heteromethod triangle. The third requirement is correlations in validity diagonal should be higher than heterotrait-monomethod values. That is, correlations between different methods of the same trait should be higher than correlations between different traits measured with the same method. Finally the fourth requirement is that trait interrelationships in all the heterotrait triangles should be parallel in both monomethod and heteromethod blocks. Requirements two to four show discriminant validity. Correlations among the study variables are examined according to the four requirements stated above.

3.5.1 Convergent Validity

Convergent validity was assessed through examination of the relationships among correlations between the corresponding measures of the same constructs using different methods. As stated above convergent validity corresponds to the second requirement (i.e., examination of validity diagonals) according to Campbell and Fiske (1959). This criterion was met for the SJT and the corresponding NEO-PI-R dimensions. All the correlations in validity diagonals were stronger than the other correlations in their rows and columns (see Table 3.3). Correlations between the SJT factors and NEO-PI-R dimensions were -.52 for Neuroticism (the direction is expected to be negative higher scores in the SJT indicate Emotional Stability), .40 for Openness to Experience, .40 for Extraversion, .28 for Conscientiousness, and .22 for Agreeableness.

This criteria, however, was only met for Openness to Experience and Conscientiousness factors in the SJT and BFI relationships. Correlations between the corresponding SJT and BFI dimensions were .24 for Openness to Experience, .15 for Conscientiousness, .14 for Extraversion, .03 for Agreeableness, and .00 for Neuroticism.

Hypothesis 2 suggested that the same/corresponding personality factors assessed with NEO-PI-R and BFI and with the SJT methodology would have higher correlations than the correlations between uncorresponding personality factors assessed with NEO-PI-R and BFI and the SJT methodology. Regarding the results it can be concluded that Hypothesis 2 was supported by the SJT-NEO-PI-R relationships but not by the SJT-BFI relationships.

3.5.2 Divergent Validity

Divergent validity was assessed using the third and fourth requirements suggested by Campbell and Fiske (1959). The third requirement involves a comparison of validity diagonals and heterotrait-monomethod triangles. This comparison is expected to show that different traits measured with the same method have lower correlations than the same traits assessed with different methods, suggesting low method variance. This criterion was nearly met for the SJT and NEO-PI-R correlations. Correlations of Neuroticism, Openness to Experience, and Extraversion in validity diagonal were higher than any other correlation in heterotrait-heteromethod block of the SJT. On the other hand, correlations of Conscientiousness and Agreeableness in validity diagonal were lower than the correlation between Openness to Experience and Extraversion factor of the SJT. In terms of the SJT and BFI, heterotrait-monomethod correlations were higher than validity correlations, suggesting existence of some method bias.

The last requirement suggests comparing heterotrait-monomethod correlations with heterotrait-heteromethod correlations. This comparison shows whether there is method variance. As can be seen in Table 3.3, the SJT method yielded high intercorrelations among the traits measured. However, considering the related nature of the personality factors this correlation was hard to interpret as a significant method variance.

Table 3.3 *Multitrait-Multimethod Matrix for Personality Factors Assessed via SJT, BFI, and NEO-PI-R*

		SJT					NEO-PI-R					BFI				
	Variables	O ₁	C ₁	E ₁	A ₁	ES ₁	O ₂	C ₂	E ₂	A ₂	N ₂	O ₃	C ₃	E ₃	A ₃	N ₃
SJT	Openness	(.37)														
	Conscientiousness	.07	(.10)													
	Extraversion	.35**	.09	(.51)												
	Agreeableness	-.20**	.07	-.19**	(.34)											
	Emotional Stability	.08	.12*	.20**	-.02	(.03)										
NEO-PI-R	Openness	.40**	<u>.00</u>	<u>.24**</u>	<u>-.07</u>	<u>.07</u>										
	Conscientiousness	<u>.09</u>	.28**	<u>-.02</u>	<u>-.11</u>	<u>-.05</u>	.16									
	Extraversion	<u>.10</u>	<u>.12</u>	.40**	<u>-.07</u>	<u>-.06</u>	.07	.06								
	Agreeableness	<u>.09</u>	<u>-.01</u>	<u>.15</u>	.22*	<u>-.02</u>	.09	.03	-.01							
	Neuroticism	<u>-.14</u>	<u>.02</u>	<u>-.22*</u>	<u>.08</u>	-.52**	-.05	.35**	.13	.08						
BFI	Openness	.24**	<u>.07</u>	<u>.18**</u>	<u>-.05</u>	<u>-.08</u>	.28**	<u>-.03</u>	<u>-.08</u>	<u>.02</u>	<u>.08</u>	(.77)				
	Conscientiousness	<u>.10</u>	.15**	<u>.09</u>	<u>.03</u>	<u>-.03</u>	<u>.15</u>	.45**	<u>.10</u>	<u>.26</u>	<u>-.30**</u>	<u>.40**</u>	(.73)			
	Extraversion	<u>.07</u>	<u>.11*</u>	.14*	<u>-.16**</u>	<u>-.03</u>	<u>.16</u>	<u>.13</u>	.19	<u>.17</u>	<u>.02</u>	<u>.39**</u>	<u>.27**</u>	(.73)		
	Agreeableness	<u>-.20**</u>	<u>.07</u>	<u>.11</u>	.03	<u>-.07</u>	<u>-.01</u>	<u>.15</u>	<u>-.23*</u>	.16	<u>-.08</u>	<u>.25**</u>	<u>.41**</u>	<u>.13*</u>	(.51)	
	Neuroticism	<u>-.07</u>	<u>-.13*</u>	<u>-.04</u>	<u>.02</u>	.00	<u>.07</u>	<u>.01</u>	<u>-.08</u>	<u>.12</u>	.21*	-.25	<u>-.45**</u>	<u>-.22**</u>	<u>-.41*</u>	(.74)

Note. Bolded numbers show validity diagonal correlations, italicized numbers show heterotrait-monomethod correlations, underlined numbers show heterotrait-heteromethod correlations. * $p < .05$, ** $p < .001$.

3.6 Divergent Validity Evidence in terms of Relationship with a Nonverbal Reasoning Test

In order to further examine divergent validity of the developed SJTs, an additional multitrait-multimethod matrix analysis was conducted using a nonverbal reasoning test as the alternative trait. As suggested as an exploratory hypothesis, another indicator of divergent validity of the current test would come from its relationship with the reasoning test scores. This analysis was conducted by correlating the SJT dimension scores with the Personnel Multiple Reasoning Test (PMRT) scores. As an evidence for divergent validity, the SJT factors were expected not to correlate significantly with the PMRT score, which is believed to reflect general cognitive ability. The results of these analyses are presented in Table 3.4

Table 3.4 *Correlations among the SJT Factors and the Personnel Multiple Reasoning Test*

SJT Variables	PMRT
Openness to Experience	.11
Conscientiousness	.14
Extraversion	.02
Agreeableness	-.05
Emotional Stability	.11

As expected, the PMRT score was not significantly correlated with any of the five personality dimensions assessed by the SJTs, yielding some evidence for divergent validity of the test.

3.7 Multitrait-Multimethod Matrix with Selected Items

Following the test modifications based on the correlational item analyses, an additional item selection procedure was followed to be able to conduct a confirmatory factor analysis on a shortened version of the test. Four items from each factor of BFI and SJT were selected for confirmatory factor analysis based on intercorrelations. The multitrait-multimethod matrix was repeated with the selected

items and items of NEO-PI-R. For NEO-PI-R, all the items were included in the analysis since item based results of NEO-PI-R was not available.

The multitrait-multimethod matrix is presented in Table 3.5. The table was analyzed in terms of convergent validity, divergent validity and method variance as Campbell and Fiske suggested (1959). As explained above, in this procedure, three types of correlations, validity diagonal (bolded in table), heterotrait-monomethod correlations (italicized in table), and heterotrait-heteromethod correlations (underlined in table) are examined with comparison to each other.

In terms of convergent validity the NEO-PI-R and SJT relationships were in general stronger than the BFI and SJT relationships, parallel to the previous analyses with the full item set (see Tables 3.3 and 3.5). Only the Agreeableness factor failed to show convergent validity in the NEO-PI-R and SJT relationships. In the SJT and BFI relationships, on the other hand, only the Extraversion factor showed convergent validity evidence.

In terms of divergent validity, in the SJT and NEO-PI-R relationships, only the Extraversion factor met the criteria, while all other factors failed to meet the criteria in the SJT and BFI relationships.

Finally, when the table is examined in terms of method variance, it was ascertained that monomethod correlations were higher than heteromethod correlations, signaling existence of method variance.

To conclude, this MTMM analysis on the reduced item sets showed slightly poorer results than the MTMM analyses with all items included. Evidence of convergent and divergent validity was not strong and method variance was detected. In the following step, these selected items were used in confirmatory factor analysis to assess construct validity of the SJT dimensions.

