5-2020

Different paths, same destination? Comparison between two approaches to developing situational judgment tests on cross cultural competency

Xiaowen Chen

Different paths, same destination? Comparison between two approaches to developing situational judgment tests on cross cultural competency

by

Xiaowen Chen

A dissertation submitted to the College of Psychology and Liberal Arts of
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Industrial/Organizational Psychology

Melbourne, Florida
May 2020

We the undersigned committee hereby approve the attached dissertation,
"Different paths, same destination? Comparison between two approaches to developing situational judgment tests on cross cultural competency."
by
Xiaowen Chen

Gary Burns, Ph.D,
Professor
Industrial/Organizational Psychology
Major Advisor

Lisa Steelman, Ph.D.
Professor and Dean
Industrial/Organizational Psychology
College of Psychology and Liberal Arts

Charles Bryant, D.B.A.
Assistant Professor
Business

Patrick Converse, Ph.D.
Professor
Industrial/Organizational Psychology

Abstract

Title: Different paths, same destination? Comparison between two approaches to developing situational judgment tests on cross cultural competency

Author: Xiaowen Chen

Advisor: Gary Burns, Ph. D.

This dissertation focuses on developing SJTs to measure an individual's cross-cultural competency, and comparing the two SJT development approaches in terms of development costs, reliability, validity, susceptibility to social desirability, and test-taker reactions. In the first phase, the two 3C SJTs were developed with the model-based approach and the SME-driven approach respectively. In the second phase, data were collected to examine the reliability and validity of the two SJTs. Both 3C SJTs demonstrated acceptable reliability (α_{SME} = .72; α_{model} =.70), and convergent to CQS (r_{SME} = .35, p < .01; r_{model} = . 24, p < .01). The SJTs psychometric properties were further examined in the third phase, wherein the SJTs displayed similar reliability and were convergent to CQS. Both SJTs predicted satisfaction with overseas life (β_{SME} = .24, p < .01; β_{model} = .18, p < .05) and sociocultural adaptability (β_{SME} = -.20, p < .05; β_{model} = -.21, p < .05), meanwhile, only having none or small correlation with satisfaction with general life (r_{SME} = .10, *n.s.* and r_{model} = .19, p < .05). The SME-driven SJT outperformed the model-based SJT and CQS in predicting the actual multicultural team performance that was rated by peers (β_{SME} = .26, p < .05; β_{model} = -.04, *n.s.*; β_{CQs} = .01, *n.s.*). The utility of the two SJT development approaches, implications, future research directions and limitations were discussed in the end.

Table of Contents

List of Figures

List of Tables

Acknowledgement

This dissertation becomes a reality with the generous support and help of many faculty members and graduate students in and out of I/O Psychology Program at Florida Institute of Technology. I would like to extend my sincere gratitude to them.

First, I would like to express my deepest appreciation to my committee chair, Dr. Gary Burns, for his kind offer when I got stuck in the middle of this research, for his supervision, and for his imparting his excellent statistic knowledge in this research.

Second, I would like to express my special gratitude to Dr. Richard Griffith, who guided me to seek for the research ideas, to formulate the research questions and to write the SJTs scenarios and response options.

I am grateful to faculty members, staff members and graduate students in Florida Tech, who voluntarily worked as SMEs for the SJTs development and/or helped me to collect the data for this research. They are Dr. Andrew Cudmore, Dr. Dzmitry Yuran, Dr. Jessica Wildman, Dr. Keith Gallagher, Dr. Kenneth Gibbs, Dr. Lisa Steelman, Dr. Pat Converse, Dr. William Gabrenya, Linda Khan, Mina Milosevic, Petra Brnova, Anthony Belluccia, Garret Kilmer, Jean-Paul Martes, Jesse Caylor, Julianna Fischer, Katherine Rau, Kyi Phyu Nyein, Lavanya Kumar, Mara Hesley, Ngoc Duong, Nicholas Moon, Sarah Almansour, Sherif al-Qallawi, Trevor Fry, Yuer Peng, etc.

Last but not the least, I am highly indebted to Dr. Charles Bryant, whose heroic help guaranteed that the high-quality data were collected in a timely manner.

Dedication

The dissertation is dedicated to my beloved parents, who have been my source of courage, strength and comfort to go through every hard time.

To Dr. Richard Griffith, who gave me the great opportunity to initiate and lead the Cross-Cultural Competence Training Program for Florida Institute of Technology. This is fundamental for my today's achievement.

To Dr. William Gabrenya for his retirement. He is a great advisor for my life and career, and my reliable and lovely friend.

Chapter 1
Introduction

In the past three decades the Situational Judgement Test (SJT) has regained popularity as a personnel assessment and selection tool, thanks to its criterion-related validity, face validity, moderate to small group differences, and relatively easy development (Ployhart & Weekley, 2006; Weekley, Ployhart, & Holtz, 2006; Whetzel & McDaniel, 2009). Since the 1890s, SJTs have been created not only to measure an individual's personal attributes that are associated with overall job performance but also to measure specific constructs like leadership (File, 1945; Garman & Johnson, 2006; Grant, 2009; Peus, Braun & Frey, 2013), practical intelligence (Sternberg, 2009; 2015), emotional intelligence (Libbrecht & Lievens, 2012; MacCann, Fogarty, Zeidner & Roberts, 2011; Sharma, Gangopadhyay, Austin & Mandal, 2013), integrity (Meijer, Born, Zielst, & Molen, 2010), personal initiative (Bledow & Frese, 2009), interpersonal skills (Golubovich, Seybert, Martin-Raugh, Naemi, Vega & Roberts, 2017; Lievens, 2013), and teamwork (Mumford, Van Iddekinge, Morgeson & Campion, 2008; Wang, MacCann, Zhuang, Liu & Roberts, 2009).

However, few attempts have been made to apply SJT methods to assessing cross-cultural competency (3C), an individual capability of functioning effectively in culturally diverse situations (Ang, Dyne, Ng, Templer, Tay & Chandrasekar, 2007; Chen 2017; Gabrenya, Moukarzel, Pomerance, Griffith & Deaton, 2012; Gertsen, 1990; Trejo, Richard, van Driel & McDonald, 2015). 3C has proven to be critical in the success of oversea missions and multicultural teamwork (Black & Gregersen, 1999; Johnson, Lenartowicz & Apud, 2006; Arthur & Bennett, 1995; Caligiuri, 2000; Shaffer, Harrison, Gregersen, Black & Ferzandi, 2006). Unfortunately, existing self-report 3C measures are beset with validity issues (Gabrenya & Chen, 2019; Gabrenya et al., 2012; Matsumuta & Hwang, 2013). Applying SJT methods in 3C measurement may produce a more valid while less controversial measure for 3C. Therefore, the first purpose of this dissertation is to develop situational judgment tests to measure an individual's cross-cultural competency.

One interesting fact in SJT research is that review and meta-analytic studies hold a high ratio in SJT published studies. More than 13 peer-review articles of review or meta-analysis have been published since 2001, while the number of the primary studies on SJT is relatively small. The meta-analytical findings rely on a small number of studies. McDaniel, Morgeson, Finnegan, Campion and Braverman (2001) were only able to trace six studies for their meta-analysis on SJT predictive validity, and Ployhart and Ehrhart (2003) only found eight studies for SJT test-retest reliability. This fact indicates that SJT research is still in the primary stage of describing and summarizing measurement properties and overt phenomena. As stated by Ployhart and MacKenzie (2011), to understand SJTs, researchers should go beyond meta-analysis methods. More theoretical and empirical studies are needed to explore the nature of SJTs, its underlying mechanism, and effective design strategies.

Notably, in the review or meta-analytic articles, the researchers spent pages to discuss the future directions of SJT research, and their claims for future research largely overlap in design methods as well as SJT's psychometric properties such as reliability, validity, and utility (Campion, Ployhart & MacKenzie, 2014; Lievens, Peeters & Schollaert, 2008; McDaniel, Hartman, Whetzel & Grubb III, 2007; Oostrom, De Soete & Lievens, 2015; Ployhart & MacKenzie, 2011; Weekley, Ployhart & Holtz, 2006; Whetzel & McDaniel, 2009; etc.). The high overlap between those review and meta-analytical articles published between in 2000s and in 2010s, to some degree, reveals the fact of slow development and the existing vacancy in SJT research. Many important research areas discussed since 2006 are still untouched. One important research blank is psychometrical benefits of SJTs developed by model-based versus SME-driven approaches. No empirical studies have compared the psychometrical properties of SJTs developed by these two approaches despite researchers' repeated calls that this research would be very important in SJT design and development (Chen, Fan, Zheng & Hack, 2016; Weekley et al., 2006). In consideration of the significance of comparing the model-based approach and the SME-driven approach in SJT research, I developed 3C SJTs with the two approaches and compared their psychometrical strengths and weaknesses, that is the second purpose of this dissertation.

Chapter 2 reviews the history of SJT development and current trends in SJT studies, presents the reasons for SJTs resurgence, summarizes the major psychometric properties and discusses the research gaps in SJT studies and the purposes of the current dissertation research. Chapter 3 presents the key components of SJT development and the existing methods in designing the components. The strengths and weaknesses of the methods are compared, and the methods utilized to designing the 3C SJTs are justified. Chapter 4 shifts the research focus from SJT to 3C research and measurement. 3C research is briefly reviewed, and a theoretically sound 3C model is presented to serve as the model of model-based 3C SJT development. Chapter 5 describes the procedures of developing the model-based 3C SJT and the SME-driven 3C SJT. Chapter 6 presents the research questions and hypotheses of the current dissertation studies. Chapter 7 describes the methodology used for the dissertation studies and the discussion of each study result. Chapter 8 is general discussion about the dissertation research, and implications, limitations and future research are presented in Chapter 9.

Chapter 2
Situational Judgement Test

As a measurement method, SJTs are a low-fidelity simulation that simulates decisions making during a series of situations happening at work or associated with constructs of interest. The test taker is required to respond to a list of response options following each scenario according to the instructions. SJTs have appeared in various forms like scenarios with open questions, situational behavior interviews, assessment center scenarios, and video-based scenarios; however, the typical SJT consists of written scenarios, response actions, and response instructions. An illustrative item is listed below.

> Alan helps Trudy, a peer he works with occasionally, with a difficult task. Trudy complains that Alan's work isn't very good, and Alan responds that Trudy should be grateful he is doing her a favor. They argue.
>
> > What action would be the most effective for Alan?
> >
> > (a) Apologize to Trudy
> >
> > (b) Stop helping Trudy and don't help her again
> >
> > (c) Try harder to help appropriately
> >
> > (d) Diffuse the argument by asking for advice.
> >
> > (From Libbrecht & Lievens, 2012, *p.* 441)

A SJT scenario is presented as a dilemma or a problem reflecting realistic situations of interest, and the solution to the dilemma or the problem requires a test taker to apply one's knowledge, skills, abilities, and experience. The test taker is instructed to choose the appropriate response options to the respective simulation. The response instruction usually falls into two categories: knowledge-based instruction and behavioral tendency instruction (McDaniel & Nguyen, 2001). The knowledge-based instruction requires the test taker to decide which action is most effective or should be conducted based on one's knowledge, while the behavioral tendency instruction asks about the test-takers most likely behavior when facing the given situation. The response options are plausible courses of action, targeting the range of judgment performance. The primary assumption of a SJT is that people's choice reveals their acquired knowledge, cognitive intention, or behavioral preferences. Their judgement performance, the decision on the courses of action, is believed as proximal causes of job performance or other job-related criteria (Chan & Schmidt, 2017; Whetzel et al., 2008).

2.1 SJT History

The use of SJTs can be tracked back to the 19th century (Whetzel & McDaniel, 2009). The earliest SJT had scenario descriptions with open-ended questions, and the test takers had to present their own solutions to the scenarios. It resembled the modern situational interview, and the test takers were asked about their actions to the situation like "*When a person has offended you, and comes to offer his apologies, what should you do?*" (from Benet Child Intelligence Scale, cited by Whetzel & McDaniel, 2009). The George Washington Social Intelligence Test is one of the first widely used SJTs, which was used to assess an individual's interpersonal skills. It appears as a set of scenarios with solutions in a multiple-choice format (Hunt, 1928; McDaniel et al., 2001; Moss, 1926).

The first surge of SJT development was during World War II when SJT techniques were applied to assess soldier's judgment ability. The military recruits were required to decide the most effective option among a list of reactions to the scenarios which were detailed descriptions of threats and challenges encountered by soldiers in realistic military situations. Those SJTs were successful in selecting competent soldiers as they not only assessed the recruits' experience, common sense, and general knowledge, but also served as a realistic job preview which discouraged those with romantic illusions of military career and those unfit for a military environment, thereby increasing retention rates (Lievens & De Soete, 2015; Northrop, 1989). Driven by the success of military SJTs in soldier selection in WWII, SJTs were widely used in personnel selection and performance assessment in the workplace. SJTs were used as a part of selection test batteries in large organizations for job promotion and for developing potential talent pipelines. For instance, Early Identification of Management Potential was used by the Standard Oil Company of New Jersey, and Test 905 used by U.S. Office of Personnel Management (McDaniel et al., 2001).

Modern SJTs are also applied to assess specific personal attributes. SJTs have been designed to measure an individual's supervising ability and leadership since the 1940s (Bruce & Learner, 1958; File, 1945; Garman et al., 2006; Grant, 2009; Kirkpatrick & Planty, 1960; Mowry, 1964; Oostram et al., 2012; Peus et al., 2013). SJTs were also designed to measure practical intelligence (Cardall, 1942: Sternberg, 1990; 2015). More recently, there is new interest in applying SJT to measuring other psychological constructs like emotional intelligence (Libbrecht & Lievens, 2012; MacCann et al., 2011; Sharma et al., 2013), integrity (Chen, 2009; Meijer, van der Sanden, Snijders, et al., 2010), social initial and interpersonal skills (Bledow & Frese, 2009; Golubovich et al., 2017; Lievens, 2013), and teamwork (Mumford et al., 2008; Wang et al., 2009). Although these SJTs are targeted at specific constructs, heterogeneity is still one of the typical characteristics prevalent in most construct-specific SJTs (Lievens et al., 2008; Ployhart & MacKenzie, 2011).

Outside the workplace, SJT methods are sometimes used in high-stake settings like educational selection and evaluation, especially medical fields, to assess students' or applicants' academic and practical performance. In the United Kingdom SJTs have been incorporated into high-stake healthcare selection. Medical students must pass specific SJTs in order to obtain certificates, and the medical SJT is one of the aptitude tests completed when people apply for medical education (Lievens, Buyse, & Sackett, 2005; Patterson, Baron, Carr, Plint, & Lane, 2009; Plint & Patterson, 2010).

2.2 New Trends in SJT Research and Development

There are two newly-emerging trends of SJT research and development. One trend is that SJTs are now starting to be applied in training and training evaluation applications. The idea of using SJT for training purpose was forwarded by Hunter, who suggested that the George Washington Social Intelligence Test serve as guidance to prepare students who were short of social experience for the vocational world (Hunter, 1928). Although suggested long ago, few researchers subsequently discussed or studied the application of SJT in training. Visual evidence shows only one SJT, the Cultural Assimilator, that was used for training purposes in 1970s. In the Cultural Assimilator program, trainees were instructed to select the best interpretation on the short episode of cross-cultural encounters. The trainer would explain the reason(s) why the trainee's selection was correct or incorrect, and if the trainee's selection was wrong, they would be asked to re-select until they got the correct answer (Brislin, Cushner, Cherrie, & Yong, 1986; Fiedler, Mischel, & Triandis, 1971). Researchers and practitioners have since started to consider using SJT items as training stimulus materials, for training need analysis, and training evaluation (Hanson, Horgen, & Borman,1998; Hedge, Borman, & Hanson, 1996). Fritzsche, Stagl, Salas, and Burke (2006) published the first article which systematically discussed the concept of scenario-based training and the ways to design and deliver such training and evaluation methods. The research on the similarities and differences of SJTs between training and selection has also been reviewed and discussed (Hauenstein, Findlay, & McDonald, 2010).

The second trend in SJT research and development is to study SJTs from cross-cultural perspective in two aspects, *that is*, (1) generalizing and validating SJTs across cultures and among different cultural groups (Lievens, 2006; Lievens, Corstjens, Sorrel, Abad, Olea, & Ponsoda, 2015; Krumm, Lievens, Huffmeier, Lipnevich, Bendels, & Hertel, 2015), and (2) developing SJTs to measure people's cross-cultural competency (Evelin, Schleicher, & Born, 2008; Rockstuhl & Ng, 2015). Most SJTs are developed in an emic approach, whereby they are theoretically constrained in a limited range of cultural contexts. It may be problematic when an emic SJT is transported to other cultural contexts or to the cultural groups different from the context or the group the SJT is originally developed for. Recently some attempts have been made on cross validating SJTs across different countries. For instance, Krumm et al. (2015) validated an integrity SJT among different ethnical groups in Turkey. Lievens et al. (2015) examined the generalizability of an American-based integrity SJT to the workplace and the job application context in Spain. Those researchers had to remove 6 out of 19 scenarios in order to optimize SJT generalization to other cultural context, which indicated that culture could be a factor hindering SJT generalization.