Table 3.5 *Multitrait-Multimethod Matrix with Selected Items*

		SJT					NEO-PI-R					BFI					
		O ₁	C ₁	E ₁	A ₁	ES ₁	O ₂	C ₂	E ₂	A ₂	N ₂	O ₃	C ₃	E ₃	A ₃	N ₃	
SJT	Openness	(.36)															
	Conscientiousness	<i>-.03</i>	(.38)														
	Extraversion	<i>.28**</i>	<i>-.09</i>	(.35)													
	Agreeableness	<i>.13*</i>	<i>.28**</i>	<i>-.20**</i>	(.34)												
	Emotional Stability	<i>.12*</i>	<i>.16**</i>	<i>.20**</i>	<i>.21**</i>	(.35)											
NEO-PI-R	Openness	<i>.27**</i>	<i>.03</i>	<i>.18</i>	<i>-.07</i>	<i>.03</i>	-										
	Conscientiousness	<i>.14</i>	<i>.10</i>	<i>.02</i>	<i>-.11</i>	<i>-.17</i>	<i>.16</i>	-									
	Extraversion	<i>.16</i>	<i>.11</i>	<i>.40**</i>	<i>.07</i>	<i>.08</i>	<i>.07</i>	<i>.06</i>	-								
	Agreeableness	<i>.11</i>	<i>.07</i>	<i>.10</i>	<i>.22*</i>	<i>.24*</i>	<i>.09</i>	<i>.03</i>	<i>.01</i>	-							
	Neuroticism	<i>-.17</i>	<i>.18</i>	<i>-.19</i>	<i>.08</i>	<i>.28**</i>	<i>-.05</i>	<i>.35**</i>	<i>.13</i>	<i>.08</i>	-						
BFI	Openness	<i>.07</i>	<i>-.04</i>	<i>.09</i>	<i>.01</i>	<i>-.04</i>	<i>.28**</i>	<i>.03</i>	<i>.22*</i>	<i>-.12</i>	<i>.07</i>	<i>.62</i>					
	Conscientiousness	<i>.05</i>	<i>.05</i>	<i>.09</i>	<i>.01</i>	<i>.00</i>	<i>-.08</i>	<i>.41**</i>	<i>.10</i>	<i>-.04</i>	<i>.12</i>	<i>.31**</i>	<i>.75</i>				
	Extraversion	<i>.08</i>	<i>.03</i>	<i>.20**</i>	<i>.13*</i>	<i>-.01</i>	<i>-.06</i>	<i>.18</i>	<i>.22*</i>	<i>-.19</i>	<i>-.01</i>	<i>.39**</i>	<i>.44**</i>	<i>.77</i>			
	Agreeableness	<i>-.06</i>	<i>.03</i>	<i>.03</i>	<i>.06</i>	<i>.03</i>	<i>.01</i>	<i>.23*</i>	<i>.11</i>	<i>.23*</i>	<i>.01</i>	<i>.16**</i>	<i>.33**</i>	<i>.05</i>	<i>.53</i>		
	Neuroticism	<i>-.07</i>	<i>-.03</i>	<i>-.09</i>	<i>.05</i>	<i>-.02</i>	<i>0.12</i>	<i>-.34**</i>	<i>-.03</i>	<i>-.01</i>	<i>.19</i>	<i>.22*</i>	<i>-.30**</i>	<i>-.41**</i>	<i>-.27**</i>	<i>.66</i>	

Note. Bolded numbers show validity diagonal correlations, italicized numbers show heterotrait-monomethod correlations, underlined numbers show heterotrait-heteromethod correlations. * $p < .05$, ** $p < .001$.

3.8 Construct Validity: Confirmatory Factor Analysis Approach

Validity of the SJT was tested with models for construct validity via CFA as suggested by Widaman (1985). This procedure involves comparing nested models to examine convergent validity, discriminant validity, and method variance. The SJT and BFI dimension scores of 302 participants were used in this procedure. There are six models tested in this method. The first model is the null model. The second model is the trait model (in the current study this model is composed of five trait factors). The third model is the method model, which is composed of two method factors, namely, SJT and BFI. The fourth model is the general trait model with method factors; in this model in addition to the two method factors, there is also a general trait factor instead of five different traits. The fifth model is the trait-method model with orthogonal factors. In this model there are two uncorrelated methods and five factors. Finally, the sixth model is the trait-method model with oblique factors. As the name implies in this model there are two correlated methods and five factors (for more information see Widaman, 1985). Table 3.6 presents the titles of the six models, and the models are graphically presented in Appendices F through K.

Table 3.6 *Models and Characteristics of Models*

Models	Characteristics
Model 1	Null Model
Model 2	Trait Model
Model 3	Method Model
Model 4	General Trait Model
Model 5	Trait Method Model with orthogonal factors
Model 6	Trait Method Model with correlated method factors

Table 3.7 *Correlations among Manifest Variables*

	Item	1	2	3	4	5	6	7	8	9	10
1	SJT	O1	-	-	-	-	-	-	-	-	-
2		O2	.03	-	-	-	-	-	-	-	-
3		O3	.12*	.15**	-	-	-	-	-	-	-
4		O4	.17*	.22**	.19**	-	-	-	-	-	-
5		C1	.02	-.06	.08	-.01	-	-	-	-	-
6		C2	.05	-.11	-.01	-.08	.14*	-	-	-	-
7		C3	-.07	-.01	.05	.05	.13*	.12*	-	-	-
8		C4	.09	.05	.04	-.07	.11	.12*	.20**	-	-
9		E1	.28**	.06	-.02	.12*	.13*	.01	.00	-.04	-
10		E2	.17**	.01	.01	.09	.04	-.01	-.06	-.07	.11*
11		E3	.15**	-.03	.11	.12*	-.11	-.05	.00	-.16**	.14*
12		E4	.25**	.06	.04	.14*	-.10	.00	-.06	-.03	.22*
13		A1	.00	-.13*	-.16**	-.06	-.04	.04	.07	.11*	-.16**
14		A2	-.03	-.17**	.04	-.11	.18*	.03	.16*	.08	.04
15		A3	.03	.07	.12*	.06	.03	.06	.10	.03	.00
16		A4	-.13*	-.02	.06	-.05	.05	.12*	.15*	-.02	-.14*
17		ES1	.07	-.05	.09	.05	-.03	.06	.14*	.12*	.06
18		ES2	.07	-.01	.02	.04	.15**	-.03	.00	.03	.15**
19		ES3	.03	-.01	.09	.04	.11	-.02	.01	-.02	.17**
20		ES4	.12*	-.01	.10	.12*	.07	.02	.08	.13*	.05
21	BFI	O1	-.05	.09	.02	.13*	-.07	-.01	-.05	.00	-.01
22		O2	.01	.05	.04	.03	-.04	.03	-.01	.10	.05
23		O3	.08	.00	-.05	.00	-.12*	-.05	.05	.03	.04
24		O4	.09	.00	.01	.03	-.14*	.02	-.01	.01	.00
25		C1	-.01	.06	.10	-.01	-.02	.09	-.06	.04	-.08
26		C2	.02	-.01	.14*	.04	-.05	.02	.00	-.01	.02
27		C3	.01	.07	.01	.00	.00	.07	.04	.07	.10
28		C4	-.02	.02	.06	-.04	-.05	.06	.03	.11	-.03
29		E1	.05	.11	.05	.02	-.07	.14*	-.02	.04	.05
30		E2	.05	.07	-.02	.02	-.09	.07	-.11	-.07	.11*
31		E3	.00	.06	-.03	.02	.00	.09	-.01	.03	.17**
32		E4	.04	.04	.02	.02	-.01	.12*	.11	.09	.10
33		A1	.04	.04	-.04	.01	-.06	.07	.00	.05	.07
34		A2	.02	-.06	.06	-.03	.00	.03	-.07	.06	.01
35		A3	-.03	-.07	.00	-.05	-.01	.00	.03	.04	-.01
36		A4	-.04	-.07	-.02	-.05	-.09	-.04	.03	.10	.03
37		N1	-.07	.07	.06	.03	-.11	.04	-.05	-.03	.03
38		N2	.07	.07	.05	.11	-.01	.12*	.04	.11	.07
39		N3	.03	.04	.04	.08	-.02	.04	-.03	.06	.00
40		N4	-.04	-.01	-.06	-.03	-.06	.08	-.02	.02	.10

Table 3.7 (continued)

	11	12	13	14	15	16	17	18	19	20
12	.30**	-	-	-	-	-	-	-	-	-
13	-.90	-.07	-	-	-	-	-	-	-	-
14	-.10	-.13*	.14*	-	-	-	-	-	-	-
15	.02	.12*	.00	.16*	-	-	-	-	-	-
16	-.21**	-.17**	.06	.14*	.19**	-	-	-	-	-
17	.00	.06	-.05	.04	.13**	.19**	-	-	-	-
18	.00	.09	.07	.18*	.18**	.01	.17**	-	-	-
19	-.03	.04	-.10	.08	.10	.08	.13**	.17**	-	-
20	.06	.14*	.04	.02	.07	.11	.14*	.04	.07	-
21	.02	.06	-.02	-.09	-.03	-.05	.01	-.03	-.11*	.05
22	.07	.08	.02	-.04	.03	-.02	.03	.04	-.07	.02
23	.12*	.04	.03	.00	.08	-.09	-.02	-.08	-.09	.02
24	.03	.14	.01	.00	.08	.07	.05	-.04	-.06	.02
25	.05	.02	-.05	-.08	.09	-.01	-.04	.02	.01	.12*
26	.08	.13*	.04	-.07	.06	.05	.05	-.08	-.12*	.12*
27	.15*	.10	-.04	-.01	.20*	.00	.01	.00	.00	.05
28	.00	.01	-.08	-.07	.07	-.03	.00	-.04	-.06	.00
29	.07	.12*	-.05	-.12*	.06	-.09	.04	-.08	.00	.06
30	.06	.11	-.05	.09	-.06	-.10	.00	-.08	-.02	-.02
31	.07	.06	-.08	-.06	-.08	-.04	.07	-.09	-.05	.04
32	.09	.18	-.07	-.01	.06	-.07	.08	-.04	-.04	.12**
33	-.04	.08	-.09	.04	.02	.05	.06	.03	.05	-.02
34	.03	.14*	-.10	.01	.12*	-.01	.01	-.05	.03	.10
35	.00	.06	.06	.05	.08	-.05	.04	-.04	-.04	.04
36	.00	.03	.05	.02	.08	.02	.08	-.08	-.02	.06
37	.02	.08	.18**	-.01	.09	-.04	-.04	.05	.05	.06
38	-.05	.08	-.11	.02	.15*	.00	.02	.05	.01	.03
39	.08	.02	-.16*	.03	.03	.00	-.06	-.01	.06	.06
40	.04	.14*	-.12*	-.08	.03	-.06	.07	-.05	.07	.05

Table 3.7 (continued)

	21	22	23	24	25	26	27	28	29	30
22	.43**	-	-	-	-	-	-	-	-	-
23	.16**	.25**	-	-	-	-	-	-	-	-
24	.20**	.21**	.56**	-	-	-	-	-	-	-
25	.32**	.17**	.11	.13*	-	-	-	-	-	-
26	.23**	.18**	.16**	.21**	.49**	-	-	-	-	-
27	.18**	.19**	.15*	.11	.42**	.43**	-	-	-	-
28	.24**	.16**	.14*	.08	.50**	.41**	.40**	-	-	-
29	.36**	.36**	.12*	.13*	.39**	.35**	.30**	.39**	-	-
30	.32**	.35**	.09	.17**	.14*	.23**	.22**	.19**	.47**	-
31	.30**	.26**	.08	.03	.16**	.15**	.20**	.32**	.45**	.43**
32	.24**	.22**	.24**	.23**	.28**	.25**	.28**	.34**	.52**	.36**
33	.06	.06	.00	.02	.08	.06	.08	.04	.02	.00
34	.16**	.14*	.08	.11	.50**	.48**	.43**	.43**	.28**	.13*
35	-.09	.00	.10	.15**	.15*	.21**	.12*	.15**	.02	-.06
36	-.06	.04	.16**	.18**	.08	.15**	.14*	.09	.04	.00
37	.04	.03	.11*	.23**	.25**	.13*	.07	.22**	.30**	.18**
38	.19**	.16**	.01	.12*	.17**	.04	.14**	.25**	.30**	.31**
39	.06	.06	.13**	.16**	.19**	.08	.24**	.21**	.23**	.25**
40	.11	.09	.09	.15**	.09	.05	.18**	.25**	.22**	.11

Table 3.7 (continued)

	31	32	33	34	35	36	37	38	39
32	.52**	-	-	-	-	-	-	-	-
33	.08	.01	-	-	-	-	-	-	-
34	.19**	.21**	.23**	-	-	-	-	-	-
35	-.12*	.00	.21**	.16**	-	-	-	-	-
36	.10	.04	.20*	.14*	.41**	-	-	-	-
37	.14*	.21**	.14**	.10	.22**	.12*	-	-	-
38	.26**	.23**	.15**	.09	.03	.03	.32**	-	-
39	.20**	.25**	.08	.07	.18**	.18**	.32**	.35**	-
40	.14*	.23**	.18**	.13*	.12*	.25**	.24**	.35**	.41**

Note. * $p < .05$, ** $p < .001$. O = Openness to Experience, C = Conscientiousness, A = Agreeableness, E = Extraversion, N = Neuroticism, ES = Emotional Stability.