The second cross-cultural aspect in SJT research, the one most relevant to this dissertation, is to develop SJTs measuring people's cross-cultural competency, a critical capability for an individual to function effectively under culturally diverse situations. In addition to the Cultural Assimilator, the earliest 3C SJT, Cross Cultural Dialogues (Stori, 2017) is an SJT-alike tool to facilitate intercultural communication and understanding, wherein the scenarios are dialogues and no interpretation options are provided. The readers interpret the dialogues by themselves, and then compares their interpretation to the decoding of dialogues provided by the developer. 3C SJTs for selection purpose have emerged more recently. Evelina et al. (2008) designed an SJT to measure cross-cultural social intelligence in two dimensions: ethnocentrism and empathy. It is comprised of written scenarios specific to pairs of countries in comparison, and four response options of each scenario reflect the variance in the degree to which the two dimensions are expressed. Ang et al. (2014a) also created a cultural intelligence SJT with multimedia scenarios for high-stake personnel selection and evaluation. The popularity of 3C SJTs may be prompted by increasing demands for qualified personnel with oversea missions. Another trigger for 3C SJTs could be the plausible assumption that SJTs have the potential to outperforms context-free surveys in assessing 3C because 3C is a situated construct per se (Rockstuhl et al., 2015).

2.3 The Reasons for Resurgence

SJT research was stagnant in the 1970s and 1980s. Breakthroughs in theory, positive meta-analytic results on psychometrical properties and lessons from selection practice brought a resurgence of research interest in SJTs in the 1990s. Lack of theory was one major obstacle in SJT development. SJTs were originally created in practice and rooted from applied usage, so it lacked solid theoretical basis from its inception, which largely hindered its later systematical development. The first theory-wise foundation that SJT relies on is the convention wisdom, "*the best indicator of future performance is the past performance*". That is, human behavior tends to be consistent over time in similar situations.

Motowidlo, Dunnette, and Carter (1990) suggested that SJTs were a low fidelity measure, and its scenarios were hypothetical work situations which partially reflects job realism. They proposed the utility idea that a low-fidelity simulation was predictive of job performance because of behavior consistency. The hypothetical work situation description can arouse human memory about their past behaviors in the same or similar situations. Even when people don't experience the same situations, they can extrapolate the important features from other situations they have experienced. People's judgment on the best reaction to a given situation is formed from their speculation on the hypothetical situations with their previous experience. Motowidlo's utility idea illuminates the promising application values of SJT in personnel selection practices: as a low-fidelity simulation, SJT requires much lower cost in designing than high-fidelity selection tools while providing high levels of criterion-related validity. More importantly, the utility idea serves as the first SJT theory, and largely changed the embarrassing fact that SJT had no theoretical foundation.

Implicit trait policy theory (ITP) is regarded as the first real SJT theory, which was also proposed by Motowidlo and his colleagues (Motowidlo, Hooper, & Jackson, 2006a; 2006b; Motowidlo & Peterson, 2008; Oostram et al., 2012; Whetzel & McDaniel, 2009). ITP theory is concerned with the causal relationship between individuals' personal attributes and their judgment on reaction effectiveness in various situations. People frame the situations they encounter with their ITP. For instance, if an individual has a high level of openness, they are more likely to believe that the course of action conveying a high degree of openness is more effective in dealing with the given situation. The judgment reveals one's strengths or weakness in personal attributes. ITP theory provided one theoretical foundation for SJT, and fundamentally advanced SJT development.

The second reason for SJTs resurgence is the favorable results reported by meta-analyses on SJT criterion-related validity. McDaniel et al. (2001) conducted the first meta-analysis on SJT criterion-related validity among non-student samples, and the meta-analytic result demonstrated that SJTs were a good predictor of job performance over a wide range of jobs (*r* = .34). Another meta-analysis study investigating SJT criterion-related validity in construct level revealed that SJTs were criterion-valid in measuring different constructs: .43 in personality composites, .38 in teamwork skills, .28 in leadership skills, .25 in interpersonal skills, .24 in conscientiousness, and .19 in job knowledge (Christian et al, 2010). Those results strongly support that SJT is a valid tool which encourages their use in workplaces and selection procedures.

The third reason for SJTs resurgence was lessons learned from practice and increasing concerns on the issues caused by overreliance on self-report personality tests in selection. Self-report questionnaires are decontextualized measurement instruments, which actually measure self-concept and does not necessarily reflect actual behaviors (Spencer & Spencer, 1993). Self-report personality assessment is criticized for its susceptibility to faking and for distortion of scores due to social desirability (Griffith, Frei, Snell, Hamill, & Wheeler, 1997; Griffith, Chmielowski, & Yoshita, 2007; Murphy, 2005; Peterson, Griffith, & Converse, 2009). It was estimated that 30%-50% of job applicants consciously distorted their responses to obtain more favorable scores on self-reported personality-based selection measures (Griffith, Chmielowski, & Yoshita, 2007). The test takers can easily elevate their scores, and the response distortion was found to be between 0.5 and 1.0 SD (Ones, Viswesvaran, & Korbin,1995). Additionally, researchers also argue against using self-report personality tests in selection contexts and point out that self-report methods attenuate criterion-related validity (Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt, 2007). Along with the increasing claims for measurement alternatives for self-report personality tests in selection settings, empirical evidence indicates that SJTs are less impacted by social desirability in criterion-related validity (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004) and more resistant to faking because they are less transparent (Chan & Schmitt, 2017).

Another practical concern was the need for alternatives to cognitive ability tests. Cognitive ability tests are questioned for their high risk of adverse impact when used in selection settings. SJTs have comparable validity but far less adverse impact than cognitive tests, and have incremental validity over personality tests and cognitive tests (Jensen, 1998; Lievens et al., 2008; Whetzel & McDaniel, 2009). All of these factors indicate that SJT might be a good alternative to cognitive tests. Finally, the job-relatedness makes SJTs appear face and content valid. The face validity enhances favorability of test takers to the tests, and high content validity evidences SJT validity (Chan & Schmitt, 1997; Chan & Schmitt, 2017; McDaniel & Nguyen, 2001; Motowidlo, Hanson, & Crafts, 1997; Ployhart & MacKenzie, 2011; Salgado, Visweswaran, & Ones, 2001; Whetzel et al., 2008; Whetzel & McDaniel, 2009).

2.4 Psychometric Properties of SJT

SJT methods demonstrate psychometric advantages over self-report personality tests and cognitive tests. However, like other measurement methods, SJTs have their own psychometric strengths and weaknesses. In the section, SJT psychometrical properties are discussed.

2.4.1 Reliability

Like other questionnaire-based measures, internal consistency reliability indexed by coefficient alpha is most frequently used to assess SJT reliability. However, compared with other measures, SJTs internal consistency reliability features low alpha values (mean alpha of .57) with large variance (ranging .24 ~ .94) (Campion et al., 2014; McDaniel et al., 2001; Lievens et al., 2008; Patterson et al., 2012; Polyhart & Erhart, 2003). Length, response instruction formats, and heterogeneity of the items are regarded as the major factors causing these two issues of SJT reliability. SJTs which contain more items tend to display higher internal consistency reliability than those with fewer items (Lievens et al., 2008; Patterson et al., 2012). Response instruction format was found have a moderation effect on SJT reliability with rating instruction format producing higher reliability estimates than most/least and ranking instruction formats (Polyhart & Erhart, 2003).

A high proportion of SJTs measure a composite of various performance-related personal attributes (Christian et al., 2010), and even the construct-specific SJTs contain heterogenous content in scenarios and response options. Internal consistency reliability is only applicable for unidimensional SJT measures (Cortina 1993; Cronbach, 1949). Incompatibility of applying internal consistency reliability to factorially complex SJT likely leads to the low and widely varying alpha values.

The facts call for using other types of reliability estimates for SJTs (Campion et al., 2014; Lievens et al, 2008; Oostram et al., 2015). Researchers recommended that test-retest reliability, parallel-form reliability, or split-half reliability would be more appropriate for SJT (Campion et al., 2014; Lievens et al., 2008; McDaniel et al., 2007; Whetzel & McDaniel, 2009). Both parallel-form and split-half reliabilities require extra effort to ensure the equivalent construct or constructs across the comparative groups of items. Test-retest reliability might be the best method to estimate SJT reliability (Lievens et al., 2008; Whetzel & McDaniel, 2009).

2.4.2 Validity

Criterion-related validity is one of the most appealing psychometric properties of SJT to researchers and practitioners. Studies demonstrated that SJT was predictive of job performance (.26) and exhibited incremental validity beyond personality (3%~5%), cognitive ability (6%~7%), and their combination (1%~2%) (Weekley et al., 2006). It is noteworthy to point out some features of SJT criterion-related validity. First, most studies of SJT criterion-related validity were concurrent designs, and its predictive validity could be decreased because observed correlation in predictive designed studies have been found .04~.15 smaller than in concurrent designed studies (Van Iddeking & Ployhart, 2008; Whetzel & Reeder, 2016). Second, when SJT targets different constructs, its criterion-related validity is likely to fluctuate. The SJTs measuring leadership and teamwork skills have relatively higher correlations with job performance than the SJTs measuring other constructs (Christian et al., 2010). Third, video-based SJTs show stronger criterion-related validity than written SJTs, and the latter is more cognitively loaded (Christian et al., 2010; Weekley et al., 2006). Fourth, response instruction was found to moderate SJT criterion-related validity; however, the moderating direction is not conclusive. Weekley et al. (2006) suggested that behavior tendency instruction displayed higher criterion-related validity than knowledge-based instruction, which, however, is opposite McDaniel et al.'s finding (2003).

The main challenge of SJT validity may be its construct-related validity. Compared with criterion-related validity, far less data or information on SJT construct validity has been published. According to Christian et al.'s (2010) investigation, one third of SJT studies didn't report any detailed construct information and little or no construct validity evidence is provided in published SJT studies (Ployhart & MacKenzie, 2011). An individual's performance in SJT is associated with his/her cognitive ability, personality, and experience (Oostrom et al., 2015). Factorial complexity makes SJTs fail to strongly relate to any specific constructs (Ployhart & MacKenzie, 2011). Its heterogenous nature makes it hard to strip away extraneous factors and uncover the main construct, which sets obstacles to select compatible measures to collect convergent and discriminant evidence. Heterogeneity is detrimental to the internal structure consistency of SJTs, which is an important evidence source for construct validity (SIOP, 2003).

2.4.3 Group differences/adverse impact

Adverse impact is one of key criteria when evaluating a selection measure, especially in United States. Most studies on the adverse impact of SJT focus on race and gender. Evidence has supported that SJTs have much less adverse impact than cognitive tests, but more than self-report personality tests (Jensen, 1998; Motowidlo & Tippins, 1993; Whetzel & McDaniel, 2009). According to the meta-analysis by Whetzel et al. (2008), SJTs with less cognitive loading display smaller group difference than those with high cognitive loading, and video-based SJTs have less adverse impact than written SJTs. The instruction type also moderates the adverse impact of SJT, *that is*, knowledge-based instruction leads to larger race difference than behavior tendency instruction does. Whetzel et al.'s (2008) meta-analysis study also suggested that SJTs showed more group differences in race rather than in gender.

2.4.4 Applicant perception

In recent years the concept of social validity – judgements concerning the social importance of a measure (Wolf, 1978) – has been gaining more attention from researchers and practitioners in selection procedures. Applicant perception is an important index to the social validity of a selection measure (Ryan & Ployhart, 2000; Thibodeaux & Kudisch, 2003). Applicant perception of a selection measure influences one's test performance and intention toward the organization, and potentially impacts legal challenges (Bauer et al., 2006; Ryan & Huth, 2008). When a selection measure is highly favored by test-takers, it has high social validity. Overall, SJT are positively perceived by test takers because of job-relatedness, and video-based SJTs are more favorable than written SJTs because the former show higher fidelity and less cognitive loaded (Lievens et al., 2008; Patterson et al., 2012; Whetzel & McDaniel, 2009).

2.4.5 Coachability

Coachability in tests and measure refers to whether and to what extent an individual's performance on a test can be improved through training or experiencing. One frequently cited study on SJT's coachability was conducted by Cullen, Sackett, and Lievens (2006). The researchers investigated the coaching effect of *avoiding extreme responses* in two SJTs. Their results indicated that the SJTs constructed from SME judgment were less susceptible to coaching. Another relevant issue is whether experience can improve SJT performance. Incumbents outperform job applicants in SJTs, which reflects job experience does help to improve an individual's SJT performance (Motowidlo & Beier, 2010). Thus the retest effect exists which is a potential threat to validity. However, the retest effect is similar with other cognitive tests (Lievens et al., 2005b).

2.4.6 Fakability

Fakability distorts the scores of a measure and hence attenuates its criterion-related validity. Resistance to faking is one important criterion to evaluate the quality of a measure. SJTs are far less susceptible to faking than personality tests (Hooper et al., 2006). The construct targeted, response instruction types, and the development of response options are all factors in SJT's susceptibility to faking (Nguyen et al., 2005; Oostrom et al., 2015;

Whetzel & McDaniel, 2009). When an SJT measures personality or integrity, it is more susceptible to faking than when measuring cognitive ability. Knowledge-based instruction is more resistant to faking than behavior tendency instruction. SJT is more fakable when the response options are written transparently and clearly relate to social desirability.

2.4.7 Utility

Utility, also called as economic utility, is generally used for developing and evaluating selection systems and to demonstrate the value of a selection measure. The utility of a measure is a function of validity, cost, potential adverse impact and applicant reactions. As an important psychometric property, utility has been long overlooked in SJT research. Only a few researchers have discussed the utility of SJTs. Motowidlo et al. (1990) mentioned that SJTs have a higher utility than high-fidelity selection tools because they cost less time and money in developing but yield similar levels of criterion-related validity. No empirical research investigates the economic utility of using SJT in practice (Lievens et al., 2008).

In review, SJTs have advantages in criterion-related validity, applicant perceptions, and utility over other measures and selection tools. It also outperforms cognitive ability tests and personality measures respectively in lower adverse impact and stronger resistance to fakability. Therefore, a SJT is a promising, effective selection measure although more effort is needed to improve its reliability and construct validity.

2.5 SJT Gap Analysis and the Purposes of the Current Research

Several prominent review and meta-analytical articles on SJTs have emerged since 2001 (Campion, 2014; Chan & Schmitt, 2017; Christian et al., 2010; Lievens et al., 2008; McDaniel, 2001; McDaniel, 2007; Oostrom, 2015; Ployhart & Mackenzie, 2011; Weekley et al., 2006; Whetzel et al., 2008). Those articles conducted systematic qualitative or quantitative research on SJTs or one of its properties, summarized the research progression, and identified future directions in SJT research. The future research directions are the interest of the current dissertation. SJT future research directions are mainly categorized into two types: future research on SJT design and design methods, and future research on SJT psychometric properties.

Weekley et al. (2006) discussed the future research on the SJT design components in detail. They summarized five research gaps in SJT scenario generation: 1) no research on the impact of the critical incident sources (SMEs) on SJT effectiveness, 2) no research on comparing the SME-driven and model-based approaches, 3) no research on the best way to write and present scenarios, 4) no research on the best way to present scenarios, and 5) no research on the influence of scenario content upon SJT construct validity. Unfortunately, more than ten years have passed and there has been little progress in those areas except scenario presentation format, wherein video-based scenarios were found to be more favored by test takers and less cognitive loaded than written scenarios (Kanning et al., 2006; Lievens & Sackett, 2006; etc.).

In SJT response options, the main concern of future research is on psychometrical advantages and disadvantages between SME-driven approach and model-based approach (Weekly, et al., 2006). Weekly et al. also called for research on the impact of response complexity upon SJT performance and faking resistance research. Although there have emerged several studies on the complexity of response options and faking in SJTs, no empirical research is available on the comparison of SME-driven and model-based approaches in developing SJT response options. Overall, in the past decade the research on SJT scenarios and response options design has been quite limited and overall SJT research has progressed slowly. There are no evident studies conducted to examine the psychometric merits and demerits of SJT scenarios and response options developed respectively with the two approaches.

Researchers have also been calling for more studies on improving SJT psychometric properties, especially in reliability, predictive validity, construct validity, and utility since the beginning of the century (Christian et al., 2010; Lievens et al., 2008; McDaniel et al., 2007; Ployhart & McKenzie, 2011; Whetzel & McDaniel, 2009). Studies have been conducted to investigate alternatives to internal consistency reliability like test-retest reliability and parallel-form reliability. Some studies have been conducted on SJT predictive validity research, only a few studies investigated SJT construct validity, but no studies have been conducted on economical utility of SJT.

In sum, some future directions for SJT development identified over a decade ago have been studied; however, the research on comparison between different development approaches, SJT utility, and construct validity have lagged behind so far. One main purpose in the current dissertation research is to fill some of SJT research gaps, especially in comparison between the development approaches, construct validity, and utility (see Table 1).