Since NEO-PI-R was administered only to 95 participants, because of the sample size requirements of CFA, this measure was omitted from CFA analyses. The SJT and BFI data both met the sample size criterion for CFA analysis ($N = 304$). On the other hand, SJT and BFI included 101 items in total. The required sample size to be able to use all the items of the tests in CFA model would be much higher than the current sample size. Thus, following the procedure used by Trippe (2002), four items with the highest intercorrelations in each factor were selected for the SJT and BFI, and the CFA was conducted on these items.

3.9 CFAs: Testing for the Trait and Method Effects

All the confirmatory factor models were tested using LISREL 8.51 (Jöreskog & Sörbom, 2006) using the data from 304 participants. Maximum likelihood estimation, with variance covariance matrix serving as input, was used for evaluating the model. Table 3.7 presents correlations of manifest variables and Table 3.8 presents χ^2 , df and selected fit statistics for the models tested.

According to generally accepted criteria (Schreiber, Stage, King, Nora, & Barlow, 2006) a good fit can be claimed if the ratio of chi-square to degrees of freedom is less than three, goodness of fit index (GFI) is .95 or higher, root mean square residual (RMR) is closest to 0, root-mean-square error of approximation (RMSEA) is .06 or below, comparative fit index (CFI), incremental fit index (IFI),

and normed fit index (NFI) are all .95 or higher. The results showed that though χ^2 was significant, χ^2 : df ratio was above 3:1 ratio.

Table 3.9 shows χ^2 difference tests between hierarchically nested confirmatory factor models. The comparison between Model 6 and Model 3 shows convergent validity. That is, the model with no trait but only method factors should provide poorer fit than the model with trait and method factors to be able to show convergent validity. As can be seen in Table 3.9, Model 6 provided significantly better fit to the data than Model 3 providing some evidence for convergent validity.

Comparison between Model 6 and Model 4 yielded divergent validity evidence. That is, if the model with two method factors and a general trait factor provides poorer fit than the model with method and trait factors, then divergent validity evidence is obtained. Results showed that Model 6 provided significantly better fit than Model 4, yielding evidence for divergent validity.

Comparison of Models 6 and Model 5 shows methods' covariance among each other while comparison of Model 5 and Model 2 shows method variance. That is, if the model with trait but no method factors provides poorer fit to the data than the model with method and trait factors, method variance is inferred. Results showed existence of a significant method factor. However covariation among methods was not significant, suggesting that covariation among measures was uniquely attributable to trait factors rather than method factors, supporting convergent validity.

Table 3.8 Selected Fit Statistics and χ^2 Values for Models Tested

Model	GFI	RMR	CFI	NNFI	RMSEA	χ^2	χ^2 :df	χ^2 :df
1						2692.91	780	3.45
2	.80	.06	.67	.64	.06	1466.14	730	2.00
3	.77	.06	.60	.53	.07	1747.45	739	2.36
4	.81	.06	.68	.64	.06	1389.14	699	1.99
5	.85	.05	.81	.78	.04	1076.28	690	1.55
6	.85	.05	.81	.78	.04	1073.62	689	1.55

Note. Model 1 = Null Model, Model 2 = Trait Model, Model 3 = Method Model, Model 4 = General Trait Model, Model 5 = Trait Method Model with Orthogonal Factors, Model 6 = Trait Method Model with Correlated Factors.

Table 3.9 χ^2 Difference Test Between Hierarchically Nested Models

Model Comparison	χ^2 difference	χ^2 difference	χ^2 critic	$p <$	Issue Addressed
6 vs.3	673.83	50	86.661	.001	Convergent Validity
6 vs. 4	315.52	10	29.588	.001	Divergent Validity
5 vs. 2	389.86	40	73.402	.001	Method Variance
6 vs. 5	2.66	1	10.82	n.s.	Method Covariance

3.10 Predictive Validity Analyses: Correlation of the SJT of Personality with Job Performance

In order to examine the criterion-related validity of the SJT developed in this study, the relationships between the SJT scores and job performance dimensions were examined. These analyses were conducted with the original version of the SJT since higher number of items might have some advantages. Job performance data

used was part of the official performance evaluation system and was obtained from the personnel files of the Human Resources Department. Four performance dimensions were selected for the validation purposes; *self development*, *innovation*, *teamwork*, and *team leadership*. According to the evaluation system of the organization, not all the employees are evaluated in all criteria. Criteria used for a given employee's evaluation depended on the job title, department, and experience. As a result, the number of participants in each performance dimension varied.

The results of correlational analyses are presented in Table 3.10. As can be seen from the table, correlation coefficients for the SJT and performance dimensions were ranged between $-.11$ and $.36$. Though most of them were close to zero, negative correlations were obtained for some relationships. Emotional Stability factor had the highest correlations with performance dimensions ($r_s = .06, .03, .32, .36$, for Self Development, Teamwork, Innovation, Team Leadership, respectively). Conscientiousness had relatively low correlations ($r_s = .16, -.11, .06, .11$ for Self Development, Teamwork, Innovation, Leadership, respectively). The other three dimensions had relatively lower correlations with performance dimensions. Among the correlations of other personality factors, the correlation between Extraversion and Teamwork ($r = .16$) and the correlation between Agreeableness and Self Development ($r = .10$) were in a positive trend.

Table 3.10 *Correlation between Performance Dimensions and the SJT Factors*

	Self Development	N	Teamwork	N	Innovation	N	Team Leadership	N
SJT Openness to Experience	-.08	178	-.08	69	-.08	56	-.09	31
SJT Conscientiousness	.16*	178	-.11	69	.06	56	.11	31
SJT Extraversion	-.04	178	.16	69	.07	56	-.01	31
SJT Agreeableness	.10	178	-.18	69	-.09	56	-.06	31
SJT Emotional Stability	.06	178	.03	69	.32*	56	.36*	31
NEO-PI-R Openness to Experience	.22*	83	.16	28	.28	31	.18	21
NEO-PI-R Conscientiousness	.08	83	-.49**	28	-.16	31	.16	21
NEO-PI-R Extraversion	.10	83	.04	28	.04	31	-.10	21
NEO-PI-R Agreeableness	.04	83	.16	28	.00	31	.34	21
NEO-PI-R Neuroticism	-.05	83	-.34	28	-.06	31	.20	21
BFI Openness to Experience	-.04	178	-.10	69	.03	56	-.13	31
BFI Conscientiousness	.03	178	-.12	69	.06	56	.02	31
BFI Extraversion	-.09	178	.03	69	.06	56	-.20	31
BFI Agreeableness	.00	178	-.25*	69	-.14	56	-.21	31
BFI Neuroticism	.08	178	.17	69	.00	56	.20	31

Note. * $p < .05$, ** $p < .001$.

CHAPTER 4

DISCUSSION

4.1 Overview

The aim of the study was to develop and validate an SJT specifically designed to assess the Big Five personality factors, i.e. the SJT of personality. The study incorporated both the development and validation procedures. The hypotheses were related to reliability, convergent validity and divergent validity of the developed SJT.

Results of reliability analyses (both internal consistency and test-retest estimates) were acceptable only for some but not all dimensions of personality. In terms of convergent validity, divergent validity and method variance, different analyses yielded slightly different results. However, in general, the findings provided support for the validity of the SJT of personality. In the following sections, first, results are going to be evaluated with respect to the hypotheses of the study. This discussion is followed by sections on contributions and practical implications of the study. Finally, limitations and suggestions for future research are presented.

4.2 Discussion of the Results Concerning Reliability and Validity

For reliability analysis, both internal consistency and test-retest reliability estimates were calculated. Cronbach's alphas for the developed SJT factors were in general not satisfactory, ranging from .17 to .52. These estimates were lower than the alphas found for the BFI scales (from .51 to .77) in the present study. This finding was not very unexpected given the previous studies in the literature that also failed to report satisfactory internal consistency reliability estimates for the SJTs (e.g., Meijer et al., 2010; O'Connell, 2007; Oswald et al., 2004). As a final note concerning the internal consistency of the SJT, lower estimates found in the present study may also be

partially related to the small number of items in each factor since the number of items has direct effect on internal consistency reliability (Cook, 2009).

Test-retest reliability estimates were obtained from a subsample of ($N = 59$) the original sample approximately two months after the first administration of the test. These estimates ranged from .22 to .75. Test-retest reliabilities for Openness to Experience (.72) and Extraversion (.75) were quite satisfactory whereas for Agreeableness (.22) it was the lowest. For both Emotional Stability (.48) and Conscientiousness (.43) the reliabilities were low but perhaps marginally acceptable. Overall, the analyses provided support for the reliability of at least Extraversion and Openness to Experience scales of the SJT, yielding only partial support for the first hypothesis.

The second hypothesis was related to the validity of the SJT of personality. Both construct validity and criterion-related validity of the test were examined. In examining the construct validity both multitrait-multimethod matrix procedure developed by Campbell and Fiske (1959) and the confirmatory factor analysis approach, in which hierarchically nested models are examined to identify trait and method effects, were employed. For the Campbell and Fiske's method, two personality measures, BFI and NEO-PI-R were used. Multitrait-multimethod matrix analysis suggested that, in general, the shortened version of the SJT of personality have acceptable levels of convergent and divergent validity. Considering the relationship between the SJT and NEO-PI-R dimensions, all five factors appeared as having convergent validity with Openness to Experience, Emotional Stability, and Extraversion factors higher convergent validity levels. Furthermore, Emotional Stability, Openness to Experience, and Extraversion factors had evidence of divergent validity as well.

On the other hand, when the relationship between the SJT and BFI was examined Openness to Experience and Conscientiousness factors appeared to have convergent validity whereas no factor displayed evidence of divergent validity.

Hence, results in general indicated that Openness to Experience, Extraversion, and Emotional Stability factors assessed by SJT methodology had satisfactory convergent and divergent validity levels.

The results of the analysis of hierarchically nested models supported the hypotheses regarding convergent and divergent validity of the SJT of personality. The analysis comparing the trait method model with correlated factors and the method model showed that the test has convergent validity. The analysis comparing the trait method model with correlated factors with general trait model showed that the test had divergent validity. However, there was a significant amount of variance attributable to method factor, which was showed by method variance analysis conducted by comparing the trait method model with orthogonal factors and trait model.

As an exploratory analysis, hypothesis regarding divergent validity was also tested with a general cognitive ability test (i.e. the PMRT). Literature points that as a method of measurement SJTs have a “judgment” component which is related to cognitive ability (McDaniel et al., 2001). In addition, studies often find relationship between SJT scores and cognitive ability scores (Clevenger et al., 2001; McDaniel & Nguyen, 2001). However, within the current study, the SJT was specifically developed to assess personality, a construct that conceptually has no relationship with cognitive ability. Hence, it was expected that personality factor scores would not have significant relationships with the PMRT score. Correlational analyses showed that none of the SJT dimensions had significant correlations with the PMRT yielding further evidence for divergent validity of the SJT in general.

When the first two hypotheses considered together, three factors of the SJT of personality appeared to have satisfactory reliability coefficients and construct validity evidence. These factors are Openness to Experience and Extraversion, and partially, Emotional Stability. Conscientiousness, on the other hand, appeared as a relatively weaker factor in terms of psychometric properties while Agreeableness was a problematic factor in terms of almost all psychometric indices. The observed differences among the SJT factors may have several methodological and/or conceptual explanations. First of all,

Openness to Experience and Extraversion that appeared as psychometrically sound factors within the current study, are the factors that have lowest relationship with social desirability (Ones, Viswesvaran, & Reiss, 1996). In addition, in the current study the SJT items aiming to assess these factors seem relatively less job-related on the surface. That is, in addition to purely job related and critical incident derived items, such as dealing with new situations in the workplace, there were items related to social life, such as dealing with food choice in an unknown country. Hence, it is possible to expect that participants engaged in lower levels of socially desirability in responding to Openness to Experience and Extraversion items. Genuine responding may have created a psychometric advantage for these two factors. On the other hand, Conscientiousness, Agreeableness, and Emotional Stability are known to have higher correlations with social desirability (Ones, Viswesvaran, & Reiss, 1996). In addition, the SJT items that aimed to assess these factors were relatively more job-related, which could have motivated participants to respond in a socially desirable manner.

Personality assessment through the SJT method is not frequent in the literature. Thus, more research in the area is needed to draw firmer conclusions as to which personality factors are more appropriate to be assessed via SJT. However, the findings of the present study suggest that SJT methodology may be differentially effective in measuring certain traits.

As a note, it is important to mention that, unexpectedly, not all the relationships among factors of BFI and factors of NEO-PI-R were in the expected and satisfactory levels. As an assessment tool NEO-PI-R has strengths over BFI such as higher number of items and higher reliability estimates.

The last hypothesis was about the relationship between the scores of the SJT of personality and job performance. To begin with, congruent with the purposes of the current study, four performance dimensions were selected for this criterion-validity examination. The organization for which the SJT was developed has a comprehensive performance evaluation system that includes both objective goals and relatively subjective work related competencies.

Among them, not all the competencies were directly related to personality factors (e.g. client orientation, system and quality enhancement), thus four of the competency evaluations were selected for the current study by the researcher herself. In addition, as explained above, not all the employees are evaluated for all the criteria. Considering the different numbers of participants who were administered NEO-PI-R and the SJT, there was also a mismatch for the same performance criteria in terms of number of participants. This mismatched numbers resulted in different numbers of participants in each cell with some of them being too low to conduct a reliable correlation analysis. The selected performance criteria for the validation study were self development, innovation, teamwork, and team leadership. These criteria and their detailed descriptions are provided in the competency dictionary of the organization for both employees and raters. Summary of these definitions are as follows; *self development*: being aware of one's own strengths and weaknesses, desire and effort for continuous learning and self development; *innovation*: developing and implementing ideas that will enhance the productivity of the organization, evaluating the situations with different and questioning what to do and how to do differently; *teamwork*: effort and desire to collaborate with others to attain a shared goal; and *team leadership*: ability to coordinate, motivate and direct the team members and to create team spirit and integration.

The correlations among personality factors assessed with the SJT and performance dimensions ranged between -.18 (Agreeableness-Teamwork) and .36 (Emotional Stability-Team Leadership). As can be seen in Table 3.10 some negative correlations were obtained in the correlation analysis. Though, infrequent in the personality and job performance literature, Barrick, Mount and Judge (2001) reported negative lower 90% credibility values for all personality factors, except for Conscientiousness, in their mega meta-analysis. The findings show that negative correlations between personality and job performance, although infrequent, in fact, exist. Conscientiousness factor, which appear as the strongest predictor of job performance across various criteria in previous findings, did not emerge as a strong predictor in the current study. Conscientiousness correlated significantly with only one performance

dimension, which is self development. Openness to Experience and Extraversion factors, though appeared as relatively strong in terms of reliability and validity, were not significantly related with job performance dimensions either. Emotional Stability was significantly related with two dimensions of job performance, innovation and team leadership, consistent with results reported by Salgado (1997), in which Emotional Stability was found as a valid predictor of job performance.

When the correlations among the performance dimensions themselves were examined it was noted that although some of these coefficients were significant and positive (e.g., $r_{\text{Emotional Stability - Innovation}} = .32$; $r_{\text{Conscientiousness - Self Development}} = .16$) they were below the expected levels reported in the relevant literature. The reason of low correlations may be rooted in the nature of the jobs, organizational context, or the constructs themselves. These low correlations made it difficult to treat these competencies as parts/components of an overall job performance construct in the present study. Hence, the quality of the criterion measurement may have contributed to the observed low correlations between the SJT dimensions and performance dimensions in this study.

It is important to note that data quality in such criterion-related validity studies need to be examined from both predictor and criterion perspectives. In terms of predictors, the data in the present study were gathered from current employees of the organization. It is clear that current employees and candidates differ in terms of test taking motivations (Cook, 2009). In terms of criterion, one must rely on the evaluation system and judgments of raters within the organization.

The observed correlations between the SJT factors and performance indices must also be examined in the light of broader personality and job performance research. In terms of performance outcomes, the current study employed supervisor rated competencies as the criterion. Studies in general suggest that personality is more likely to predictive of citizenship behaviors and attitudes than task performance (Berry, Ones, & Sackett, 2007; Borman et. al 2001). Hence, it is plausible that the observed correlations could have

been much higher had more citizenship-behavior or citizenship-attitude related performance dimensions been included.

4.3 Strengths and Contributions of the Study

The current study is believed to have a number of contributions to the related literature. Current SJT literature advises us to use this type of tests as measurement tools to assess predetermined constructs, and accordingly, such studies have accumulated considerably in recent years (McDaniel & Whetzel, 2006). However, there are still relatively few studies in which the SJT methodology is used to assess specific personality related characteristics (e.g., Christian, Edwards, & Bradley, 2010). The current study has potential to contribute to this literature by presenting a comprehensive attempt to measure the Big Five dimensions using the SJT approach.

This study is also believed to represent a methodologically and theoretically sound attempt to develop an SJT of personality. To start with, the three-step approach to test development suggested by McDaniel and colleagues (1990) was rigorously followed in the development of the SJT of personality. Also, critical incidents technique was used in item and response option development. The development of the scoring key, experts representing both the job context (i.e., mostly supervisors) and the theoretical background (i.e., clinical psychologists) were involved. Furthermore, in developing the items, response options, and the scoring key, the theoretical basis set by McCrea and Costa (1992) was used as the overarching frame of reference. In addition, characteristics of the test items were formed and edited according to guidance suggested by the prominent researchers in the area (e.g., McDaniel & Nguyen, 2001), in terms of item length, instruction types, complexity and fidelity. All of these are believed to contribute to the thoroughness of the SJT developed in this study.

Originally a 60-item SJT was developed. Following the initial analyses, however, a shorter version of the test was formed by eliminating nearly half of the items. This shorter version had higher correlations with NEO-PI-R and BFI

since items were selected based on their correlations with the relevant NEO-PI-R dimensions. The shortened version of the SJT had higher convergent and divergent validity coefficients. This construct validity advantage suggests that the shorter version can be treated as more of a “personality test.” On the other hand, criterion-related validity analyses were more satisfactory with the longer version of the test than they were with the short version of the test. Criterion-related validity advantage of the original the SJT indicates that it can be treated as more of a “selection test.” Future research is needed to further clarify these assertions.

In addition to the contributions described above, the SJT of personality is also believed to contribute to the existing personality assessment in the personnel selection literature because of its two characteristics: high contextualization and low transparency. The SJT developed in this study is a highly contextualized one. The items were derived from critical incidents, representing the situations/problems that are critical to the jobs in question within the organization, reflecting the organization’s own dynamics. Assessing the personality of an applicant via a contextualized test to the work situations is suggested to lead to better results than a general personality assessment (Cook, 2009; Hunthausen, Truxillo, Bauer, & Hammer, 2003; Robie, Schmit, Ryan, & Zickar, 2000).

Low transparency of the items is an additional strength of the SJTs in general. Typical SJT items are likely to reduce social desirability effects observed in traditional personality inventories. The degree of transparency was not directly assessed in the current study. However, when compared to BFI and NEO-PI-R items, it was probably more difficult to guess the underlying personality factor for the SJT items. However, future research is needed to empirically test this assertion.