Table 1 — SJT research questions the dissertation focuses on

"Future research" advocated since 2006	Studies on "future research" since 2006	Dissertation focus
Future research on SJT design and design methods		
About scenario		
Potential influence of critical incident sources (SMEs) on SJT effectiveness	N.A.	
Psychometrical strengths of model-based SJTs? (Model-based vs. Critical incident-based approach)	N.A.	√
The best way to write SJT scenarios (reading level, scenario length and complexity)	N.A.	
The impact of scenario presenting format on responses	Kanning et al., 2006 Lievens & Sackett, 2006	
The influence of scenario content on SJT construct validity	N.A.	Somewhat
"Future research" advocated since 2006	Studies on "future research" since 2006	Focus
Future research on SJT design and design methods		
About response options		
SME-driven approach vs. model-based approach in developing response options (their influences on SJT validity)	N.A.	√
The influence of response option complexity on judgment performance	Arthur et al., 2014 Rasmussen 2009	
More research on faking and SJTs - The impact of response content on faking - To what extent are knowledge instructions faking resistant	Ramsay 2006 Oostrom et al., 2017	Somewhat
About response instruction		

- Do response instructions have an impact on SJT measurement properties in the applicant setting?	Klassen et al., 2014	
- How do response instructions differ while controlling for the content of the SJT?	Stagl 2006 Lievens et al., 2009	
About scoring key development		
Do different groups of SMEs yield rational keys of varying validity (incumbents, supervisors, trainers, customers, etc.)? - Systematic research on SMEs	Motowidlo & Beier, 2010 LIevens & Motowidlo, 2016	
About scoring method		
Is one scoring strategy substantially better than other (e.g. consensus)? - Is knowing what to do fundamentally different than knowing what not to do? - Does what one is most likely to do indicate meaningfully different info than what one is least likely to do	Bergman et al., 2006 Legree et al., 2010 Legree & Psotka, 2006 De Leng et al., 2017 McDaniel et al., 2011 McDaniel & Weekly, 2012 Sorrel et al., 2016 Rijmen 2011	
- What are the advantages and disadvantages of multistage SJT	N.A.	
Future research on psychometric property improvement		
About reliability		
Alternatives to internal consistency reliability	Catano et al., 2012 Shi, 2012	Somewhat
About validity		
- Validity generalization to applicant samples	N.A.	
- Convergent and discriminant validity	N.A.	Somewhat
- Predictive validity	Fertig 2009; etc.	Somewhat
- General factor exploration	N.A	
About utility		
Research on economic utility of using SJT	Koczwara et al., 2012	√
About SJT theory		

Why SJT predict work behavior? - Substantial variance in SJTs remaining unexplained except cognitive ability and personality. - New insights are needed to understand the constructs assessed by SJTs	Lievens & Motowidlo, 2016 Lievens & Sackect, 2012 Motowidlo et al., 2009 Motowidlo et al., 2006 Motowidlo & Beier, 2010	Somewhat
The process of analysis that lead to the solution	Oostrom et al., 2010 Motowidlo et al., 2016 Ployhart 2006 Ayal et al., 2015 Reinerman-Jones et al., 2016	Somewhat

The judgment of situations	Brown et al., 2016;
	Rockstuhl et al., 2015;
	Krumm et al., 2015
	Motowidlo et al., 2016	

About SJT generalizability		
Cross-cultural transportability	Lievens et al., 2015 Lievens 2006 Lievenes & Sackett, 2007 Prasade et al., 2017 Evelina et al. (2006) Ang et al., 2014a	Somewhat

Another purpose of the current dissertation is to develop and validate SJTs to measure people’s 3C. Until now there were three SJTs related to 3C, *that is*, Triandis’ Cultural Assimilator, Evelina’s SJT, and Ang’s Intercultural SJT (iSJT*). 3C SJT is still be at a nascent stage and 3C SJTs developed in the present dissertation will be valuable additions to 3C measurement.

Chapter 3
SJT Design and Development

As discussed in Chapter Two, SJTs have appeared in various forms and with different names since their inception. The current dissertation creates and develops two SJTs composed of written scenarios with written response options in multiple choice from. The design mainly involves five aspects of consideration: (1) how to generate scenarios, (2) how to generate response options, (3) how to design instructions, (4) how to establish scoring keys, and (5) what scoring method is adopted (Weekley et al., 2006). The design of each aspects can influence the efficacy of a SJT (Arthur et al., 2014; Weekley et al., 2006). The commonly-used methods of SJT development will be discussed, and the methods used for the current SJTs design will be justified.

3.1 Scenario Development

Scenarios are hypothetic work situations that reflect common situations and encounters in workplaces or other contexts of interest. To simulate response action, scenarios should present dilemmas or problems which need to be solved. Scenarios should also be contextualized in the relevant situations in order to arouse the proper course of action. For instance, if a SJT is designed to measure leadership performance, it is better to contextualize the scenarios in manager-involved situations like leader-employee interaction in organizations (see the illustrative scenario below).

> A member of your department has been employed for three years, but his original project expired after two years. Thus, your manager assigned him to a new job. However, in the last months your manager noticed that the employee regularly shows up in the office quite late and does not work longer than absolutely necessary. In his current project the employee achieves very little progress. (From Leadership Situational Judgement Test by Peus, Braun, & Frey, 2013, *p*. 792).

Scenarios generation is the starting point as well as a major part of the development of SJT (Motowidlo et al., 1997). The traditional approach to scenario generation is SME-driven, which mainly relies on SMEs in collecting critical incidents. SMEs can be incumbents, supervisors, and experts in the relevant field(s). Critical incidents are usually collected from SMEs via interviews or focus group discussions. Critical incidents are experience or stories about situations encountered in the workplace, which convey how the job should be performed as well as the specific behaviors and KSAOs essential for successful performance. Critical incidents also illustrate excellent and poor performance in the work situations (McDaniel & Nguyen, 2001; Smith & Kendall, 1963).

Sometimes SMEs are given the guidance to provide their experience or stories within a delimited scope or only focus on a particular construct and competency in order to generate more relevant scenarios. For instance, to develop leadership scenarios, managers and leaders were interviewed about their experiences only in management. When critical incidents are collected, the developer will screen out the incidents which are not critical by checking content overlap. The incidents which do not overlap will be removed. The resulting critical incidents are grouped and categorized by themes. The representative situations are selected from each theme and written into scenarios. For instance, the SMEs may contribute many stories about time conflict and the way they manage the conflict. The developers first categorize those stories about time conflict in one group, then select or rewrite a situation representative for those time conflict situations. The scenarios of Leadership SJT by Peus and his colleagues (2013) were developed with the traditional SME-driven approach.

Differentiating from the traditional SME-driven approach, some researchers rely on a specific theory or a theoretical model to develop scenarios. The scenarios are written to reflect a theoretically sound model which is composed of personal attributes or competencies extracted from literature reviews, relevant theories, and/or job analysis results. Those personal attributes or competencies are theoretically or empirically supported as important determinants to effective performance in the work situations. Researchers call this approach a model-based approach or theory-driven approach (Weekley et al., 2006; Bledow & Freses, 2009). In this dissertation I use model-based SJT to refer to the SJT in which the scenarios and response options were created based on a model. It is noteworthy to clarify that the construct-driven SJTs are, in essence, model-based SJTs, because their development and validation relies on the nomological network of the target construct (Chen et al., 2016).

A theoretically rigorous model and clearly defined personal attributes or competencies are crucial for SJT developers to create scenarios with the model-based approach. When the model is well established and defined, the developers can create scenarios to reflect personal attributes or competencies of interests. The Teamwork SJT developed by Stevens and Campion (1999) is a typical model-based SJT, where scenarios were developed from the theoretical model built on the wealth of literature. The scenarios were written to reflect the knowledge, skills, and abilities required for effective teamwork derived from the literature on groups and teamwork.

It is atypical for the clear-out application of either SME-driven approach or model-based approach to developing SJT scenarios. Only a small number of SJT scenarios have been developed with the model-based approach without any assistance from SMEs. In most cases, researchers used SMEs to collect critical incidents in reference with some theoretical guidance (Weekley et al., 2006). For instance, when Motowidlo et al. (1990) developed their management performance SJT, they first summarized the shared core managerial skills from the review of the documented job analysis on some managerial positions. Then they organized focus group discussions among incumbents and supervisors about critical incidents of effective and ineffective management performance, especially of the shared core managerial skills extracted in the preliminary review. The scenarios of the construct-specific SJTs are developed by the hybrid approach such as Integrity SJT (Chen, 2009), Personal Initiative SJT (Blew & Frese, 2009), and Cross-Cultural Social Intelligence SJT (Evelina et al., 2006).

The three approaches have their own pros and cons in generating SJT scenarios. The traditional SME-driven approach can generate the most authentic scenarios in the workplace, and the scenarios generated by this approach are comprehensive and mostly reflective of the workplace reality. However, the effectiveness of the approach is largely influenced by the SMEs and the interview/discussion questions. Also, the SME-driven approach has several drawbacks. First, the scenarios by SMEs are heterogenous and contains a plethora of factors, which makes it hard to detect the key factor(s) associated with performance. Therefore, the assessment results of SME-driven SJTs are less applicable for research purposes. Second, the approach is time consuming. SMEs recruitment, interviews, and transcription of interview data costs large amounts of time, human resources and money. The developers also need to ensure information saturation in order not to miss important situations happening in the workplace.

Comparatively speaking, the model-based approach is less time consuming and allows the measure developers more autonomy. The developers can write the scenarios without any SMEs. Another merit is that the model-based scenarios are more construct-focused because they are developed to project specific person attributes or competencies. The biggest problem of the approach is scenario authenticity, where the scenarios could fail to reflect the real workplace situations or the full picture of the reality due to the limited scope of the developers on the situations. The hybrid approach seems to be able to make up the weaknesses of the two approaches by setting targeted constructs for SMEs interview, and SMEs can generate situational information only associated with those constructs. However, it should be pointed out that while sharing some merits with both approaches, the hybrid approach bears both the weaknesses of SME-driven and model-based approaches: first, it consumes much more time than either of the two approaches because the developers have to spend time on SMEs interviews as well as on the model establishment; second, the authenticity of scenarios will be attenuated because of the intervention of targeting at specific constructs; third, because SMEs may lack relevant theoretical knowledge support, it is hard for them to generate construct-specific situation information.

Unfortunately, there exists no empirical research on the three approaches, and we can't make confident conclusions on which an approach is superior to the others. All the discussions are from researchers' practical experience and informed guesses. Therefore, in the current dissertation research, I will develop two sets of SJT scenarios, with the SME-driven approach and the model-based approach, compare the psychometrical properties of the two SJTs, and check which approach is better according to empirical evidence.

3.2 Response Option Development

Response option development is the following step once the scenarios are created. In this step the possible courses of actions reacting to each scenario are collected. Ideally those actions can cover the full range of effectiveness in dealing with the scenario situations. However, in reality it is impossible to cover all possible responses, and the more practical way is to select the sample responses which are most representative for different level of effectiveness.

Like scenario development, the SME-driven approach is commonly used to generate response options. The difference from scenario generation is that SMEs in response option generation includes not only those with expertise in the relevant fields (real SMEs) but also poor performers and/or the novices who have no or little experience. The options produced by real SMEs are assumed as effective actions while those provided by poor performers and novices are treated as ineffective actions. This strategy attempts to span a range of effectiveness in performance (Lievens et al., 2008). The response options of military SJTs used in WWII were developed in that approach (Northrop, 1989).

The model-based approach is also used in creating response options. Like model-based scenarios, the responses are created based on theories and literature findings. The response options could either be a composite performance like those created by SME-approach, or reflect a specific construct or a dimension of a higher-level construct. The response options of Leadership SJT (Peus et al., 2013) was developed with model-based approach. All response options were created with reference to the Full Range of Leadership Model (Bass, 1985; Bass & Avolio, 1994), and each response option reflects the different leadership style (see the illustrative response options below, from Peus et al., 2013, p 792.).

> a. (IS) Manager discusses with the employee how he could push his project forward by new impulses. (S)he encourages the employee to voice his own ideas and makes suggestions himself/herself, for example regarding cooperation with other projects.
>
> b. (IM) Manager motivates the employee to put more effort into the project again and explains to the employee how he can thereby make a substantial contribution to the vision of success of the entire department.
>
> c. (IC) Manager asks the employee in a personal conversation why he only makes little progress in his project at present and offers to support him with the further project design by providing specific feedback.
>
> d. (II) Manager points out to the employee how important his full commitment is to him/her. (S)he openly communicates his/her criticism of the employee's current work ethics, but emphasizes that (s)he highly valued his performance on former projects.
>
> e. (CR) Manager agrees with the employee that from now on he will work hard on this project again. (S)he explains to the employee that in return he might receive a bonus if the project develops positively.
>
> f. (MBA) Manager announces to the employee that from now on (s)he will actively control whether the employee is not keeping the regular working

> hours. Moreover, (s)he will control the work results of this employee on a daily basis in order to check it for mistakes.
> g. (MBP) Manager waits and sees in which way the employee's job performance develops in the next months. (S)he only interferes if the employee is on the verge of giving up the project and quitting his position.
> h. (LF) Manager does not attend to the employee and his progress in the project. (S)he leaves the responsibility for the success or failure completely to the employee and only engages in his/her own projects.[1]

Response options produced by the model-based approach can also appear to reflect different levels of the constructs of interest. Each option indicates a competent level of the specific construct. A typical example is the response options of Personal initiative SJT (Bledow & Frese, 2009). All response options were framed from high personal initiative to low person initiative, reflecting the continuum of personal initial competence level. The options indicating high and low personal initials were first written based on the theoretical findings, and typical examples collected from the incumbent survey. Thereafter, the options derived from the high and low personal initiative responses were constructed by the developers to fill the middle range of continuum (see the below illustrative item, from Bledow & Frese, 2009, *p.* 233).

> You are under enormous pressure to accomplish your tasks on time. Yesterday, new trainees started in your department. They are unfamiliar with the workflow in your department. You have to interrupt your work to answer trainees' questions and to correct their mistakes. You are expected

[1] Intellectual stimulation (IS), inspirational motivation (IM), individualized consideration (IC), idealized influence (II), contingent reward (CR), management by exception active (MBA), management by exception passive (MBP), and laissez-faire leadership (LF).

to do both, to finish your work on time and to take care of the trainees. What would you do?

Least likely --- most likely

a. I tell the trainees that I am available after work to answer their questions.

b. I openly say that I cannot take care of the trainees and work for better initial training of the trainees.

c. I send the trainees to my colleagues when they have questions.

d. I try to get by without becoming stressed and worn out.

Until now, few researchers have discussed the pros and cons of the approaches on response option generations. Free style seems prevailing in SJT response option generation. In order to call for more attention in response option generation approaches, I will use both SME-driven and model-based approaches to create the response options to examine the efficacy of the two approaches in SJT response option generation.

3.3 Response Instruction Design

The existing studies on SJT response instruction fall into two aspects: response instruction research and response instruction format research. The former focuses on what types of information an SJT is expected to collect, people's knowledge on good and bad behavior, or their actual behavior preference. The latter is regarding the three types of SJT scales, *i.e.* rating, ranking, and most/least.

3.3.1 Response instruction

SJT response instruction is the guidance for test takers to make judgement among a set of response options. McDaniel and Nguyen (2001) summarized that SJT instructions typically involve two types of information: the test takers' knowledge on the best way to deal with the given situations and their actual behavioral tendency or preference to reacting to the given situations. The two types of information are called as knowledge-based instruction and behavior-tendency instruction, respectively. The knowledge-based instruction measures test takers' knowledge by asking them for the best or worst response to the given situation. This type of instruction is usually formulated as "should/shouldn't do". The behavior-tendency instruction is often constructed as "most/least likely to do" and "would do" to obtain test takers' intention and behavioral preference in the give situations. This categorization is widely accepted by experts (Lievens et al., 2009; Weekley et al., 2006).

Compared with other aspects of SJT development, more research has been conducted on response instruction and more consensus have been achieved. Knowledge-based instruction is found more cognitive-loaded and less susceptible to faking, while behavioral instruction is more related to personalities and less resistant to faking (Lievens et al., 2008; Lievens et al., 2009; Whetzel & McDaniel, 2009). SJTs with knowledge-based instruction measure maximum performance while SJTs with behavioral instruction measure typical performance (Lievens et al., 2008). Researchers claimed that SJT instruction was an important moderator on SJT criterion-related validity, but they held different opinions on the directionality of the moderation. Some believed SJTs with knowledge instruction had higher criterion-related validity, and behavioral tendency instruction attenuated SJTs' criterion-related validity due to high likelihood of faking (Chan & Schmitt, 2017; McDaniel & Nguyen, 2001; Nguyen et al., 2003; Oostrom et al., 2015; Stagl, 2006; Whetzel & Reeder, 2016). However, other researchers argued for higher criterion-related validity produced by SJTs with behavioral tendency instructions based on a small sized meta-analysis results

(McDaniel et al., 2007; Lievens et al., 2009). In addition, Ployhart and Ehrhart (2003) found a relatively low correlation between the judgment performance with the two response instructions for the same SJT items. Both moderation and low correlation findings call for caution when researchers and practitioners select the response instruction. Choice of response instruction could likely influence an SJT's criterion-related validity and construct validity.