Finally, this study is believed to have potential to contribute to the local literature in personnel selection in Turkey as well. It is hoped that this study will lead the way for the development of other context specific SJTs tapping critical, job-related attributes so that a more comprehensive evaluation of the

utility of the SJT methodology for selection purposes could be made in the Turkish context.

4.4 Practical Implications

This study has also some implications for the personnel selection practices. Personality inventories are widely used in personnel selection in many countries including Turkey (Piotrowski & Armstrong, 2006; Sözer, 2004). In addition, the inventories used in personality assessment are generally standard all-purpose measures developed and used for various purposes. SJT method, on the other hand, is infrequently employed in selection practices despite convincing empirical evidence favoring them. Assessing personality via SJTs is expected to have a two-fold advantage over other tests. First, SJT's are tailor-made tests by nature. A tailor made test designed for an organization's selection battery is expected to function better than a generic test developed for other purposes rather than personnel selection. Secondly, SJT's have predictive power with respect to their development procedure. Thus, it is possible to increase predictive ability of a selection battery by adding an SJT of personality.

4.5 Limitations of the Study and Suggestions for Future Research

The current study has several limitations regarding number of participants, design of the study and methods employed. In this section, limitations are discussed followed by suggestions for future research.

An important limitation of the present study is related to the number of participants. Consistent with the aim of the study, which was to develop and validate an SJT of personality, participants were administered two other personality inventories: BFI and NEO-PI-R. Although the number of participants who took SJT and BFI was satisfactory to conduct most of the analyses, analyses involving the comparison of the SJT and the NEO-PI-R could not be made under ideal conditions because NEO-PI-R was administered to only a subset of participants ($N = 95$). Observed lower reliability estimates of BFI and the small sample size receiving NEO-PI-R

constituted an important limitation of the current study. Future studies should include larger number of participants receiving all measures. In addition, number of participants with performance data was relatively low, ranging from 31 to 178. Again, small sample size for job performance may have negatively affected the correlations obtained.

Possibly another major limitation of the study is related to the design of the study. The current study evaluated the criterion-related validity, by using a concurrent validity strategy. That is, already working, veteran employees rather than job applicants were administered the tests and the performance data were obtained concurrently from the files. However, ideally the tests should have been applied to real candidates (not to veteran employees) and predictive validity assessment should have been conducted later when performance data of the hired individuals would become available. Future research is needed to assess the predictive validity of the developed SJT.

In terms of predictive ability of the SJT of personality, only available job performance data were used. It is important to include other performance indices in criterion-related validity analyses of the SJT of personality in the future studies.

Assessment of validity of the current test was conducted with two methods; multitrait-multimethod matrix suggested by Campbell and Fiske (1985) and hierarchically nested models suggested by Widaman (1985). Campbell and Fiske propose several factors to take into consideration in evaluating correlation matrix formed according to multitrait-multimethod procedure. First, reliability estimates of two measures which give input for validity diagonal are important, since a low reliability of a test might exaggerate the method variance of the other test. The two scales of current study, BFI and the SJT did not produce compatible reliability estimates due to lack of satisfactory reliability estimates of the SJT. However, it is important to note here that test re-test reliability estimates were found as slightly higher than inter item reliability estimates. Second, having an adequate sample size is crucial since limited sample size for one or more traits would depress reliability coefficients. For the present study, though sample size for each trait

in a given test was equal, sample size for NEO-PI-R was smaller than the other tests.

It is also important to acknowledge that the method proposed by Campbell and Fiske (1985) to assess convergent and discriminant validity is subject to many criticism (Widaman, 1985). Since the correlations lack independence from each other, testing the statistical significance of the overall pattern is not possible. Amount of variance for measures are not estimated precisely. Another criticism, which was also pointed by Campbell and Fiske, is about the differences in reliability estimates of the tests used. It is stated that difference among reliability levels will result in distorted correlations among measures (Kenny & Kashy, 1992; Widaman, 1985). In response to these criticisms, alternative procedures are suggested such as the procedure involving testing hierarchically nested models (Widaman, 1985). This is why in the current study, in addition to Campbell and Fiske's method, hierarchically nested models method was also used.

Although there are two additional personality assessments other than the SJT, the analysis of validity with hierarchically nested models was conducted with BFI only. In addition, sample size was not high enough to conduct confirmatory factor analysis with all the items of BFI and SJT. Hence, only four items from each factor of these measures were selected based on intercorrelations. In the future, the analyses may be replicated by including all the items of SJT and BFI as well as by including NEO-PI-R dimensions or other personality assessments tools.

Finally, although it is believed that the test development procedure followed in this study was very rigorous, future studies may combine different techniques in test development, such as employing empirical based scoring techniques or deriving questions from job analytic information.

REFERENCES

- Bauer, T. N. & Truxillo, D. M. (2006). A theoretical basis for situational judgment tests. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgment Tests*. Mahwah, NJ: Erlbaum.
- Barrick, M. R., Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1-26.
- Barrick, M. R., Mount, M. K., Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?, *International Journal of Selection and Assessment*, 9-30.
- Becker, T. E. (2005). Development and validation of a situational judgment test of Employee integrity. *International Journal of Selection and Assessment*, 13, 225- 232.
- Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment test: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223-235.
- Bery, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis. *Journal of Applied Psychology*, 92, 410-424.
- Benet-Martinez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75, 729-750.
- Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment*, 9, 52-69.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2).
- Chan, D & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, 15(3), 233-254.

- Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *Handbook of personnel selection*. Oxford: Blackwell.
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta analysis of their criterion related validity. *Personnel Psychology*, 63, 83-117.
- Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example, *Personnel Psychology*, 51, 193- 208.
- Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 86(3), 410-417.
- Cook, M. (2009). *Personnel Selection: Adding value through people*. Wiley-Blackwell, Oxford, UK.
- Gülgöz, S. (2002). Five factor model and NEO-PI-R in Turkey. In R. R. McCrae, and J. Allik (Eds.), *The five-factor model of personality across cultures*. New York: Kluwer Academic/Plenum Publishers.
- Hunthausen, J. M., Truxillo, D.M., Bauer, T. N., & Hammer, L. B. (2003). A field study of frame of reference effects on personality test validity. *Journal of Applied Psychology*, 88, 545-551.
- Hooper, A. C., Cullen, M J., Sackett, P. R. (2006). Operational threats to use of SJT's: Faking, coaching, and retesting issues. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgment Tests*. Mahwah, NJ: Erlbaum.
- Jöreskog, K.G. & Sörbom, D. (2006). LISREL 8.80 for Windows [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc
- Kanning, U.P., Grewe, K., Hollenberg, S. & Hadouch, M. (2006). From the subjects' point of view – reactions to different types of situational judgment items". *European Journal of Psychological Assessment*, 22 (3), 168-76.
- Kenny, D. A., & Kashy, D. A. (1992) Analysis of multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165-172.
- Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology*, 69, 569- 573.

- Lievens, P., & Patterson, F. (2011). The validity and incremental validity of knowledge test, low fidelity simulation and high fidelity simulations for predicting job performance in advanced level high stakes selection. *Journal of Applied Psychology, 96*(5), 927-940.
- McCrae, R. R., & Costa, P. T., Jr. (1992). Discriminant validity of NEO-PI-R facet scales. *Educational and Psychological Measurement, 52*, 229-237.
- McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment test to predict job performance: A clarification of the literature. *Journal of the Applied Psychology, 86*(4), 730-740.
- McDaniel, M. A. & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-113.
- McDaniel, M. A. & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*, 515-525.
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Lee Grubb III, W. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91.
- McDaniel, M.A. & Whetzel, D.L. (2007). Situational judgment tests. In D.L. Whetzel & G. R. Wheaton (Eds.). *Applied measurement: Industrial psychology in human resources management*. Mahwah, NJ: Erlbaum. 235-257.
- Meijer, L. A. L., Born, M. P., Zielst, J. V., & van der Molen, T. (2010). Construct-driven development of a video-based situational judgment test for integrity. *European Psychologist, 15*, 229- 236.
- 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.

- Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, 91, 749-761.
- Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgment tests. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgment Tests*. Mahwah, NJ: Erlbaum.
- O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Lee Grubb III, W., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment*, 15, 19-29.
- Ones, D.S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81, 660, 679.
- Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessments for personnel selection. *Human Performance*, 11(2/3), 245-269.
- Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*. 89, 187- 207.
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, and R. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 224-239). New York: Guilford Press.
- Phillips, J. F. (1992). Predicting sales skills. *Journal of Business and Psychology*, 7(2), 151-160.
- Phillips, J. F. (1993). Predicting negotiation skills. *Journal of Business and Psychology*, 7(4), 403-411.
- Piotrowski, C., & Armstrong, T. (2006). Current recruitment and selection practices: A national survey of Fortune 1000 firms. *North American Journal of Psychology*, 8(3), 489-496.
- Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational

- judgment test. *International Journal of Selection and Assessment*, 11(1), 1-16.
- Robie C., Schmit, M. J., Ryan, A. M., & Zickar., M. J. (2000). Effects of item context specificity on the measurement equivalence of personality inventory. *Organizational Research*, 3, 348-365.
- Roth, P.L., Bobko, & P., McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, 58, 1009- 1037.
- Sackett, P. R., (2011). Integrating and prioritizing theoretical perspectives on applicant faking of personality measures. *Human Performance*, 24, 379-385.
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology*, 82-(1), 31-43.
- Salter, N. P., & Highhouse, S., (2009). Assessing managers' common sense using situational judgment tests. *Management Decision*, 47(3), 392-398.
- Scheiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review, *Journal of Educational Research*, 99(6), 323-337.
- Schmitt, N. and Chan, D. (2006), "Situational judgment tests: method or construct?", in Weekley, J. and Ployhart, R.E. (Eds), *Situational Judgment Tests*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 135-56.
- Sözer, S. (2004). *An evaluation of the current human resources management practices in the Turkish private sector*. Unpublished master's thesis. Middle East Technical University, Ankara.
- Sümer, N., & Sümer, H. C. (2002). Adaptation of BFI in a Turkish sample. Unpublished manuscript.
- Sümer, C. S., Er, N., Sümer, N., Ayvaşık, B., Mısırlısoy, M., Erol-Korkmaz, H. T., (2012). Development of personnel selection battery for blue collar workers. Unpublished manuscript. Department of Psychology, Middle East Technical University, Ankara, Turkey.
- Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist*, 11, 912-927.

- Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management*, 25(2), 207- 226.
- Trippe, D. M. (2002). *An evaluation of the construct validity of situational judgment tests*. Unpublished master's thesis. Virginia Polytechnic Institute and State University, Blacksburg.
- Viswesvaran, C., Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59(2), 197-210.
- Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, 52, 679-700.
- Weekley, J. A., & Jones, C. (1997). Video-based situational judgment testing. *Personnel Psychology*, 50(1), 25-49.
- Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationship with performance. *Human Performance*, 18, 81- 104.
- Weekley, J. A., & Ployhart, R. E.(2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgment Tests*. Mahwah, NJ: Erlbaum.
- Weekley, J. A., & Ployhart, R. E. (2006). Introduction to situational judgment testing. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgment Tests*. Mahwah, NJ: Erlbaum.
- Whetzel, D.L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21, 291-309.
- Whetzel D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resources Management Review*, 19, 188-202.
- Widaman, K.F., (1985). Hierarhically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9(1), 1-26.

APPENDICES

APPENDIX A: Critical Incident Questionnaire

Kritik Olaylar Anketi

Değerli Coşkunöz Holding Çalışanı,

Bu çalışma Coşkunöz Holding “Personel Seçme Sistemi Geliştirme Projesi” kapsamında ODTÜ Endüstri ve Örgüt Psikolojisi Yüksek Lisans programı öğrencisi Ayda Eriş tarafından yürütülen Durumsal Muhakeme Testi geliştirme çalışmasının bir parçasını oluşturmaktadır. Sizden istenen, aşağıda sunulan açıklamaları okuduktan sonra soruları cevaplamanızdır. Ankette sizin ya da başka birinin kimliğini belirleyecek herhangi bir bilgi istenmemektedir. Verdiğiniz bilgiler gizli tutulacak ve sadece test geliştirme çalışması kapsamında kullanılacaktır. Cevaplarınız bireysel olarak değerlendirilmeyecek, sadece problem alanların tespit edilmesi için kullanılacaktır. Katkılarınız için teşekkür ederiz.

İş Olayları ve Kişilik Özellikleri

Çalışanlar iş hayatında çeşitli olaylarla/durumlarla/problemlerle karşılaşır. Herhangi bir olayla/durumla/problemlerle karşılaştığında farklı kişiler farklı tepkiler gösterebilirler. Bazı davranışlar bir olayın ya da sorunun daha kolay çözülmesini sağlayabilir. Bazı davranışlar ise sorunun büyümesine ya da etkisiz bir şekilde çözülmesine yol açar.

Kişilik özelliklerimiz, hayatın her alanında olduğu gibi, iş yerinde de yaşadığımız olaylar ile çok yakından ilgilidir. Sahip olduğumuz belirli kişilik özellikleri bazen yaşanan olayların olumlu gelişmesine katkıda bulunurken bazen de olayların olumlu gelişmesini engelleyebilirler.

Aşağıda kişileri birbirinden farklı kılan kişilik özelliklerinden bazıları örnek olarak verilmiştir.

Dışadönüklük: Dışadönüklüğü yüksek kişiler aktif, genelde insanlarla birlikte olmayı seven, sıcak ve sosyal kişilerdir. Konuşmayı başlatan taraftırlar.

Duygusal Denge: Duygusal dengesi yüksek kişiler, güvenli, stresli durumlarda bile rahat, kolay sinirlenmeyen kişilerdir.

Uyumluluk: Uyumluluğu yüksek kişiler iyi huylu, geçimli, iş birliğine açık, çatışmaları engelleyen kişilerdir. Yardımcı olmaktan keyif alırlar.

Sorumluluk Bilinci: Sorumluluk bilinci yüksek kişiler sorumlu ve düzenlidir. Standartları her zaman yüksektir ve hedeflerine ulaşmak için çok çalışırlar.

Gelişime Açıklık: Deneyime açıklığı yüksek kişiler yeni tecrübelerle açıktır, ilgi alanları çok geniştir, hayal güçleri kuvvetlidir. İşleri yapmanın yeni ve değişik yollarını araştırırlar.

Lütfen aşağıdaki bölümleri doldurunuz.

Demografik Bilgiler

Yaşınız: _____ Cinsiyetiniz: Kadın Erkek

Coşkunöz Holding' te çalışma süreniz: _____

Toplam iş deneyiminiz: _____

Mesleğiniz/İşiniz : _____

Göreviniz/Pozisyonunuz: _____

Bağlı olduğunuz birim: _____

Soru 1:

İş yerinde yaşanan olumlu ya da olumsuz bir çok olayda olaya dahil olan tarafların "kişilik özellikleri" durumun ortaya çıkmasında ya da nasıl sonuçlandığında kritik bir rol oynar. Kişilik özellikleri bazen yardımcı, bazense engelleyici faktör olabilir.

Çalışma ortamında son zamanlarda yaşadığınız ya da tanık olduğunuz ve taraflardan birinin kişilik özelliğinin son derece olumlu bir rol oynadığı bir olayı/durumu düşününüz.

Bu olay/durum ne idi? Lütfen ayrıntılı bir şekilde yazınız.

Sonuç ne oldu?

<p>Sizece bu kiři/kiřiler ne yapmıř olsa durum daha <u>olumsuz</u> sonulanırdı?</p>
<p>Sizece bu kiřinin/kiřilerin hangi kiřilik zellięi bu durumda kritik bir rol oynadı?</p>
<p style="text-align: center;">Soru 2</p>
<p>alıřma ortamında son zamanlarda yařadığınız ya da tanık olduęunuz ve taraflardan birinin kiřilik zellięinin <u>son derece olumsuz</u> bir rol oynadıęı bir olayı/durumu dřününüz.</p>
<p>Bu olay/durum ne idi? Ltfen ayrıntılı bir řekilde yazınız.</p>

Sonuç ne oldu?

Sizce bu kiři/kiřiler ne yapmıř olsa durum daha olumlu sonuçlanırdı?

Sizce bu kiřinin/kiřilerin hangi kiřilik özellięi bu durumda kritik bir rol oynadı?

*Eklemek istedięiniz olay(lar) var ise arka sayfayı kullanabilir ya da yeni bir anket isteyebilirsiniz.

APPENDIX B: Example Item of SJT

Puan		Amiri olduğunuz birimde kullanılan raporlama formatında ekip olarak bazı zorluklar yaşıyorsunuz. Halen kullanılmakta olan format hatasız olmasına rağmen çoğu zaman işlerin gereğinden fazla uzamasına neden oluyor. Biriminizde geçen hafta işe başlayan bir çalışan yeni bir yöntem önerdi. Bu yöntem işlerinizi kısaltabilecek olsa da nasıl sonuç vereceğinden emin değilsiniz. Böyle bir durumda ne yaparsınız?
3	a	Çalışanımın hevesini kırmamak için bu fikri incelemesi için daha tecrübeli bir çalışana yönlendiririm.
4	b	Çalışanımdan hazırlık yaparak bu yöntemi bana daha detaylı anlatmasını isterim.
2	c	İşleri tam anlamıyla kavrayabilmesi için daha zamana ihtiyacı olduğunu düşünürüm, fikri için teşekkür ederim.
1	d	Hatasız işleyen bir format varken değiştirme ihtiyacı duymam.
5	e	Fikri uygun prosedürlerle denemeye alırım.

APPENDIX C: Screenshot of computerized version of SJT

1	<p>Aynı ofisi paylaştığınız çalışma arkadaşlarınızdan biri bir müşteri ile ofiste görüşme yapmakta. Konuşmalarına şahit olduğunuz kadarı ile arkadaşınız hazırlıksız olduğu için müşterinin bir sorusuna cevap veremedi. Bunun üzerine müşteri, arkadaşınıza sesini yükselterek çıkışıyor ve neredeyse azarlamaya başlıyor. Başlangıçta arkadaşınız elinden geldiğince alttan almaya çalışmış olsa da giderek onun da sinirlendiğini görüyorsunuz. Böyle bir durumda ne yaparsınız?</p>
<input type="radio"/>	Sorunun çözülebilmesi için amirimi bulup konuya müdahale etmesini isterim.
<input type="radio"/>	Konu benimle ilgili olmadığından karışmam.
<input type="radio"/>	Konuya dahil olup yardımcı olabileceğim bir şey olup olmadığını anlamaya çalışırım.
<input type="radio"/>	Arkadaşım benden yardım isteyene kadar sessizliğimi korurum.
<input type="radio"/>	Ortamı yumuşatmak için esprili sözlerle lafa girerim.

APPENDIX D: Items of Big Five Inventory (BFI)

"Aşağıda sizi tanımlayan ya da tanımlamayan bir takım özellikler sunulmaktadır. Lütfen aşağıda listelenen her bir özelliğin sizi ne ölçüde tanımladığını belirtiniz. Cevaplarınızı samimiyetle vermeniz önemlidir. Her bir özellik için "Kendimi biri olarak görüyorum." ifadesine ne derece katıldığınızı düşünerek değerlendiriniz. İşaretlemenizi ilgili kutucuğa dokunarak yapınız.

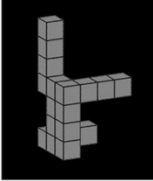
"

"Kendimi biri olarak görüyorum."