Although the previous meta-analysis studies concluded the instruction of behavioral tendency had lower validity and more susceptible to fake, such a conclusion was unwarranted due to lack of empirical support (only one empirical study explicitly compared the two types of instructions), unmatched comparison in meta-analysis, and an empirical finding that behavioral tendency instruction had higher validity (Chan & Schmidt, 2017; Polyhart & Ehrhart, 2003). Also, what one thinks and what one behaves are not necessary the same. The 3C SJTs are expected to assess an individual's capability of dealing with culturally complex situations, which is more associated with people's immediate reaction to cross-cultural encounters rather than their cross-cultural knowledge, therefore, behavioral tendency instruction is more appropriate for the two 3C SJTs.

3.3.2 Response instruction format

The response instruction format is another important factor, which can't be ignored in SJT response instruction research. The format is viewed as "potentially a critical design feature" in SJT development, and influences SJT's construct validity and measure effects on racial and gender groups (Authur et al., 2014; Ployhart & Ehrhart, 2002). Rate, rank and most/least are the common response instruction formats of the existing SJTs. Rate instructs the test-takers to rate the effectiveness for each response options in terms of a Likert scale. Rank format requires the test takers to place all response options in the order of effectiveness. Most/least format demands test-takers to decide the most and least effective options among a set of response options.

Only a few researchers have systematically investigated the three response formats. Ployhart and Ehrhart (2003) compared the best/worst format with rate format, and found that rate format showed better internal consistency than best/worst but no significant variance was found in criterion-related validity between the two response formats. Arthur et al. (2014) compared rate, rank, and most/least formats with an integrity SJT among a large group of job applicants. Their study revealed that when compared with the other two formats, the SJT with rate response format displays lower group differences, higher internal consistency, higher test-retest reliability, stronger correlation with personal traits while weaker correlation with cognitive ability, and higher level of response distortion. The rank and most/least formats display similar in most psychometric properties, but most/least shows higher internal consistency, less cognitive-loaded, more favorable to test-takers and less completing time than rank (see Table 2).

Table 2 — Properties comparison among the three response formats

Response format	Rate	Rank	Most / least
Internal consistency	Highest	Lowest	Medium
Test-retest reliability	Highest	Lowest	Medium
Alternative form reliability	Lower	Higher	Higher
Criterion-related validity	No sig. diff.	N.A.	No sig diff.

Cognitive-loading	Lowest	Highest	Medium
Correlation with personality	Higher	Lower	Lower
Group difference	Small	Large	Large
Response distortion	Most susceptible	Less susceptible	Less susceptible
Test-taker response	Most favorable	Less favorable	Medium
Time to completion	Lest	Most	Medium
Complexity of scoring method	Complex	Complex	Less complex
Free from extreme respondence	Heavily influenced	Free from influence	Free from influence

When selecting an instruction format, test developers also need to consider what scoring method they want to use, which will be discussed in detail in Section 3.5. Generally speaking, the developers will have to face more complicated calculating and converting issues when adopting rate or rank response formats. The scoring method of most/least format is relatively simple, stable, and causes less controversial than rate and rank formats. Most/least format will be utilized in the current 3C SJTs regarding the trade-offs among the three response formats.

3.4 Scoring Key Development

Scoring key refers to the weight or score produced by SJT items. Unlike most instruments, of which each item has a definitely correct answer, a SJT doesn't have an objectively right or wrong answer to how to deal with a scenario effectively (Bergman et al., 2006; Legree & Psotka, 2006). In reality, there could be different ways to solve a dilemma or a problem, and different actions may produce either similar or differential effects.

Therefore, it is hard to conclude that one response option is definitely better than the other options. The inherently ambiguous nature of SJTs demands researchers use extra caution when developing a sound key, which directly influences the usefulness of SJTs.

Four methods have been created and applied to the SJT scoring key development: empirical, theoretic, rational, and hybrid. Each scoring method has different characteristics (see Table 3). Empirical method determines the key by examining the relationship between each response option with a criterion measure, and usually the option with highest correlation with the criterion measure is set as the key to the scenario. Empirical keying is less theory-grounded, and the quality of criterion measures and the sample response largely determine the quality of the key. Its criterion-bounded nature also limits the generalizability of the empirical key. Opposite this, the theoretical method heavily relies on theories to determine the best and the worst reaction to SJT scenarios. Theoretical keys have higher generalizability because they are unconstrained to criterion measures and sample responses.

Table 3 — Comparison of scoring key development methods

	Empirical	Theoretical	Rational	Hybrid
Reference sources	Criterion measure	Theories	SMEs, test-taker	Criterion measure, theories, SMEs and test-taker
Generalizability	Low	High	High	Medium
Validity	Significant	Not significant	Significant	Significant
Gender effect	No	No	No	No

The rational method is called an SME-driven method, because it relies on the judgment performance of experts or test-takers to decide the scoring key. There are several ways to generate rational keys. One way is to let experts decide the best and the worst reaction options to the scenarios, then the correctness of answers is decided in terms of interrater agreement among those experts. Another way is to compare the rating results between experts and novices. The options rated as the best answer by the experts are treated as the correct answer, and the option rated as the best answer by the novices while not by the experts is treated as incorrect answer. The rational key is determined by SMEs and unrelated to any particular criterion measures, so it is more generalizable than empirical scoring key.

The hybrid scoring key is developed by combining different methods to develop scoring keys or by combining scoring keys developed by independent methods. For instance, empirical method and theoretical method can be used together to develop a scoring key, wherein researchers use theoretical method to do primary assessment on response options, then use empirical method to examine the primary assessment. When the empirical outcome is consistent with primary theoretical assessment, the scoring key is finalized; otherwise, the discrepancies will be investigated, and the option decisions will be adjusted based on the follow-up theoretical and empirical research. The hybrid scoring key could potentially increase the predictivity of SJTs (Mumford, 1999).

The scoring key developed with each of four methods produces different effects on SJT validity, internal consistency, and adverse impact (Berman et al., 2006; Legree & Psotka, 2006; De Leng et al., 2017). Berman and his colleagues utilized the Leader Skill Assessment to compare the four types of scoring keys in terms of validity and group differences. The results reveal that all the four keys display good discriminant validity and were free from gender differences. Empirical, rational, and hybrid keys demonstrate significant validity and incremental validity for supervisory ratings and promotion rates, and rational keying yields the highest incremental validity over cognitive ability and personality scores.

Considering the advantages of rational scoring keys in generalizability, validity, and gender differences, as well as the availability of reference sources for the current dissertation research, I will adopt the rational approach to develop the scoring keys for both 3C SJTs.

3.5 Scoring Method Selection

Scoring method is another important factor which can influence SJT psychometric properties (Leng et al., 2017; Weekley et al., 2006). Differentiated from the scoring key development, scoring method refers to the way to calculate and convert the gap of judgment performance between SMEs and the test taker. Because SJTs lack clear-cut correct or wrong answers to scenarios, the traditional way to assign a score only to the correct answer is not feasible for SJTs in most cases. The choice of scoring method for a SJT depends on its instruction format and the type of the scoring key (Leng et al., 2017). Most existential studies on scoring methods focus on the rate format with reference to rational keying (Leng et al., 2017; Weng et al., 2018). Those studies proposed a variety of statistic interventions to control for systemic errors, distance, and central tendency. Since the 3C SJTs will utilize most/least format with rational keying, those interventions are not feasible for the current dissertation research.

Little theoretical and empirical guidance is available about how to score SJTs of most/least response instruction format with rational keying. However, some findings in rational keying and the scoring practice of most/least formatting SJT provide some suggestions on the best way to score judgement performance of 3C SJTs. First, consensus scoring method is required when rational keying is used. Consensual scoring is a type of profile matching (McDaniel et al., 2011), wherein a test-taker's response profile is compared with the reference profile. The test taker's profile consists of his/her item responses, and the reference profile is a profile of item means generated from SMEs. Less variance indicates better performance on the test, and the variance between the two profiles is converted to the final score for the test taker's performance. Consensus scoring strategy will be utilized for developing the reference profile in the current 3C SJT development.

Second, the method to calculate and convert the variance largely impacts the psychometric properties of the given SJT. For rate format, there are five commonly used methods, *that is*, raw, standardized, dichotomous, mode, and proportion consensus scoring, which differentiate from each other on statistic interventions and also influence internal consistency and criterion-related validity of an SJT (Leng et al., 2017; Weng et al., 2018). Notably, the score of SJT with rate format is very likely to be distorted by extreme responding habits, *that is*, choosing the extreme scores will lower scores while avoiding selecting the extreme scores will increase the scores. For rank format, it is more complicated to calculate and convert the variances of rankings between SMEs and test takers, so there is little research available in the previous literature. It is relatively easy to calculate and convert the variances in most/least judgement performance between SMEs and test takers, because the most and the least options generated by SMEs serve as the objectively correct answer and exclude the remaining options from score calculation. When test takers choose the same most/least options as SME consensus, they will earn points; if they don’t choose the right option, they won’t get any points; and if they chooses the reversed option, *that is*, choose the SME least option as most or choose the SME most option as least, they will lose points. The scoring method eliminates the influence of extreme response on measure results, the score can be calculated in a much easier way and the score even doesn’t need to be converted. Therefore, I will use that way to score the two 3C SJTs in the current dissertation.

In sum, the two 3C SJTs in the present research will adopt behavior tendency instruction and most/least response format. The scores will be calculated with a rational key with consensus scoring strategy. The only differences of the two 3C SJTs is the way to develop their scenarios and response options: one will use SME-driven approach and the other will use model-based approach.

Chapter 4
3C and its Antecedent Model

3C research started in the 1950s from the discussion on effective intercultural communication and interaction. Researchers worked on exploring the factors that were most important to facilitate intercultural effectiveness (*e.g*. Cleveland, Mangone & Adams,1960; Hammer et al., 1978; Ruben, 1976). The factors recommended by the early researchers varied in a wide range, including actual behaviors, behavior dimensions, personal characteristics, or their combinations. Later, researchers used 3C or similar names to refer to an individual's capability to achieve intercultural effectiveness, for instance, cross-cultural competence (Gabrenya et al., 2012; Johnson, et al., 2006), intercultural communication competence (Spitzberg & Changnon, 2009), bicultural competence (Bell & Harrison, 1996; Black et al., 1991), global competence (Adler & Bartholomew, 1992; Hunter et al., 2006), cultural intelligence (Earley & Ang, 2003; Thomas et al., 2008), intercultural sensitivity (Chen & Starosta, 1997), and intercultural competence (Deardorf, 2006; Howard-Hamilton et al., 1998; Ting-Toomey & Kurogi, 1998). Although named differently, the noticeable similarities and the substantial overlapping across the definitions of those constructs indicate that those constructs are very similar. For instance, Earley and Ang (2003) defined *cultural intelligence* as "a person's capability for successful adaptation to new cultural settings, *that is*, for unfamiliar settings attributable to cultural context" (*p*. 9), which is the same as the definitions of cross-cultural competence proposed by Gabrenya et al. (2012), Gertsen (1990), and Johnson et al. (2006), and similar to the definition of intercultural competence by Spitzberg and Changnon (2009) and Whaley and Davis (2007). In the current dissertation, I adopt the name of *cross-cultural competency (3C)* and the definition suggested by Chen (2017), "*an individual's capability to effectively function in culturally diverse contexts, which is influenced by a set of individual antecedents*" (*p*. 14).

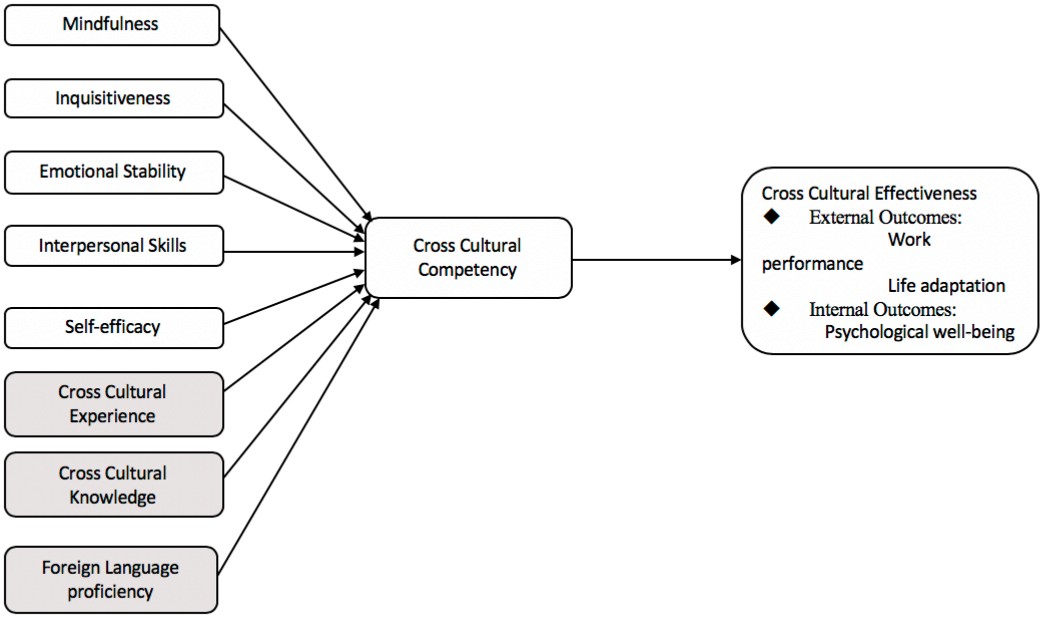

Figure 1 — 3C model

Numerous qualitative and quantitative studies on 3C offers a rich literature source for developing a model-based 3C SJT. 3C is viewed as the determinant to sojourners' success in oversea mission accomplishment, adaptation to living and working in different cultural environment, as well as psychological adjustment to foreign cultures (Leung et al., 2014). Meta-analysis also displayed that 3C had positive effects on expatriate effectiveness ($r = .36$; Li-Yueh & Alfiyatul, 2015). Therefore, in the model the right side of 3C is individuals' overall performance in cross-cultural contexts, which comprises of work performance, general and social adaptation, and psychological well-being (see Figure 1).

Disputes exist in 3C operationalization and its antecedents. Researchers confuse 3C per se and its antecedents (Johnson et al., 2006), and disagreements prevail in the proposed 3C nomological networks (Chen, 2017). Although many 3C models have been proposed by researchers, none of them are universally acknowledged. Therefore, the dissertation didn't adopt any existing 3C model but presents a model based on extensive literature review with the consideration of the way 3C is conceptualized. As stated previously, 3C is conceptualized as a capability in the present dissertation. A capability is not born, can be trained and changed, but can also be influenced by some inborn traits to some degree. It is a composite construct determined by a set of individual attributes such as cognitive style, personality traits, and experience. Those attributes jointly determine an individual's 3C level. The key 3C antecedents were derived from 3C and cross-cultural adaptation literature: „ inquisitiveness, emotional stability, interpersonal skills, self-efficacy, cultural knowledge, cross-cultural experience, and foreign language proficiency (Abbe et al., 2007; Chen, 2017; Johnson et al., 2006; Leung et al., 2014).

4.1 Cross-cultural Mindfulness

Mindfulness is a cognitive style composed of openness to novelty, alertness to distinctions, sensitiveness to different contexts, awareness of multiple perspective, and orientation in the present (Langer, 1997; Langer, 2000; Sternberg, 2000). The five components are closely related, each component "*leads to the others and back to itself*", and actually are "*different versions of the same thing*" (Langer, 1997, *p.* 6). In cross-cultural contexts, a mindful individual tends to be open to different cultural values, norms, beliefs, behaviors, and rituals. They are also aware of differences between one’s own and other’s culture, can imagine multiple perspectives resulting from cultural differences, and are able to be flexible when dealing with issues with present orientation (Chen, 2017).

Ample evidence highlights the significance of the mindful components to an individual’s 3C. Openness was found related to individuals’ flexibility in shifting their behaviors across cultures (Shaffer et al., 2006; Tarique & Weisbord, 2013). Such a flexibility enables people to do quick and accurate assessment on cross-cultural situations and to adapt based on the needs of cultural and business contexts (Caligiuri, 2008). Openminded expatriates were more capable of finding substitute entertaining activities enjoyed from the host country. Mindful people could quickly perceive cultural differences, analyze the assumption gap between themselves and local people, and facilitate culturally appropriate behaviors in cross-cultural situations (Byram, 1997; Deardorff, 2016; Kupla, 2008). Keeping aware of the local culture improved expatriate managers’ relationship with local partners and eliminated conflicts in management (Buckley et al., 2006). On the other side, mindlessness is detrimental to overseas missions. For instance, ignorance on cultural differences led to American psychologists’ failure in disaster assistance in Sri Lanka after the Indian Ocean tsunami (Christopher et al., 2014). Lack of sensitivity and awareness impeded sojourner transformation between home and host cultures, which ultimately caused failure in oversea missions or sojourner maladaptation (Mendenhall & Oddou, 1985; Shaffer et al., 2006).

4.2 Cross-Cultural Self-efficacy

Self-efficacy, as defined by Bandura and Schunk (1981), refers to an individual's judgment on how well he or she "can organize and execute courses of action required to deal with prospective situations containing many ambiguous, unpredictable, and often stressful elements" (*p*. 587). It is self-perception of one's own capability rather than the actual skills or competencies evaluated by third parties. For example, if a student has a high level of self-efficacy for the final exam of American History, this means that they believe that they can do well on the history exam.

Cross-cultural self-efficacy refers to one's own belief in their capability to manage the challenges and difficulties arising in cross-cultural situations (Wilson, 2013). Cross-cultural self-efficacy has been found to positively influence people's acculturation and sociocultural adaptation in foreign cultures (Bhaskar-Shrinivas et al., 2005; Gong & Fan, 2006; Long, Yan, Yang, & Van Oudenhoven, 2009). People with high self-efficacy are more likely to feel comfortable in culturally diverse environment, and perform more effectively in the cross-cultural missions because they believe that it is in their capabilities to handle the situations (Leiba-O'Sullivan, 1999). According to meta-analysis results, self-efficacy was moderately related to expatriate oversea performance and job satisfaction (Li-Yueh & Alfiyatul, 2015). The significant role of self-efficacy in 3C has been supported across several studies (Abbe et al., 2010; Ang et al., 2007; Wilson, 2013).

4.3 Inquisitiveness

Inquisitiveness, or curiosity, is a motivational antecedent to 3C. It refers to "an active pursuit of understanding ideas, values, norms, situations, and behaviors that are new and different" (Bird et al., 2010, *p*. 815). Inquisitiveness reflects an individual's passion to learning different cultures. It is worthwhile to point out that some researchers use motivation as a proxy to inquisitiveness (Ang et al., 2007; Arasaratnam, 2009; Gabrenya et al., 2012); however, Pinder (2008) suggested that this is not proper because motivation is a wider concept composed of a set of energetic forces originating from both within and beyond an individual's being, which initiate work-related behavior and determine its forms, direction, intensity, and duration. Motivation can be curiosity, self-efficacy, or external incentives. To be precise, the 3C model proposed by the current dissertation uses inquisitiveness, instead of motivation, to clarify learning motivation as an independent 3C antecedent which is distinguished from self-efficacy.

Inquisitiveness is a precondition of acquiring behavior and knowledge. Individual differences in cultural inquisitiveness result in 3C differences (Gabrenya et al., 2012). Sojourners with higher level of inquisitiveness are believed to learn and apply foreign cultural values, rules, and behaviors better than those who are less inquisitive. When these individuals kept learning and updating their foreign cultural knowledge, their 3C would be elevated (Abbe et al., 2007). A longitudinal study among more than one hundred executives across the globe revealed that inquisitiveness was one major characteristic of effective global leadership (Black et al., 1999). Empirical investigations also supported that inquisitiveness was a significant predictor to an individual's 3C over a variety of samples (Doutrich & Storey, 2004; Gong & Fan, 2006; Kawashima, 2008; Messelink & Thije, 2012).

4.4 Interpersonal Skills

Interpersonal skills in cross-cultural contexts refer to the skills to communicate and interact with people from different cultures and to maintain a good relationship with them. Interpersonal skills have plenty of synonyms in 3C literature such as interpersonal engagement, self/other orientation, and relational skills. Extraversion is closely associated with interpersonal skills, and extroversive people usually have better interpersonal skills and have an easier time making friends. A large number of 3C researchers proposed that interpersonal skills were a critical factor associated with 3C and global leadership (Arthur & Bennett, 1995; Byram, 1997; Deardorff, 2006; Hammer et al., 1978; Ting-Toomey, 1999). Good interpersonal skills help to foster understanding across cultures and to reduce misunderstanding caused by discrepancies in cultural values, and hence promote effectiveness in dealing with cultural complexity.

In a Delphi study among the top cross-cultural scholars and administrators, interpersonal skills were consensually rated as one of the basic elements to 3C (Deardorff, 2006). Li-Yueh and Alfiyatul's (2015) meta-analysis also showed that interpersonal skills were significantly related to expatriate oversea effectiveness (*r* = .30). Empirical studies among expatriates in multinational companies revealed that interpersonal skills were a major contributor to expatriate success over a variety of jobs, and were regarded as important as technical competency in expatriate personnel selection (Arthur & Bennett, 1995; Shaffer et al., 2006).

4.5 Emotional Stability

Negative emotion posts a negative impact on people's performance (Chi et al., 2013; Kaplan et al., 2009). Negative emotions seem unavoidable in people's cross-cultural experience and cultural shocks can trigger all types of negative feelings (Kim, 1988). Emotional stability, or neuroticism, is an important antecedent to 3C (Costa & McCrae, 1992; Kealey 1996; Shaffer, et al., 2006; Tung, 1981; Wildman et al., 2016). People, who are more stable in emotion can keep their emotions positive and control their negative emotions. They are more likely to keep functioning in culturally complex situations as normally as they do in their familiar environment. In contrast, more neurotic or less emotionally stable people easily become frustrated and stressed out; this often leads them to become aggressive or to adopt defensive reactions (like withdrawal or turnover) in unfamiliar cultural contexts.

Expatriate studies have accumulated plenty of evidence for the role of emotional stability. Ones and Viswesvaran's (1999) meta-analysis revealed emotion stability predicted expatriate job performance (β = .25), adjustment (β = .28), oversea mission completion (β = .27), and relationship with local people (β = .24). Emotional stability was negatively related to expatriate withdrawal cognition (Peltokorpi & Froese, 2014; Shaffer et al., 2006), but emotional instability magnified the negative effect of stressful situations and led to hasty decisions or misjudgment on the real situation (Caligiuri, 2000a; Ormel, et al., 2001).

4.6 Cultural Knowledge

Cultural knowledge is closely related to an individual's 3C level (Abbe et al., 2007; Arthur & Bennett, 1995; Byram, 1995; Deardorff, 2006; Gabrenya et al., 2012; Howard-Hamilton et al., 1998; Imahori & Lanigan, 1989; Johnson et al., 2006; Ting-Toomey, 1999; Ting-Toomey & Kurogi, 1998). It is self-evident that cultural knowledge is a necessary condition for an individual to start effective cross-cultural communication and cooperation. Cultural knowledge incorporates culture-specific and culture-general knowledge. Culture-specific knowledge includes knowledge of values, norms, beliefs, cognitive and behavioral styles, living habits, and the communicative and interactive rules in one's home and host countries. Culture-general knowledge refers to people's understanding of cultural values and dimensions in a global perspective and awareness of culture impact on people's beliefs, values, and behaviors which results in diversity and discrepancies across cultures.

Both culture-specific and culture-general knowledge contribute to an individual's capability to function well in cross-cultural contexts. Culture-general knowledge prepares people for cognitive adjustment, being mindful and keeping alert to shocks, conflicts, anxiety, and uncertainty happening in cross-cultural contexts (Brandl & Neyer, 2009). Knowledge of specific culture helps people to quickly know the differences between their own culture and other cultures and make them easier to understand foreign people's values and behaviors. Pre-departure cultural knowledge training was empirically supported to promote trainees continuous culture learning in the host country and their adjustment to the local culture (Tarique & Caligiuri, 2009).

4.7 Cross-cultural Experience

Cross-cultural experience has been suggested as an important antecedent to 3C (Abbe et al., 2007; Arasaratnam & Doerfel, 2005; Benet-Martinez, 2006; Black, et al., 1991; Hammer et al., 2003; O'Sullivan, 1999; Tarique & Weisbord, 2013). Cross-cultural experience includes overseas study experience, oversea working experience, social experience with people from different cultures, growing up in a multicultural family, cultural training experience, oversea travel experience, and other types of experience that provides exposure to different cultures. These experiences enable people to observe other cultural groups and to learn cultural differences by cultivating cultural awareness and increasing the ability to detect the implicit values unique to a specific culture. With those experiences people are more likely to handle cultural shocks and conflicts in a mature and well-prepared manner (Abbe, et al., 2007; Benet-Martinez et al., 2006; Hammer et al., 2003). These individuals are more tolerant to ambiguity happening in cross-cultural communication and display more cognitive flexibility in interacting with people from other cultures (Tarique & Weisbord, 2013).

Cross-cultural experience is frequently included in 3C research, wherein it is treated as a control variable or predictor variable for an individual's 3C or adaptation to a foreign culture (Ang et al., 2007; Basow & Gaugler 2017; Moon et al., 2012; Tamam, 2010). In most cases cross-cultural experience was significantly related to an individual's capability of handling cross-cultural issues or to an individual's adaptation to the host country. Besides, people's experience is positively related to their attitude to the specific culture. Good experience produces positive attitudes to the specific culture and people of that culture, and such positive attitudes help people cross-cultural adaptation. Instead, bad experience leads to negative attitudes toward the culture and cultural people, which results in maladaptation (Arasaratnam, 2009; Chao et al., 2017).

4.8 Foreign Language Proficiency

It is self-evident that mastery of the local language is a must for an individual to work and study smoothly in a foreign country, which directly determines how competent an individual is in cross-cultural communication and interaction. Speaking in the same language largely enhances communication, and being capable of speaking in the local language helps to gain favorable attitudes from the local people and helps to quickly build a relationship with them. Empirical evidence repeatedly demonstrated significant correlations between foreign language proficiency and cross-cultural competency (Basow & Gaugler 2017; Meydanlioglu et al., 2015; Paige et al., 2003; Strekalova, 2013; etc.).

4.9 The 3C Model Used for 3C SJT Development

The 3C model used for 3C SJT development in the current dissertation research is derived from the literature review on 3C, its antecedents, and cross-cultural adaptation with reference to the model proposed by Chen (2017). 3C in the model is operationalized as a capability which is jointly determined by people's cross-cultural mindfulness, self-efficacy, inquisitiveness, emotional stability, interpersonal skills, cross-cultural knowledge and experience, as well as foreign language proficiency (see Figure 1) Among those antecedents, cross-cultural knowledge, experience, and foreign language proficiency can be directly evaluated by knowledge tests or from people's bio-records. However, the psychological antecedents, mindfulness, inquisitiveness, emotional stability, self-efficacy, and interpersonal skills are hard to assess directly. The 3C SJT developed with the model-based method in this dissertation endeavor targets assessing people's 3C through measuring the five 3C antecedents in the proposed model.

Chapter 5
The Development of Two 3C SJTs

The current dissertation research is designed to fulfill two purposes: to develop two 3C SJTs respectively with SME-driven approach and model-based approach, and to investigate the efficacy of the two SJT development methods via comparing the psychometric strengths and weakness of the two 3C SJTs. The research design is comprised of three phases: (1) to develop 3C SJTs with the two approaches, (2) to validate the two 3C SJTs, and (3) to compare the psychometric properties of the two 3C SJTs. This chapter focuses on Phase (1), the development of the two 3C SJTs.

5.1 Attributes of SJT Development

Campion et al. (2014) identified SJT attributes involved in SJT use, development, and scoring methods with a descriptive summary of 59 empirical SJT studies (see Figure 2). The attributes of SJT use are dimension numbers, study purpose, sample size, SJT study context, construct assessed, and SJT research design. The SJT development and scoring attributes include response medium, response format, instruction format, number of items, situation and scenario development, key development, scoring method, scenario presentation, and stimulus medium. I use Campion's (2014) structure as the general reference framework in 3C SJT design. Because the dissertation research is to compare the efficacy of different approaches in developing scenarios and response options, most of the attributes in the structure are fixed the same when the two 3C SJTs are designed except scenarios and response options development methods in order to eliminate unnecessary "noise" caused by irrelevant or confounding factors (see Table 4). Notably, there exist evitable discrepancies of the dimension numbers and the measured constructs in the two SJTs, which are caused by the two scenario and response option development methods.

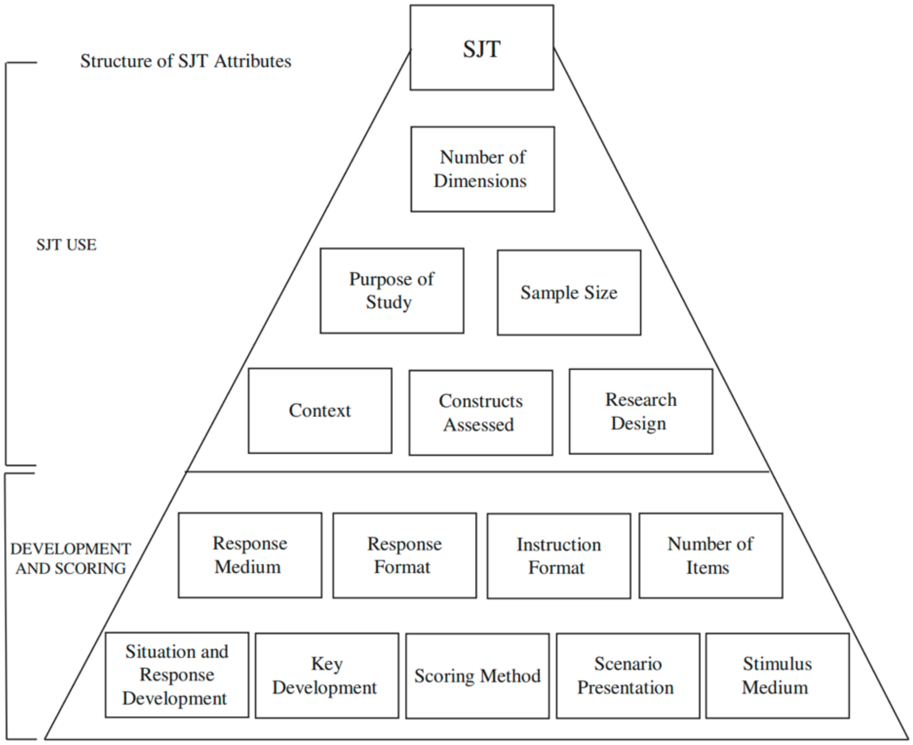

Figure 2 — Structure of SJT attributes (adapted from Campion et al., 2014)

Table 4 — Comparisons in development, scoring methods and use attributes of the two 3C SJTs

	SME-driven 3C SJT	Model-based 3C SJT
SJT use		
# of dimensions	Heterogenous	5 dimensions
Purposes of the study	1. SJT development 2. SJT Validation 3. Psychometric property comparison	
Sample size	Same size	
Study context	Same context	
Constructs assessed	Holistic performance in cross-cultural situations	self-efficacy mindfulness emotional stability inquisitiveness interpersonal skills
Research design	1. Concurrent design 2. Reliability investigation 3. Validity investigation 4. Face validity investigation 5. Utility investigation	
Development and scoring		
Response medium	Same medium: paper-and-pen	
Response format	Same: most/least	
Instruction format	Same: behavioral tendency	
# of items	Same item number	
Situation and response development	SMEs generate both scenarios and response options	Developers generate both scenarios and response options
Key development	Same: Rational key	
Scoring method	Same: 1 for right answer, -1 for reverse answer, 0 for other	
Scenario presentation	Same: Sequential	
Stimulus medium	Same: paper-and-pen	

5.1.1 SJT use

For the SME-driven SJT, the SMEs were asked about the effective and ineffective performance, which lead to heterogeneity of both scenarios and response options. However, for the model-based SJT, I led a team of SMEs developed the scenario and its response options to target specific constructs, hence the SJTs may be more construct-oriented and have a clearer structure of dimensionality. It is the same case in the current 3C SJT development. All the items in SME-driven 3C SJT are heterogenous while each combination of the scenario and its response options in model-based 3C SJT focuses on assessing one of the five 3C antecedents.

Both 3C SJTs are developed to realize the same research purposes as discussed previously. The research is concurrent design with both SJT performance data and the other performance evaluation data collected at the same time. The reliability, validity, face validity and utility will be investigated with the same procedures.

5.1.2 Development and scoring

The two SJTs will be administered to the same sample group in the same setting to ensure the same study context and to eliminate between-individual differences. Via Qualtrics, the two SJTs were presented randomly and the items of each SJTs were also presented in a random order. The response instruction of both SJTs use "most/least" behavior tendency instruction. Both SJTs contain the same number of items and each item has 5-8 response options. The keys of both SJTs were developed by the same group of SMEs (*n* = 4), who have abundant and successful experiences in cross-cultural management for more than ten years. This method was applied to score test-taker's performance; *that is*, if the individual chooses the same most/least options as the scoring key, they will gain 1 point. If the test taker chooses the reversed most/least options against the scoring key, they will lose

1 point. If they choose the other options, they will neither receive nor lose any points. The total score indicates their performance in the SJTs, which indicates their cross-cultural competence.

5.2. Development of 3C SJT with SME-driven Method

The SME-driven method mainly relies on SMEs to generate a pool of critical incidents and courses of response actions. The situations extracted from the pool of critical incidents are developed into SJT scenarios. A different group of SMEs, including both experienced and unexperienced performers, provide their response actions to each scenario. Those response actions are processed into response options which ideally cover the full spectrum of all possible response actions.

For this dissertation research, twenty-three students in a southeast internationalized university, who were engaged in several multicultural teamwork and cross-cultural interactions, were recruited as scenario SMEs. All SMEs were required to have at least three experiences in multicultural teamwork and each experience was required to last more than three weeks. These requirements ensured that each SME had had adequate opportunities to communicate and interact with people from different countries. The SME recruitment was advertised in the university forum. Unqualified students were screened out with a qualification survey via an email. In the qualification survey, those SMEs not only confirmed they met the SME requirement, but explicitly stated that they had plenty of opportunities and adequate time to communicate and interact with people from different cultures. Besides adequate involvement in multicultural teamwork, these SMEs either shared a dormitory room with students from different cultures, belonged to an ethnically diverse athletic group, or worked with different cultural students in the campus societies or off-campus organizations. The SMEs participated in one-to-one interviews with a reward of a $25 Walmart gift card.

Each interview lasted 1 to 1.3 hours and was audio recorded. The interview was semi-structured, each SME student answered a list of the same questions in detail (see Appendix A), and additional questions were asked according to their answers. The whole interview focused on SME students' cross-cultural experience, their good and poor performance, and the challenges they faced when they interacted with the foreigners at work or in the daily life. The interview audio records were transcribed with all personal identifiers removed. The transcribed interview data was periodically analyzed to check if the information was saturated. Data saturation is reached when no additional information is obtained from the study (Fusch & Ness, 2015; Guest, Bunce, & Johnson, 2006). The data saturation appeared in the 18th interviewee. The interview work continued until the 23rd SME student to ensure data saturation. The demographic information of the scenario SMEs is shown in Table 5.

Table 5 — The demographic information of the scenario SMEs

No.	Gender	Nationality	Mother language	Major	Education level
1	male	China	Chinese	Computer engineering	Grad
2	female	Iran	Persian	Mechanical engineering	Grad
3	Male	Egypt	Arabic	Engineering	Undergrad
4	female	Zimbabwe	English	Bio-engineering	Undergrad
5	male	US	English	Computer science	Undergrad
6	male	Cameroon	French	Aviation management	Undergrad

7	male	US. CA	English	Aerospace management	Undergrad
8	male	US	English	Bio-chemical	Undergrad
9	male	China	Chinese	Mechanic engineering	Undergrad
10	male	US	English	Aeronautical analysis	Undergrad
11	male	Bangladesh	English	Computer engineer	Grad
12	male	China	Chinese	Oceanological engineering	Grad
13	male	Bangladesh	English	Science education	Grad
14	Male	Iran	Persian	Engineering	Grad
15	female	Indian	English	Info system	Undergrad
16	Female	Russian	English	physics	Grad
17	Female	Venezuela	Spanish	Business	Undergrad
18	Female	Serbia	Serbian	MBS	Grad
19	Female	UK	English	Psycho	Undergrad
20	Female	US	English	Interdisciplinary education	Undergrad
21	Female	US	English	Marine biology	Undergrad
22	Female	US	English	Social science	Faculty
23	Male	Canada	English	Aerospace	Grad

Two rounds of qualitative analysis were conducted on the interview data. In the first round the data were categorized by themes, which resulted in a total of 35 themes. Then in the second round of data analysis, the critical incidents in each theme were further examined, recategorized, and combined in terms of similarity. This resulted in a final set of 13 themes covering all the critical incidents generated by SMEs. Themes included language obstacles, communication, jokes, time management, leader/leadership, slackness, conflict/conflict management, trust building/cooperation, working style, instruction/coaching style, mindset, discrimination, and other events in general life such as making friends, gossip, parties, and entertainment. The most typical situation(s) of each theme was extracted and written into scenarios. As a result of these interviews, the thirty-six scenarios were created.

In order to investigate whether the primary scenarios are typical situations in cross-cultural interaction, five students and one faculty member in an I/O psychology program -- who had oversea studying and working experience -- were invited to rate the typicality of the primary scenarios in a 5-point Likert scale (1= *strongly disagree* to 5 = *strongly agree*). The rater demographic information was listed in Table 6. Any scenario that received *disagree* or *strongly disagree* ratings were removed. All raters *agreed* or *strongly agreed* with the typicality of seven scenarios, and another thirteen scenarios got *agree* or *strongly agree* from 80% of raters. Another five scenarios with 50% of raters rating *agree* or *strongly agree* were also retained for the further investigation. As a result, a total of 25 scenarios were used for the following response action pool creation.

Table 6 — The demographic information of the raters for the scenario typicality

Rater	Gender	Nationality	Host country	Stay length	Status
Rater 1	Male	Aruba	U.S.	5 years	Graduate
Rater 2	Female	China	Canada, U.S.	5 years	Graduate
Rater 3	Female	U.S.	Italy	6 months	Graduate
Rater 4	Male	U.S.	Spain	6 months	Graduate
Rater 5	Male	Egypt	U.S.	2 years	Graduate
Rater 6	Male	U.S.	China, Spain, Austria, Italy	10 years	Faculty

A second group of SMEs ($n = 4$) were recruited to help create a response option pool for the 25 scenarios. Those SMEs were recruited from graduate students with cross-cultural experience. Their responses to each scenario served to generate keyed responses. In order to catch the complete spectrum of possible actions, including less effective and poor response actions, the scenarios were turned to an open-end questionnaire and distributed to undergraduates enrolled in one of two common classes (Introduction to Psychology and

Cross-Cultural Management). The undergraduates were required to write down their response to each scenario in 1-2 simple sentences, and were also encouraged to write down the reason about their response in 1-2 sentences. Their participation was rewarded with 2 extra class credits. A total of 53 students completed the questionnaire. All the responses were sorted in terms of similarity, and the overlapped responses were excluded. As a result, each scenario had 7 – 15 primary response options.

To reduce the response options into a reasonable size, another group of students (*n* = 63) were recruited to participant in the same open-end questionnaire about their responses to each scenario. The frequency of each primary response option was investigated with this round of response data collection, and those response options with highest frequency were retained for each scenario. My supervisor, who has plenty of expertise in SJT development, was invited to assist response option finalization. As a result, each scenario has 5-8 response options.

Four cross-cultural experts were invited to select the best and worst response from the primary response option list for each scenario. Each expert had more than 10 years of experience engaging in cross-cultural activities or oversea missions. All of them had achieved accomplishments in oversea missions or cross-cultural management. The interrater agreement was analyzed, and the scenarios, of which the best or the worst response options were not agreed by the experts, will be excluded from the final version. The best and worst responses agreed by all experts will be used as the scoring key to the SME-driven 3C SJT. Some response options in the middle (neither the best nor the worst) were removed to ensure each scenario have the similar number of response options. A sample is illustrated below.

Imagine you study at a foreign university for your master’s degree. One day you hear a Ph.D. student complain you are rude because you didn’t greet him first when you met. You also learn that the country is hierarchical and status-sensitive. What will you most/least likely do next time when you meet that PhD student?

a. I would greet him and explain that you don't mean to offend him.

b. I would greet him as if nothing happened.

c. I wouldn’t greet him and mind my own business.

d. I would tell him it is not proper to talk behind my back.

e. I would speak to him about cultural differences.

f. I would avoid him by keeping your distance from him.

5.3 Development of 3C SJT with Model-based Method

In the current dissertation research, the model-based 3C SJT was developed based on the 3C model suggested in Chapter 4. This SJT is supposed to assess an individual’s 3C in terms of its five psychometric antecedents separately. Each dimension of the model-based 3C SJT was composed of several scenarios targeting a 3C psychometric antecedent.

The response options of each scenario were designed to reflect a continuum of the targeted construct levels. For instance, for the cross-cultural mindfulness scenarios the options reflect the continuum from least mindfulness to most mindfulness. The response option reflecting the highest level of the measured construct is regarded as the best option while the one reflecting the lowest level is the worst response. When writing the scenarios and response options, reference was made to the existing scales relevant to each targeted construct and to the method of how the construct was conceptualized and operationalized by other researchers and scale developers. Although the response options were written by the developer, all the model-based scenarios were transformed to an open-ended questionnaire, which were distributed to the undergraduates taking a psychological course (*n* = 68). The procedure was the same as the one to collect the SME-driven 3C SJT response options. The response action pool was used to check if the response options written by the developers were plausible in reality and able to cover most real response behaviors. A thorough content investigation demonstrated that the response options of the model-based 3C SJT did cover most of real reactions generated by the participants.

5.3.1 Cross-cultural mindfulness scenarios and response

Langer (1997) systematically analyzed mindfulness construct, and described it as an ability to actively draw novel distinctions. Most of the existing mindfulness measures operationalized mindfulness as awareness of oneself, novelty and distinctions, non-judgment, and flexibility to situations (*e.g*. Langer's Mindfulness Scale, Five Facet Mindfulness Questionnaires, *etc*.). This operationalization was adopted to write cross-cultural mindfulness scenarios. The mindful scenarios were written focusing on awareness, non-judgment, and flexibility in cross-cultural contexts. The response options are written to reflect the various levels of people's mindfulness. The example scenario is illustrated as below.

> Imagine you are an international student, and you and your local friends decide to watch a movie in the theater on the weekend. You visit the theater website and find a new movie with the strange name Bohemian Rhapsody. What will you most/least likely do?
>
> a. I would ignore it and go on reading through the movie list.

b. I would watch the trailers and reviews available online.

c. I would search for the movie information and what Bohemian Rhapsody means.

d. I would choose this movie to watch.

e. I would ask my friends to make a decision.

f. I wouldn't watch it until my friends recommend it.

g. I would ask my friends what the film is about.

5.3.2 Self-efficacy scenarios and response options

Self-efficacy is the belief in one's own capability of doing a task successfully. Notably, self-efficacy is specific to a functioning domain and consistency across functioning domains is not expected. For example, an individual who displays a high level of self-efficacy in one task may display low self-efficacy in another. Therefore, self-efficacy scales should be constructed under the specific contexts (Bandura, 2006). Therefore, in the 3C SJT, self-efficacy scenarios were constructed under cross-cultural contexts to ensure precisely measure an individual's belief in their capability of solving cultural problems or dilemmas.

There are different ways to operationalize cross-cultural self-efficacy. Based on empirical investigation and factor analysis results, Abbe et al. (2010) adopted three criteria to measure cross-cultural self-efficacy by focusing on communication effectiveness, influence effectiveness, and preparedness. Ang et al. (2004) conceptualized cross-cultural self-efficacy under motivation scope, their Motivational CQ targets people's interest and self-confidence in cross-cultural interaction, and partially measures an individual's self-efficacy in cross-cultural socialization and adjustment. Wilson (2013) took in Ang et al.'s conceptualization of cross-cultural efficacy, and cross-cultural efficacy in that study also includes interest aspects of motivation.

However, as discussed in Chapter Five, I sought to avoid using motivation as a proxy to self-efficacy because it also includes the factor of inquisitiveness when writing the scenarios and response options. Utilizing Bandura's view that self-efficacy is discriminant from inquisitiveness and should be assessed separately (Bandura, 2006), cross-cultural self-efficacy scenarios were written to reflect how people are persistent when facing cross-cultural challenge and how much effort they decide to devote to solve the challenge. When people feel they are highly capable, they tend to devote more efforts and time to the task (Bandura et al., 1996; Bandura, 2006). Like Fan and Mak‘s (1998) Social Self-efficacy Scale, which measures an individual's difficulty appraisal and social confidence feeling in a set of social interaction settings, cross-cultural self-efficacy scenarios describe the challenges in cross-cultural preparedness and communication. And the response options were written to indicate a range of levels of efforts an individual is willing to devote to handle those cross-cultural challenges. The way to writing response options was similar with the response option design suggested by Bandura's self-efficacy scale wherein the statements representing different levels of difficulty are rated as a test-taker's confidence level. The illustrative item is presented below.

> You are given an opportunity to participate in a two-week project in which you will have to work with experts from India, China, Zambia, Russia, and France to create a marketing plan. The project team should form a report and present it to the company board in the end. Regarding techniques you

are qualified for the project requirement. What are you most/least likely to do?

a. I would not accept the offer because I don't think I can work with foreigners well.

b. I would accept the offer but I will not spend much time and effort on the project.

c. I would accept the offer and will spend time and effort on the project.

d. I prefer to wait for a while to see how other collogues react to the offer.

e. I will accept the opportunity if it is offered again.

5.3.3 Inquisitiveness

Inquisitiveness is a relatively simple psychological construct to measure because its conceptualization is less controversial. Inquisitiveness is commonly conceptualized as the desire to learn and the tendency to show interest and curiosity in unfamiliar people, things, and environment. There are few scales specific to inquisitiveness, and most available inquisitiveness scales are subscales under personality or character strengths measures (Lee & Ashton, 2004; Bernard, Mills, Swenson, Leland, & Walsh, 2006; Duan & Bu, 2017; Hogan & Hogan, 1992). Motivational CQ partially measures interest in foreign cultures. I referred to these scale items when creating the inquisitiveness scenarios. The response options reflect the behaviors responding to the different levels of people's inquisitiveness (See the illustrative item below).

Image you are an international student, and you and your local friends decide to watch a movie in the theater on the weekend. You visit the theater website and find a new movie with the strange name *Bohemian Rhapsody*. What will you most/least likely do?

a. I would ignore it and go on reading through the movie list.

b. I would watch the trailers and reviews available online.

c. I would search for the movie information and what Bohemian Rhapsody means.

d. I would choose this movie to watch.

e. I would ask my friends to make a decision.

f. I wouldn't watch it until my friends recommend it.

g. I would ask my friends what the film is about.

5.3.4 Interpersonal skills

Interpersonal skills have been measured in terms of building and maintaining relationship and exchanging information (Lievens, 2013). Interpersonal skills are also partially measured by some composite performance SJT. For instance, an SJT used to predicting college student success partially measures students' interpersonal skills (Peeters & Lievens, 2005). Interpersonal skills in 3C are conceptualized as skills at communicating and interacting with people from different cultures and maintaining a good relationship with them. The interpersonal skill scenarios are involved with communicating and socializing challenges in cross-cultural settings. The response options reflect the degree of the willingness to open or maintain communication with foreign people or how smart people deal with communicative difficulties in cross-cultural contexts as the example item illustrates below.

You move to a new apartment with a new roommate. Your roommate has different routines than you: he plays loud music, always has friends visiting, and is very untidy. What do you most/least likely to do?

a. I would tolerate his behaviors.

b. I would look to move out.

c. I would address my concerns and establish rules agreed by both of us.

d. I would ask my roommate to be more respectful and considerate.

e. I would fight back like playing loud music when he studies, bring my friends.

f. I would use expressions and postures to show that I am bothered.

g. I would leave my roommate messages on what is expected to do or not to do.

5.3.5 Emotional Stability

Emotional stability is conceptualized as a capability of staying calm and withholding the negative emotions when people experience cultural shock and cultural conflicts, and of keeping positive attitudes when they encounter frustrations arising from culturally complex situations. The scenarios focus on difficult or frustrating situations in cross-cultural settings, and the response options describe a range of possible emotional reactions triggered by those situations. Those emotional reactions reflect people's cross-cultural competent level of emotion management. The illustrative item is as below.

You are invited to a weekend party hosted by one of your friends. When you arrive, you find you do not know anyone except the host. What do you most/least likely feel in the situation?

a. I would feel nervous.

b. I would feel angry.

c. I would feel nothing special.

d. I would feel excited.

e. I would feel awkward.

f. I would feel upset.

g. I would feel calm.

5.4 Response Instruction, Rating Form, and Scoring Method

As discussed previously, both SJTs adopt behavioral tendency instruction requiring the test-takers to select their most and least likely behaviors among a list of response options to each scenario. When their most and least selection is the same to the most/least scoring keys, points are earned towards the total score or specific construct score. When participants select the reversely coded options, they lose points. Otherwise, scores do not change.

5.5 Calibrations for the Two 3C SJTs

According to the previous SJT research and studies, an individual's judgment performance and the psychometric properties of SJT are influenced by scenario complexity, reading level, the scenario length, the item numbers, the number of response options, response instruction and rating form (Weekley et al., 2006). In order to make the two SJTs more comparable and to eliminate possible confound factors caused by format inequivalence, the scenarios and response options of the two 3C SJTs will be calibrated to be consistent before psychometric comparisons.

5.5.1 Pilot study

The two 3C SJTs were administered to the same group of participants in a random sequence via Qualtrics. The reliability, internal structure and item characteristics will be investigated. Items performing poorly will be removed.

5.5.2 Calibrations between the two 3C SJTs

Scenario calibration. The scenarios of both SJT were calibrated to be the same or similar in terms of content, complexity, reading level, length, and total number of response options. Word Readability Statistic Tool was used to examine the level of readability of the two SJTs and to make sure both SJTs have the similar readability. T-test result revealed that no significant difference was found in the two SJTs (*t* = 2.12, *p* < .05).

Response option calibration. Each scenario in both SJTs has 5-8 response options. Although it is not possible to keep the similar length among response options within and between the SJTs, long or multiple sentences were avoided. All response options were written with simple words, and Word Readability Statistic Tool was used to ensure similar reading level. The response instruction and rating form were kept consistent between the two SJTs.

Chapter 6
Hypotheses and Research Questions

This chapter discusses the research hypotheses and their underlying rationale and theories. Research questions are put forward about the utility of the different SJT development methods.

6.1 SJT vs. Self-report Measures in 3C Measurement

3C has been regarded as the critical quality for successful performance in cross-cultural environment. How to measure an individual's 3C accurately is always a main concern in cross-cultural research. Most of the existing 3C measures are self-report in a Likert-scale form. More than fifty 3C self-report measures have been published and used for research and practical purposes, however, most of them are subject to validity issues. Two prominent reviews suggested serious validity issues prevalent among the frequently used 3C measures, and only one or two measures could consistently demonstrate acceptable validity (Gabrenya et al., 2012; Matsumoto & Hwang, 2013).

The poor validity of the 3C measures may be due to the deficiencies in the nature of self-report Likert-scale measures, that is, weak resistance to faking and measuring self-concepts. Those 3C measures take the form of short statements with a Likert scale where the test takers are required to rate how well each statement describes themselves or how much they agree with those statements. Therefore, those measures actually assess individuals' self-perceived capability, instead of their actual capability, of dealing with culturally complex situations. Such self-perception is susceptible to self-enhancement bias. People tend to overestimate their capabilities and see themselves better than others, known as the lake Wobegon effect or above-average effect (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Meyers 1998). Also, it is likely that the test takers deliberately elevate their ratings to be more social desired rather than reflecting their real performance (Griffith et al., 2007; Ones et al., 1995).

It is actual behaviors, rather than self-perception, that matters in cross-cultural effectiveness, therefore, self-report measures are not the best method to assess 3C but SJT, a situated measure, may be more appropriate (Rockstuhl et al., 2015). The situated nature of SJT is more likely to precisely catch the test takers' behavioral tendency or preference in dealing with cross-cultural issues, which, in turn, reflects their actual 3C level. In addition, SJT is an indirect measure, and the measured construct(s) or measured purpose is less apparent to the test takers than most self-report measures (Chan & Schmitt, 2017). It is harder for the test takers to figure out which response option in SJT is more socially desired, and thus they are less likely to adjust their judgment performance than they do in self-report measures. Unlike self-report measure of which the criterion-related validity is subject to be attenuated by social desirability (Morgeson et al., 2007; Ones et al., 1995), SJT is less impacted by social desirability in its criterion-related validity (Oswald et al., 2004). Given that SJTs are both closer to the actual behavior and more resistant to social desirability, I hypothesize:

Hypothesis 1: *3C SJTs have higher criterion-related validity than 3C self-report measures*.

6.2 SME-driven Method vs. Model-based Method

SME-driven method is also called as a “sampling” approach because it largely depends on SMEs’ experience. With the SME-driven approach, SJT developers act more like information gatherers, processers, and organizers. The scenarios and response options developed with SME-driven method appear higher level of realism because SMEs are less likely to generate and identify unrealistic situations and courses of action. A large group of SMEs can produce a large pool of realistic situations and a wide range of reactions to the given situations (Weekley et al., 2006).

The major issue of SME-driven method is the resulting scenarios and response options are unavoidably full of heterogeneous items and simultaneously tap a number of predictor constructs/factors which are hard to detect or interpret in many cases (McDaniel & Whetzel, 2005). The heterogeneity makes it hard to clear explain why and how the SME-driven SJTs work in selection or in predicting job performance. It is also a challenge to validate SME-driven SJTs with multiple evidence sources due to its heterogenous nature.

Another issue of SME-driven method in developing SJTs is that method demands more personnel, time, and economic sources in recruiting and training SMEs for SJT development. Data collection and processing is very time consuming. SMEs’ experience could largely influence the quality of SJTs. Different SME groups are likely to produce differential scenarios and response options (Weekley et al., 2006). Large amounts of time and personnel sources are also needed to conduct SME interviews and focus group discussions to generate adequate situations and responses. Additionally, a large pool of situations and responses requires additional time and effort to edit scenarios and response options based on data analysis results.

More recently, researchers call for model-based method in developing SJTs (Arthur & Villado, 2008; Campion et al., 2014; Mumford, Van Iddekinge, Morgeson, & Campion, 2008; Weekley et al., 2006). The main features of model-based SJT are construct-specific and theory-based, each scenario and/or response option taps a specific personal attribute or competency, and theory is the foundation of model formulation and justification.

With the model-driven method, the development of SJT starts with clearly defining the given construct(s), then the scenarios and/or response options are generated to only tap a given construct(s) and the dimensionality of the construct. The scenarios and/or response options are expected to adequately sample the conceptual domain of the given construct(s) or its dimensionality. Technically the developers take the major role to generate both scenarios and response option writing without or with very limited help of SMEs. The biggest advantage of model-based approach is that the content of SJT is more refined and controlled by the developers, and hence the scenarios and response options can capture the specified construct(s) or the unique dimension(s) of construct (Weekly et al., 2006).

However, because of high dependence on theories, the model-based approach is largely constrained if there is the lack of solid theories or extensive empirical research on the target construct(s). That is the biggest drawback of this method in SJT research and development. Apparently, compared with SME-driven method, the model-based method requires less time, personnel and money to develop an SJT. However, there exists a risk that such a cost saving in development is at the cost of psychometric quality. No studies have been conducted to investigate the utility of the two approaches, that is, the trade-offs between SJT development cost reduction and SJT psychometric quality decrease. The current dissertation research attempts to compare the utility of the two approaches, and explore the answers to the two research questions as follows.

Research question 1: *Which approach produces an SJT with higher psychometric properties, SME-driven approach or model-based approach?*

Research question 2: *Which approach has a higher economical utility in developing an SJT considering the tradeoffs between the amount of time, personnel and money used in measure development and measure psychometric properties, SME-driven approach or model-based approach?*

Compared with SME-driven SJTs, model-based SJTs are more construct-focused and the constructs are more detectable and interpretable because they are built within theoretical framework or with theoretical guidance (Weekley et al., 2006). The developers stick to the underlying model to control the content of scenarios and response options. In model-based SJTs each scenario and its response options only target at a specific construct, therefore, model-based SJT is likely to display higher internal consistency than SME-driven SJT. Therefore, I hypothesize as follows:

Hypothesis 2: *As a whole, model-based SJT will have higher internal consistency than SME-driven SJT.*

Hypothesis 3: *When model-based SJT targets several specific constructs, its sub-scales will have higher internal consistency than the internal consistency of the overall model-based SJT.*

When the response options of an SJT developed with model-based method reflect various levels of a specific construct, those response options are more transparent (Oostrom et al., 2015). When test takers make judgment on situations, their decisions are likely to follow social desirability and choose the option most favored for social desirability. It is hypothesized as follows.

Hypothesis 4: *Model-based SJT is more susceptible to social desirability than SME-driven SJT and will be more strongly correlated with a measure of social desirability than SME-driven SJT.*

The model-based scenarios and response options reflect the underlying theoretical or competency model instead of the real situations, which may be detrimental to the realism. The scenarios and response options of model-based SJT is less reflecting the realistic situations than those generated by SMEs, therefore, SME-driven SJT may have higher face validity than model-based SJT.

Hypothesis 5: *SME-driven SJT has higher face validity than model-based SJT.*

6.3 General Domain Knowledge vs. Job-specific Knowledge

Although SJTs are believed to be capable of measuring a specific construct, it is still restricted in the range of measured constructs (Schmitt & Chan, 2006; Chan & Schmitt, 2017). According to Chan and Schmitt (2017), the primary dominant construct measured by SJT is adaptability constructs and job contextual knowledge. Adaptability constructs "are likely a function of both individual difference traits and the result of acquisition through previous experiences", and job contextual knowledge is "gained through experience in various real-world contexts" (Chan and Schmitt, 2017, *p*. 224). They suggested that adaptability constructs and job contextual constructs are situational judgement competencies, which are caused or predicted by the traditional KSAOs (cognitive abilities, personality traits, values, and experience) with varying weights, and SJTs measure those situational judgment competencies which are proximal causes to job performance or other related criteria.

Similarly, Lievens and Motowidlo (2016) proposed that SJTs measure procedural knowledge on the way to effectively perform in work situations, which comprises general domain knowledge and specific job knowledge. Specific job knowledge corresponds to Chan and Schmitt's (2017) job contextual construct concept, which is acquired from specific job experience. General domain knowledge is similar to the adaptability constructs suggested by Chan and Schmitt, but regarded to be rooted from individual implicit trait policy. It refers to the knowledge of "the utility of expressing certain traits" which are believed importance for effective performance. For instance, when emotional stability leads to better cross-cultural teamwork, those who know that will have more general domain knowledge about the utility of emotional stability and more aware of keeping emotional stable. Unlike specific job knowledge, the general domain knowledge is context-independent, which is the result from the interaction of socialization process and personal dispositions. The personal dispositions, like emotional intelligence, interest, values, personality traits, etc., are antecedents to the general domain knowledge (see Figure 3). Lieven and Motowidlo pointed out the general domain knowledge can predict performance across work situations, and advocated developing generic SJT which measures general domain knowledge deliberately and systematically. Their empirical studies supported the idea of the generic SJT (Motowidlo & Beier, 2010; Motowidlo et al., 2016). Specific job knowledge and general domain knowledge independently contributed to variance in job performance, and general domain knowledge can predict individual performance across a variety of situations.

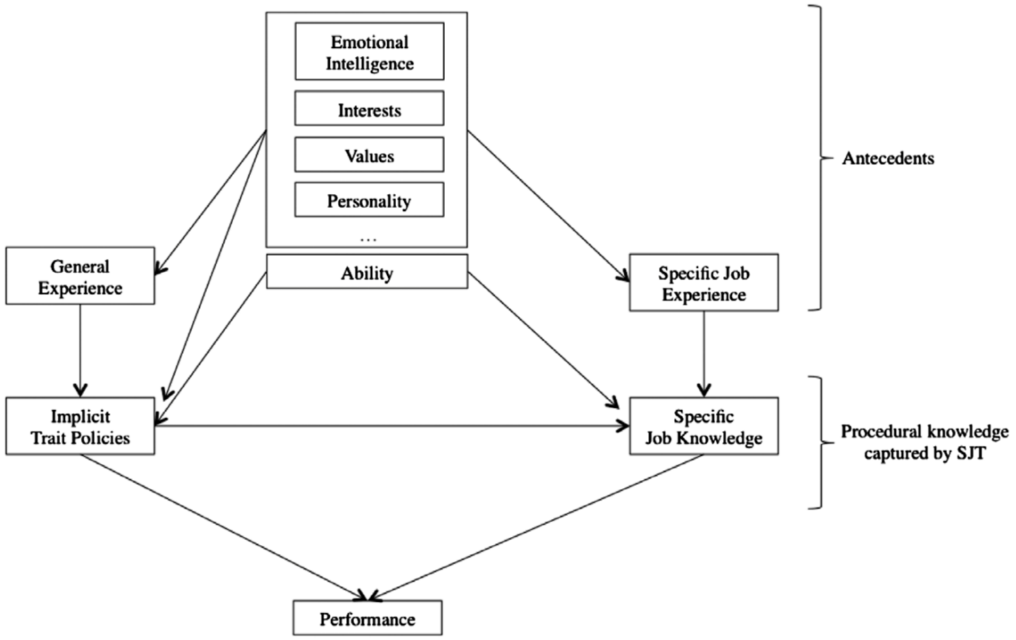

Figure 3 — Expanded model of the knowledge determinants and antecedents of situational judgment test (SJT) performance (from Lieven & Motowidlo, 2016)

When the 3C model is mapped to SJT's determinants and antecedent model proposed by Lieven and Motowidlo (2016), mindfulness, interpersonal skills, self-efficacy, emotional stability and inquisitiveness determine general domain knowledge in predicting cross-cultural performance, while cultural knowledge, cultural experience, and foreign language proficiency are the determinants of specific job knowledge, because they are culture-specific and more culturally context-dependent. The model-based 3C SJT developed in this study only measures general domain knowledge. Differently, the SME-driven 3C SJT targets at an individual's overall performance in cross-cultural situation, which doesn't discriminate general domain knowledge and job-specific knowledge. Therefore, test-takers' judgement performance on the SME-driven 3C SJT is the joint result from their general domain knowledge and job-specific knowledge. Compared with SME-driven 3C SJT, the model-based 3C SJT only measures one part of knowledge determinants of cross-cultural performance, therefore, its criterion validity is likely not as good as SME-driven 3C SJT. I hypothesize as below.

Hypothesis 6: *SME-driven 3C SJT has stronger criterion-related validity than model-based 3C SJT.*

Chapter 7
Methodology

This chapter focused on the methodology applied for the second and third phases of my dissertation research. Two studies were conducted to examine the reliability and validity of the SJTs, to investigate their psychometric strengths and weaknesses, and to compare the utility of the two SJT development approaches.

7.1 Study 1: SJT Finalization and Validation

The goal of Study 1 was to finalize the two 3C SJTs and examine their psychometric properties (reliability, construct validity and face validity). To finalize the SJT versions, the item-total correlations of the two SJTs, the content validation results, and inter-item correlations of the model-based SJT were referenced for item reduction. Once finalized, evidence of reliability, internal structure consistency, convergent and divergent validity, and face validity evidence were investigated. Hypotheses 2-5 were examined and the data analysis results were discussed.

7.1.1 Participants

The data were collected among the undergraduate and graduate students in two universities of the United States. A total of 320 undergraduates and graduates participated in the assessment survey for the SJT finalization and validation. Their participation was rewarded with extra course credits or with the 20% chance of winning 15-dollar gift cards. In order to eliminate irresponsible responses and fatigue responses, a strict control practice was implemented to ensure data quality. First, two attention check items were used to screen out the inattentive responses. Sixty-eight participants (29.6%) failed in one of the attention check items and were removed from the dataset. The length of completion was then used to remove potentially irresponsible responses. Four datasets were created and analyzed using no time limits (*n* = 252), completion time of more than 1200 seconds (*n* = 228), completion time more than 1800 seconds (*n* = 183), and completion time more than 2100 seconds (*n* = 149). The three datasets with completion time thresholds displayed almost the same statistical properties and tendencies, while the dataset with no time control had different statistical tendencies. Therefore, completion time more than 1200 seconds was used as the second data screening criterion. Twenty-four participants who completed the assessment in less than 20 minutes were removed. As a result, a final dataset (*n* = 228, Dataset 1) was used for the SJT item reduction and validation[2]. The demographic information of the participants was demonstrated in Table 7.

Table 7 — Demographic Statistics of Datasets[3]

	Dataset 1	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**
N	228	90	65	180	136
Gender (%)					
Female	60.40	24.70	26.30	57.20	41.20
Age (%)					
Less than 20	48.70	10.10	10.8	48.9	23.50
21-25	41	74.20	73.8	41.70	61.80
26-30	6.20	2.20	0	6.70	2.90
31-35	0.40	2.20	3.1	0.60	2.20

[2] There is no an established rule on the sample size of SJT validation. The sample size decision of this study is referenced with the previous SJTs development and validation studies. Bledow and Frese (2009) used 126 participants to validate their person initial SJT, and Peus et al. (2013) used 327 participants to validate their leadership SJT.

[3] Datasets 2-5 were used in Study 2.

36-40	0.40	3.40	4.6	1.10	2.20
More than 40	2.60	6.70	6.2	0.60	7.40
Education (%)					
Undergraduate	90.31	78.50	86	92.60	79.50
Graduate	6.61	15.90	7.8	5.00	14.70
Post-graduate	1.32	5.70	6.3	0	5.90
Other	1.76	0.00	0	2.20	0
Ethnics (%)					
White	73.00	62.90	60	83.20	52.70
Black	7.96	9.00	12.3	11.70	3.70
Asian	7.96	18.00	18.5	1.10	23.50
Hispanic	6.19	5.60	4.6	2.80	10.30
Other	4.87	4.50	4.6	1.10	9.60
National Status (%)					
International students	13.70	37.10	35.4	0	47.10
Domestic students （U.S.）	86.30	62.90	64.6	100	52.90
If having oversea experience (%)					
Yes	36.57	59.56	58.46	0	100

7.1.2 Procedures

The assessment survey consisted of the two 3C SJTs and the measures used for validation purposes. The assessment survey was administered via Qualtrics.com. The two 3C SJTs were first presented to the participants in a random sequence. Then the other measures were administered in a random sequence. The participants' demographic information was collected at the end of the survey.

Statistical Package for the Social Sciences (SPSS) 23.0 and R 3.5.2 were used for the data analysis. The item-total correlations were first checked with the purpose of item reduction for both SJTs. Items with a corrected item-total correlations less than .10, were removed. For the model-based 3C SJT, the procedure of finalization proceeded with two additional steps due to its different development approach. I/O psychologists provided content validity ratings, and items with poor content validity were removed. The intercorrelations of items in their own subscales were also referenced to remove the ones which showed no significant correlations with any other items in the same subscale.

The internal structure consistency of the SME-driven SJT, the model-based SJT and subscales of the model-based SJT were examined with the value of Cronbach's alpha. The Spearman Brown correction formula (Spearman, 1910; Brown, 1910) was used to ensure the comparability between the two SJTs and between the model-based SJT and its subscales. The Spearman Brown formula allows estimation of Cronbach's alpha as if the scales were the same length. Cocron analysis was used to examine whether the Cronbach's alpha values were significantly differentiated from each other (Diedenhofen & Musch, 2016). The correlation coefficients of the two 3C SJTs and other measures were examined. The Steiger *z* test (Steiger, 1980) was used to test the difference in correlation coefficients. In discussing effect sizes, I adopted the conventional view on the magnitude of correlation, that is, *r*-values smaller than .20 are regarded as small, which serves as an indicator of divergence, *r*-values between .20 and .40 as moderate, and *r*-values larger than .40 as high. When the two measure scores are moderately or highly correlated, it suggests the two measures converge. And when their correlations are smaller than .20, they will be believed to diverge from each other.

7.1.3 Measures

Demographics. The questions on the basic demographic information include the participants' age, education level, gender, ethnicity, national origin and native language.

SME-driven 3C SJT. The original version of SME-driven 3C SJT consists of 25 scenarios and each scenario has 6-8 response options. Participants are instructed to choose their most and least likely behaviors in each scenario. The sample item is listed in Appendix B. The reliability of the finalized version was estimated with *α* (.72).

Model-based 3C SJT. The current version of model-based 3C SJT consists of 17 scenarios and each scenario has 5-8 response options. The participants are instructed to choose their most and least likely behaviors in each scenario. The sample item is listed in Appendix C. The reliability of the finalized version was estimated with *α* (.70).

Cultural Intelligence Scales (CQs). CQs was designed to evaluate "an individual's capability to function and manage effectively in culturally diverse settings" (Ang, van Dyne, Koh, Ng, Templer, Tay, & Chandrasekar, 2007, *p*. 336). It is composed of 20 items with four dimensions, four items for metacognitive CQ, six items for cognitive CQ, five items for motivational CQ and five items for behavioral CQ. The test-takers are instructed to decide how the item statement describes their capabilities by indicating how much they agree or disagree on the description in a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). Its reliability was estimated with *α* (.87).

Face validity measure. Three items were designed to assess the face validity of the 3C SJTs: (1) *It would be obvious to anyone that the test content is associated with the cross-*

cultural situations; (2) *The test assesses how well a test-taker will deal with the difficult situations in cross cultural interactions*; (3) *My performance on the test is a good indicator of my ability to deal with people from different cultures.* The test takers will be instructed to rate to what extent they agree with each statement in a 5-point Likert scale (1= *strongly disagree*, 5= *strongly agree*). The reliability was estimated with α for both SME (.66) and model based SJTs (.66).

Social desirability scale (SD). The short version of Marlowe-Crowne Social Desirability scale (13 items, Reynold, 1982) will be used to assess the social desirability. The scale uses a true/false format. The test-takers will be instructed to decide whether the statement is true or false as it describes their personality. A specific scoring key was used to tally the final score of social desirability. Its reliability was estimated with *α* (.75).

Satisfaction with Life Scale (SLS). Satisfaction with Life Scale was developed by Diener, Emmons, Larsen and Griffin (1985), which assesses people's satisfaction with their general life. The scale contains 5 statements and the test-takers are instructed to rate how they agree or disagree each statement in a 7-point Likert scale (1 = *strongly disagree*, 7= *strongly agree*). Its reliability was estimated with *α* (.77).

Satisfaction with Oversea Life Scale (SLS_Oversea). The scale was adapted from Satisfaction with Life Scale (Diener et al., 1985) by contextualizing each statement under oversea environment. For instance, the item, *The conditions of your life were excellent*, in the SLS was contextualized as *The conditions of your life in the United States were excellent* for the international students in the United States and *The conditions of your life in the foreign country were excellent* for the American students with foreign living and studying experience. SLS_Oversea keeps the same item number and the same format of SLS. Its reliability was estimated with *α* (.71).

Sociocultural Adaptation Scale (SCAS). SCAS was developed by Searle and Ward (1990) to assess people's behavioral competency of sociocultural adaptation to a foreign country. The 29-item version was used in this dissertation study. Respondents are instructed to rate the degree of difficulty they experience in a foreign culture on a 5-point Likert scale (0 = *no difficulty*, 4 = *extreme difficulty*) on items such as "*making friends*," and "*understanding the local worldview*." Unlike the other scales in this dissertation research the high score of SCAS indicated the low sociocultural adaptation. The reliability was estimated with estimated α (.89).

7.1.4 Finalization of the SME-driven 3C SJT

The item-total correlation of each item in the SME-driven 3C SJT was estimated and four items, S13, S43, S51 and S53, displayed low item-total correlations (r <.10) and were removed from the original version (Table 8). One item, S41, displayed a negative correlation with the total scale, which was removed too. As a result, the final version of SME-driven 3C SJT consists of 20 scenarios with 5-8 response options. The statistics information of the original and final versions was demonstrated in Table 9. The estimated reliability of the finalized version was .72.

Table 8 — The item-total correlations of the items of the SME-driven SJT, the items of the model-based SJT, and the items of model-based SJT subscales (Dataset 1, n=228).

SME-driven items	Corrected Item-total correlations	Cronbach's Alpha if Item Deleted	Model-based Items	Corrected Item-total correlations	Cronbach's Alpha if Item Deleted
S5	0.376	0.646	M29[a]	0.076	0.723
S7	0.254	0.657	M30	0.331	0.701
S9	0.445	0.638	M31	0.158	0.717
S11	0.377	0.646	M32	0.378	0.698
S13[a]	0.055	0.672	M33	0.306	0.708
S15	0.379	0.645	M34	0.408	0.695
S17	0.376	0.642	M35	0.44	0.689
S19	0.218	0.66	M36	0.367	0.698
S21	0.369	0.648	M37	0.377	0.697
S23	0.131	0.672	M38	0.223	0.712
S25	0.347	0.645	M39	0.409	0.695
S27	0.392	0.643	M40	0.37	0.698
S29	0.283	0.654	M41	0.381	0.697
S31	0.117	0.669	M42	0.255	0.709
S33	0.204	0.661	M43	0.205	0.715
S35	0.292	0.653	M44	0.299	0.705
S37	0.155	0.666	M45	0.215	0.713
S39	0.156	0.665			
S41[a]	-0.145	0.699			
S43[a]	0.027	0.675			
S45	0.219	0.66			
S47	0.356	0.648			
S49	0.176	0.663			
S51[a]	0.085	0.673			
S53[a]	0.078	0.672			

Note. [a] indicates that the item was removed from the final versions of 3C SJTs.

Table 9 — Means, Standard Deviations and Reliabilities of the Original and Final Versions of SME-driven SJT (Dataset 1, n=228).

Versions	Item Number	Item removed	*M*	*SD*	Reliability
S_0	25	N.A.	17.30	7.04	.67
S_f	20	S13, S41, S43, S51, S53	15.19	6.73	.72

Note. S_0 refers to the original version of the SME-driven 3C SJT, and S_f refers to the finalized version.

7.1.5 Finalization of the model-based 3C SJT

The item-total correlations were investigated for each model-based 3C SJT items and only one item, M29, was dropped from the original SJT (Table 9). Next, the content validity of each item was investigated. Three professors in I/O psychology judged what construct each item taps on independently and interrater agreement was calculated. Table 10 notes the constructs the items were originally designed to measure and the content rating results from the three experts. One item, M36, was dropped due to the interrater disagreement. Finally, the interrelations among items in the same subscales were examined. One item (M38) was removed because it appeared no significant correlation with any items in its subscale. The final version of the model-based 3C SJT consisted of 14 scenarios with 6-8 response options respectively. The statistical information of each version of the model-based 3C was shown in Table 12.

Table 10 — The content validation results of Model-based SJT, and the decision of the construct targeted by each SJT items according to content validity

Model-based Items	Originally target construct	Rater 1	Rater 2	Rater 3	Final target construct
M29	Inq	Inq	Inq	Inq	Inq
M30	Inq	Inq	Inq	Other	Inq
M31[a]	Inq	Md	Md	Md	Md
M32	Se	Se	Se	Se	Se
M33[a]	Se	Is	Is	Is	Is
M34	Es	Es	Es	Es	Es
M35[a]	Se	Inq	Inq	Inq	Inq
M36[b]	Is	Md	Is	Se	--
M37	Es	Es	Es	Es	Es
M38	Md	Md	Md	Md	Md
M39	Es	Es	Es	Is	Es
M40	Is	Is	Inq	Inq	Is
M41	Is	Is	Is	Is	Is

M42	Md	Md	Md	Is	Md
M43	Es	Es	Es	Es	Es
M44 [a]	Md	Se	Se	Is	Se
M45	Md	Md	Other	Md	Md

Note. Inq=inquisitiveness, Md=mindfulness, Se=self-efficacy, Is=interpersonal skill, Es=emotional stability, and other=other unlisted constructs. [a] The target construct of the item was shifted according to the content validation results. [b] indicates the item was removed.

Table 11 — Means, Standard Deviations and Reliabilities of the Versions of the Model-based SJT (Dataset 1, n=228)

Versions	Item Number	Removed item	*M*	*SD*	Reliability
M_0	17	N.A.	12.04	6.33	.72
M_1	16	M29	11.54	6.23	.72
M_2	16	M36	11.14	5.94	.70
M_3	15	M29, M36	10.64	5.84	.71
M_f	14	M29, M36, M38	9.79	5.62	.70

Note. M_0 refers to the original version of Model-based 3C SJT, and M_f refers to the finalized version.

7.1.6 Results and discussion

The reliability of the SME-driven 3C SJT, the model-based 3C SJT and its five subscales were computed. The reliabilities of all of the SJT scales were within the conventionally acceptable limit (α > 70); the model-based subscales with fewer items showed lower levels of reliability. The reliability estimates, the adjusted reliabilities of the model-based SJT and its subscales, and the correlations of the scales were listed in Table 12. Overall, both SJTs and the SJT subscales demonstrated acceptable internal consistency. The model-based SJT showed higher internal consistency than the SME-driven SJT with Spearman-Brown formula adjustment, however, the Cocron test showed the difference was not significant (t = 1.73, df = 226, p = 0.08). Hypothesis 2, which predicted the model-based SJT has higher internal consistency than SME-driven SJT, was not supported.

I hypothesized that the model-based SJT subscales should have higher internal consistency than the overall SJT. Most of the model-based SJT subscales demonstrated higher internal consistency than the overall SJT after correction, except the Interpersonal Skill subscale. However, only the Self-efficacy and the Emotional Stability subscales demonstrated significant advantage over the overall SJT according to the Cocron test (see Table 13). Thus, Hypothesis 3 was partially supported.

The SME-driven SJT was hypothesized as being less susceptible to social desirability than the model-based SJT (Hypothesis 4). A moderate correlation was found between the scores of the SME-driven SJT and Social Desirability Scale ($r = .25$, $p < .01$), while a large correlation was found between the model-based SJT and Social Desirability Scale ($r = .43$, $p < .01$). The Steiger z test showed that the correlation coefficient of SME-driven SJT with SDS was significantly lower than the one of model-based SJT with SDS ($z = -2.11$, $p < .05$)[4], therefore, Hypothesis 4 was supported.

[4] I used the online calculator: Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common. *Computer software*. Available from http://quantpsy.org.

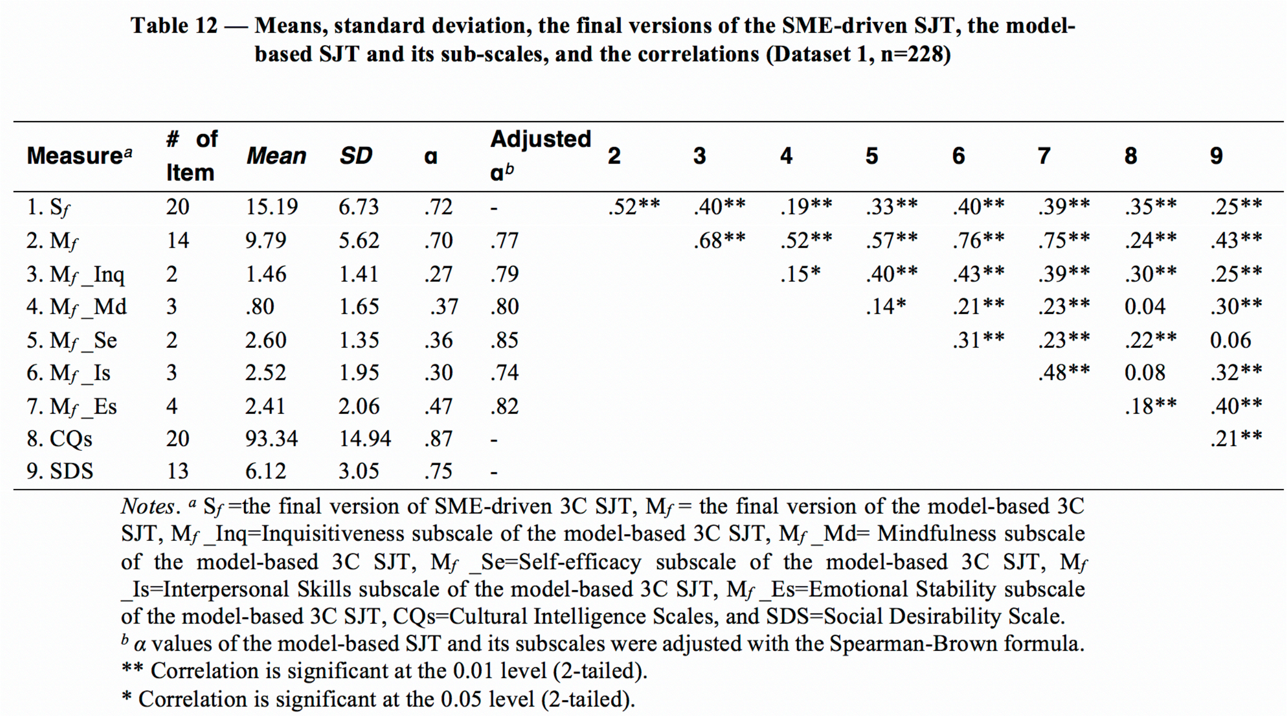

Table 12 — Means, standard deviation, the final versions of the SME-driven SJT, the model-based SJT and its sub-scales, and the correlations (Dataset 1, n=228)

Measure[a]	# of Item	*Mean*	*SD*	α	Adjusted α[b]	2	3	4	5	6	7	8	9
1. S_f	20	15.19	6.73	.72	-	.52**	.40**	.19**	.33**	.40**	.39**	.35**	.25**
2. M_f	14	9.79	5.62	.70	.77		.68**	.52**	.57**	.76**	.75**	.24**	.43**
3. M_f_Inq	2	1.46	1.41	.27	.79			.15*	.40**	.43**	.39**	.30**	.25**
4. M_f_Md	3	.80	1.65	.37	.80				.14*	.21**	.23**	0.04	.30**
5. M_f_Se	2	2.60	1.35	.36	.85					.31**	.23**	.22**	0.06
6. M_f_Is	3	2.52	1.95	.30	.74						.48**	0.08	.32**
7. M_f_Es	4	2.41	2.06	.47	.82							.18**	.40**
8. CQs	20	93.34	14.94	.87	-								.21**
9. SDS	13	6.12	3.05	.75	-								

Notes. [a] S_f =the final version of SME-driven 3C SJT, M_f = the final version of the model-based 3C SJT, M_f _Inq=Inquisitiveness subscale of the model-based 3C SJT, M_f _Md= Mindfulness subscale of the model-based 3C SJT, M_f _Se=Self-efficacy subscale of the model-based 3C SJT, M_f _Is=Interpersonal Skills subscale of the model-based 3C SJT, M_f _Es=Emotional Stability subscale of the model-based 3C SJT, CQs=Cultural Intelligence Scales, and SDS=Social Desirability Scale.

[b] α values of the model-based SJT and its subscales were adjusted with the Spearman-Brown formula.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 13 — The results of the Cocron test of the reliability comparison between two SJTs and between the model-based SJT and its subscales

Comparison	***t***	***df***	***p***
M_f *vs.* S_f	1.73	226	.08
M_f _Inq *vs.* M_f	.93	226	.35
M_f _Md *vs.* M_f	1.23	226	.22
M_f _Se *vs.* M_f	3.94**	226	.00
M_f _Is *vs.* M_f	1.42	226	.16
M_f _Es *vs.* M_f	2.79**	226	.01

The high correlation between the SME-driven SJT and the model-based SJT (*r* = .52, *p* < .01) demonstrated the convergence between the two SJTs. Both of the SJTs also converged with CQs as judged by their moderate correlations (r_{S} = .35, r_{M} = .24, *p* < .01). The moderate to large correlations with Social Desirability Scale suggested that neither 3C SJT diverged from Social Desirability Scale very well. However, the psychological properties of both 3C SJTs were still well supported with their internal consistency, content validity evidence, and convergent evidence.

The face validity of the two SJTs was examined and four data cases with missing data in the face validation were removed. The SME SJT demonstrated significant higher face validity than the model-based SJT (M_{SME} = 12.07, M_{Model} = 11.67, *t* = -2.27, *p* < .05), which was consistent with Hypothesis 5. Statistic information can be found in Table 14.

Table 14 — Means, standard deviation, and the t-test results of face validity differences of the SME-driven SJT and the model-based SJT (Dataset 1)

SJTs	N[5]	M	*SD*	Difference	*SE*	*t*	*df*	*p*	95%CI
SME-driven SJT	226	12.07	1.92						
Model-based SJT	224	11.67	1.82	-0.40	0.18	-2.27	449	0.02	-.75, -.05

7.2 Study 2: Replication, Validation and Hypothesis Testing

Study 2 attempted to replicate and expand on Study 1 by presenting criterion validity evidence for the 3C SJTs.

7.2.1 Participants

A total of 121 undergraduates and graduates participated in Study 2, and all of them were engaged in multicultural teamwork. Their participation was rewarded with extra course credits or a 20% chance to win a 15-dollar gift card. Their teammates were also invited to rate their performance in the team project(s). With the strict data quality controls, 31 data cases were removed because of failure in either attention check or short completion time (the same criteria from Study 1). The dataset (*n* = 90