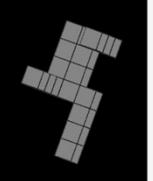
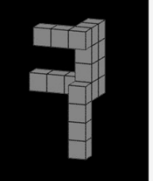
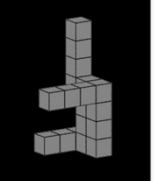
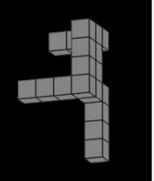
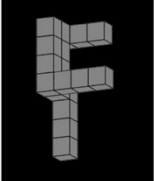
1. Konuşkan
2. Başkalarında hata arayan
3. İşini tam yapan
4. Bunalımlı, melankolik
5. Orjinal, yeni görüşler ortaya koyan
6. Ketum/ Sır saklayabilen
7. Yardımsever ve çıkarıcı olmayan
8. Biraz umursamaz
9. Rahat, stresle kolay baş eden
10. Çok değişik konuları merak eden
11. Enerji dolu
12. Başkalarıyla sürekli didişen
13. Güvenilir bir çalışan
14. Gergin olabilen
15. Maharetli, derin düşünen
16. Heyecan yaratabilen
17. Affedici bir yapıya sahip
18. Dağınık olma eğiliminde
19. Çok endişelenen
20. Hayal gücü yüksek
21. Sessiz bir yapıda
22. Genellikle başkalarına güvenen
23. Tembel olma eğiliminde olan

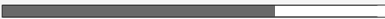
24. Duygusal olarak dengeli, kolayca keyfi kaçmayan
25. Keşfeden, icat eden
26. Atılgan bir kişiliğe sahip
27. Soğuk ve mesafeli olabilen
28. Görevi tamamlanıncaya kadar sebat edebilen
29. Dakikası dakikasına uymayan
30. Sanata ve estetik değerlere önem veren
31. Bazen utangaç ve çekingen olan
32. Hemen hemen herkese karşı saygılı ve nazik olan
33. İşleri verimli yapan
34. Gergin ortamlarda sakin kalabilen
35. Rutin işleri yapmayı tercih eden
36. Sosyal, girişken
37. Bazen başkalarına kaba davranabilen
38. Planlar yapan ve bunları takip eden
39. Kolayca sinirlenen
40. Düşünmeyi seven, fikirler geliştirebilen
41. Sanata ilgisi çok az olan
42. Başkalarıyla işbirliği yapmayı seven
43. kolaylıkla dikkati dağılan
44. Sanat, müzik ve edebiyatta çok bilgili

**APPENDIX E: Example Item of Personnel Multiple Reasoning Test
(Screenshot)**

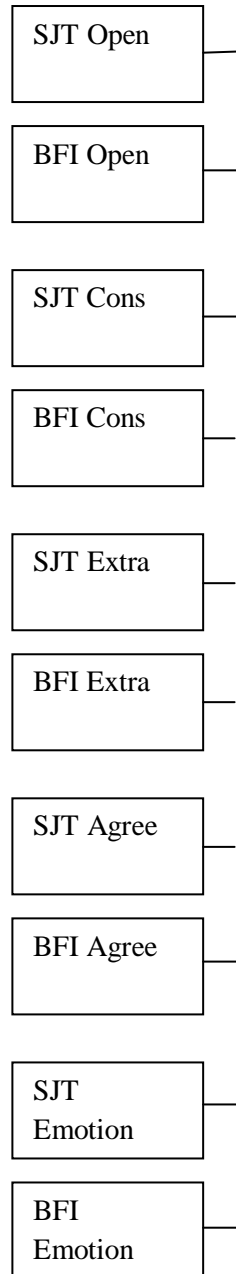


Aşağıdakilerden hangisi yukarıdaki şekil döndürüldüğünde veya şekle farklı bir noktadan bakıldığında görünen halidir?

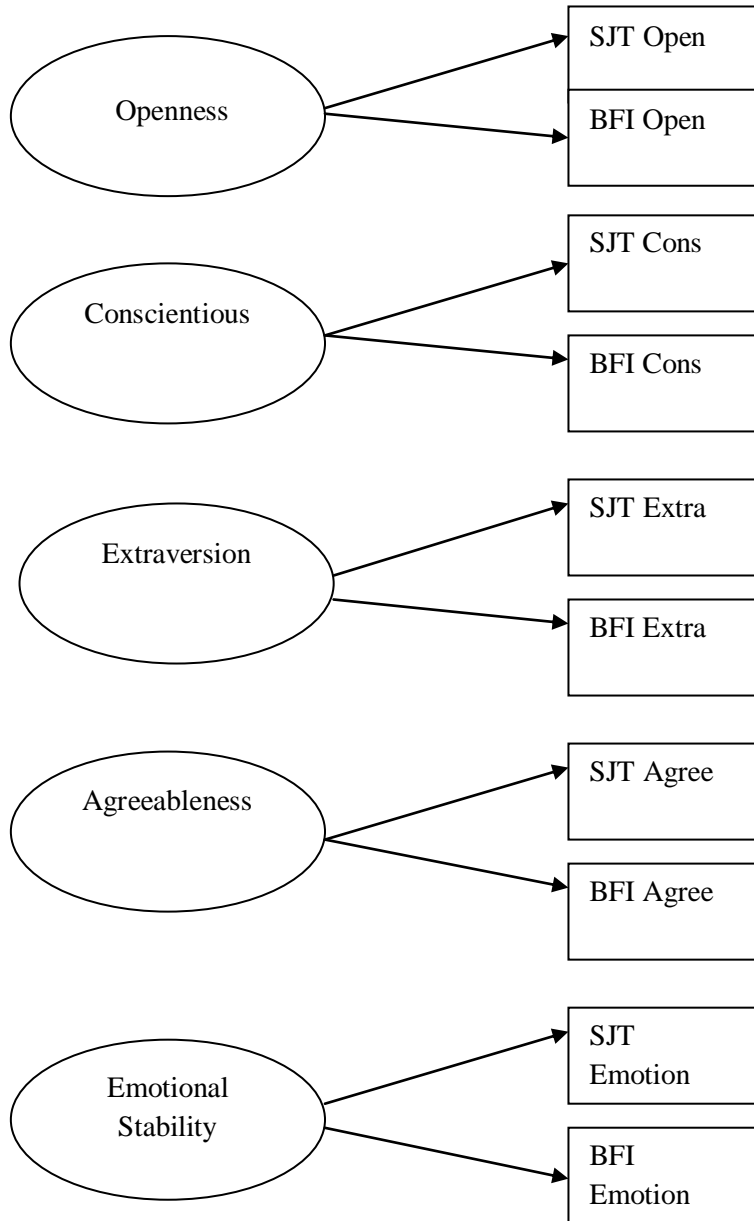




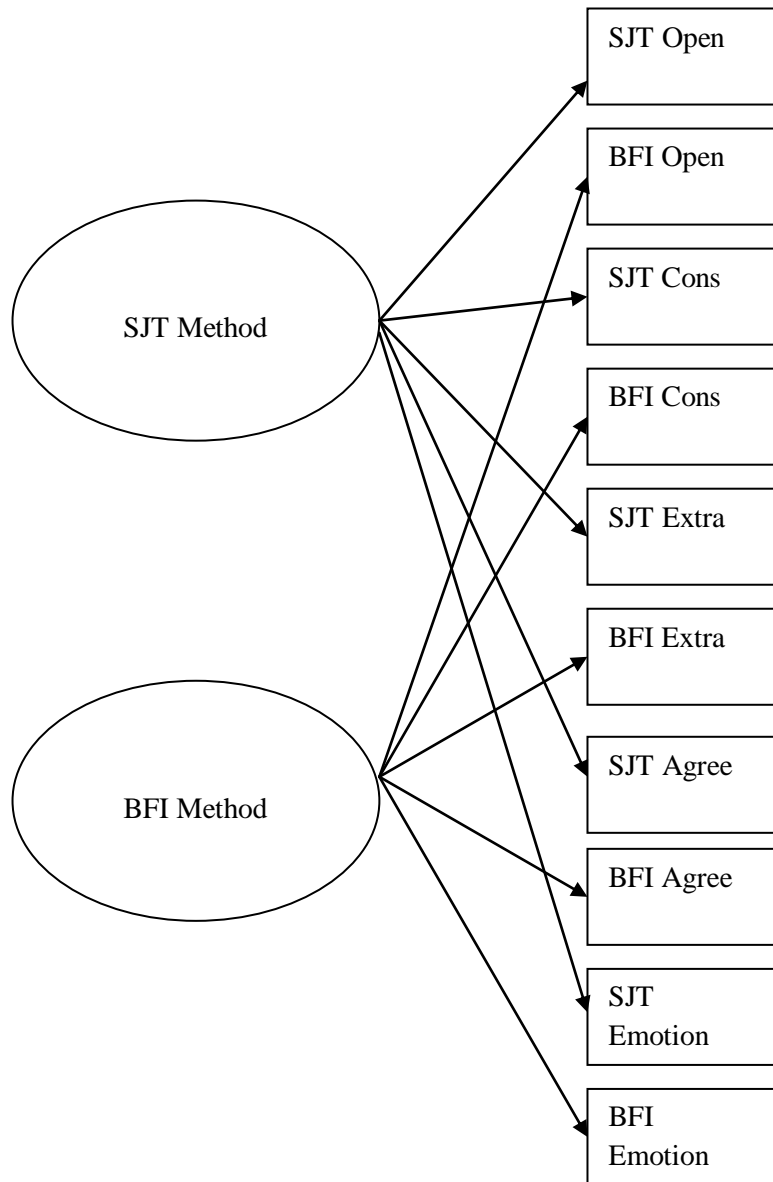
APPENDIX F: Figure of Model 1: Null Model



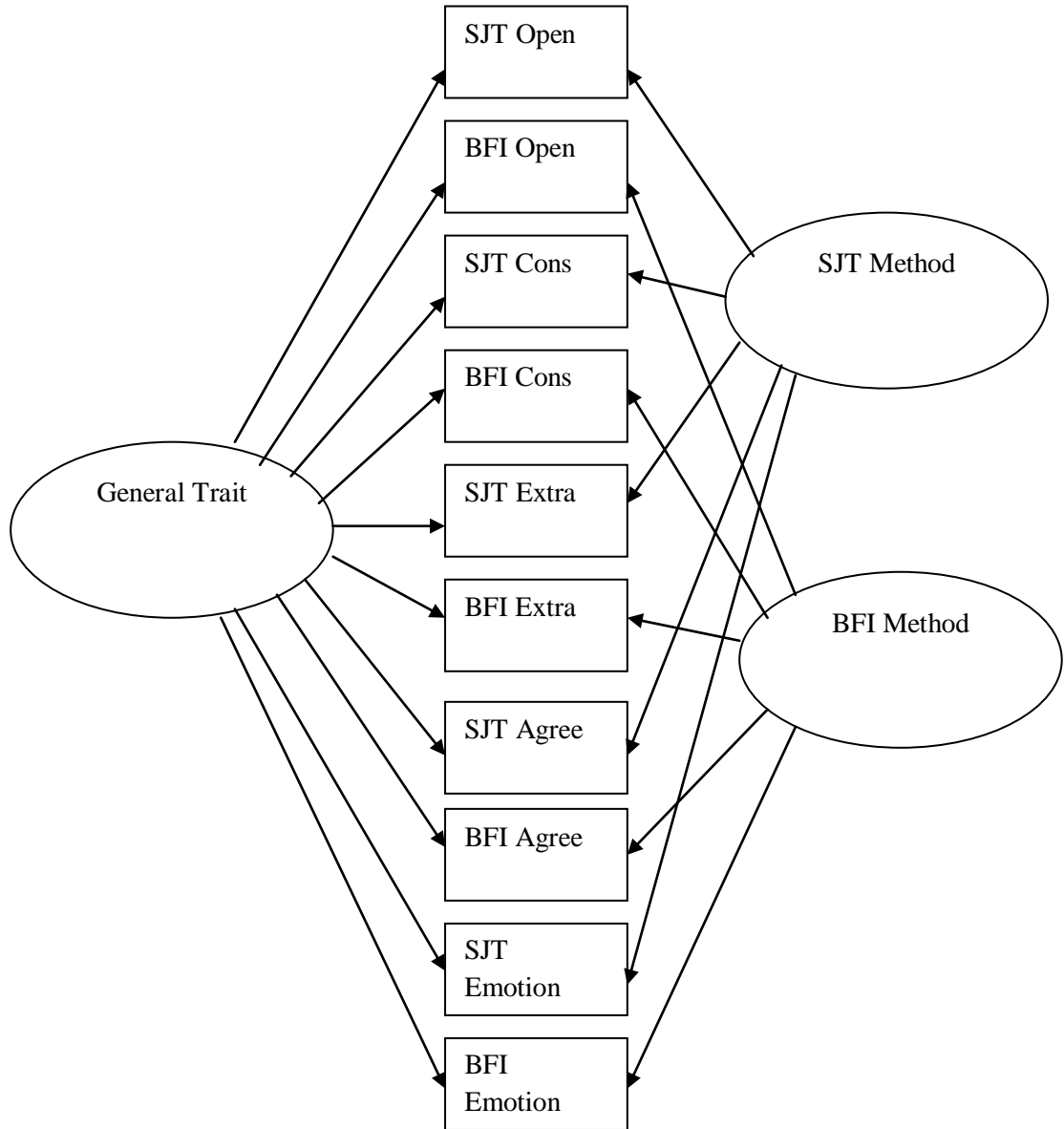
APPENDIX G: Figure of Model 2: Trait Model



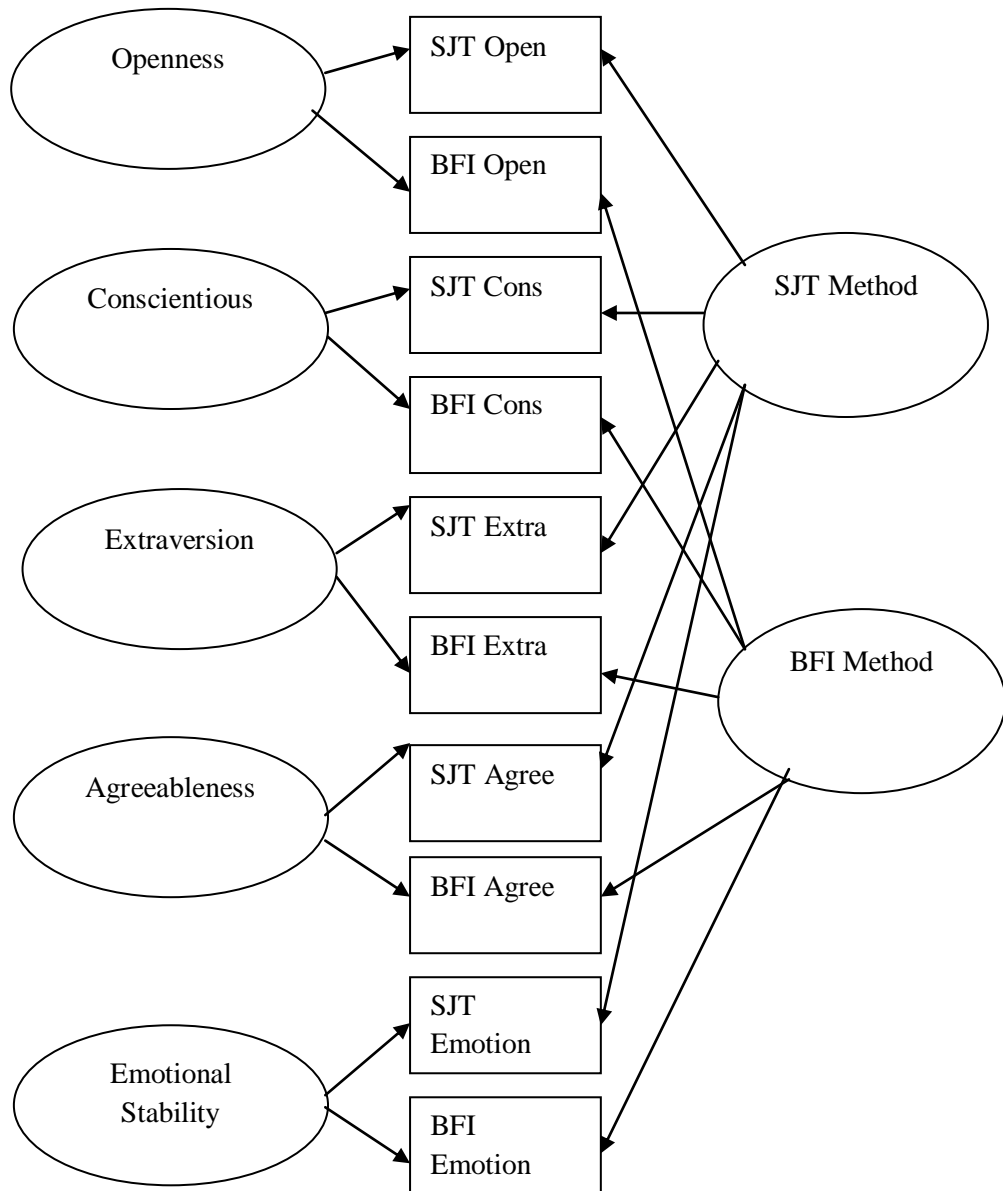
APPENDIX H: Figure of Model 3: Method Model



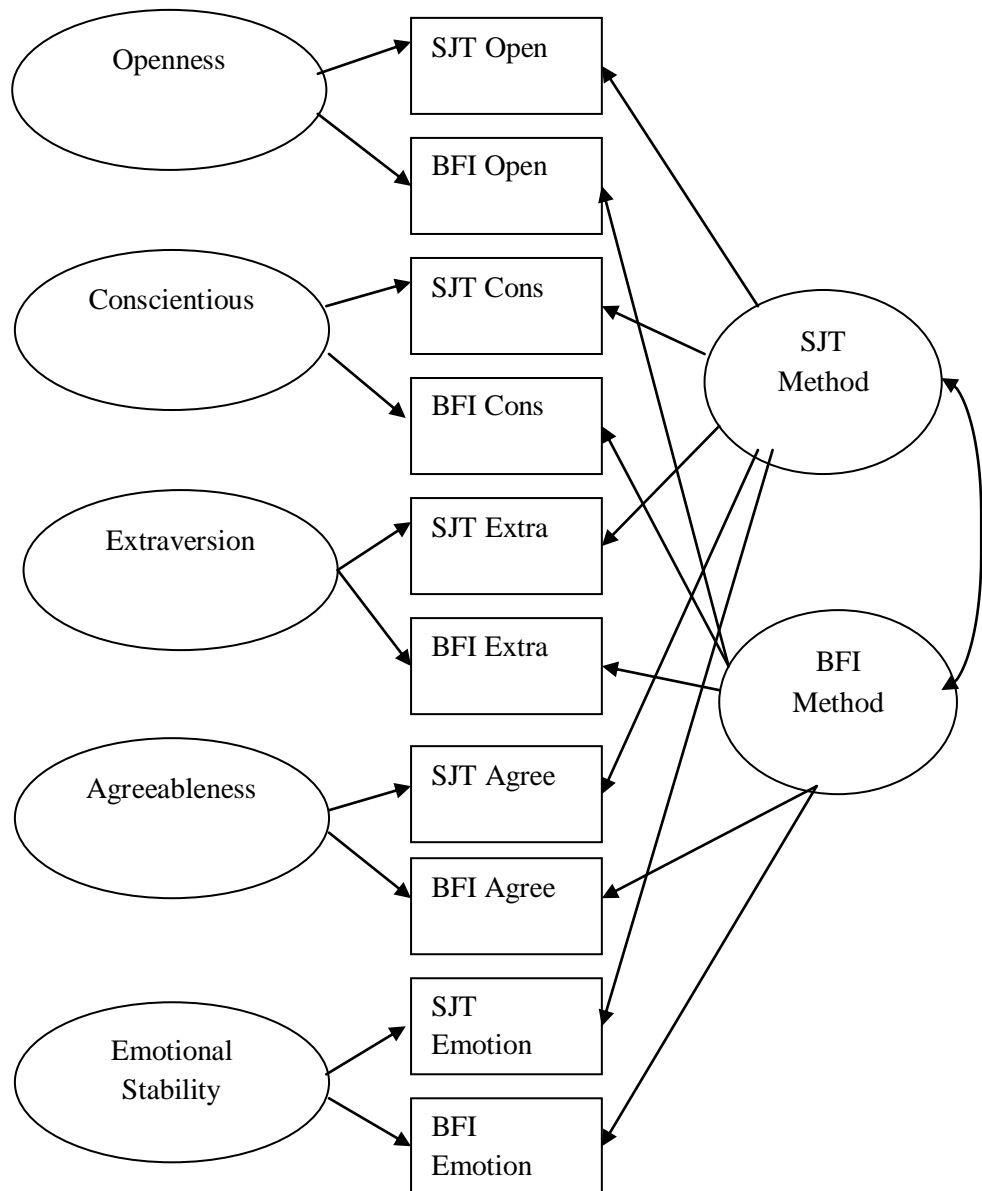
APPENDIX I: Figure of Model 4: General Trait Model



APPENDIX J: Figure of Model 5: Orthogonal Methods Model



APPENDIX K: Figure of Model 6: Correlated Methods Model



APPENDIX L: Tez Fotokopisi İzin Formu
TEZ FOTOKOPİSİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü

Sosyal Bilimler Enstitüsü

Uygulamalı Matematik Enstitüsü

Enformatik Enstitüsü

Deniz Bilimleri Enstitüsü

YAZARIN

Soyadı : Eriş

Adı : Ayda

Bölümü : Psikoloji

TEZİN ADI : Situational Judgment Tests In Assessing Specific Personality
Characteristics

TEZİN TÜRÜ : Yüksek Lisans

Doktora

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir.
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden kaynak gösterilmek şartıyla fotokopi alınabilir.
3. Tezimden bir bir (1) yıl süreyle fotokopi alınmaz.

TEZİN KÜTÜPHANEYE TESLİM TARİHİ: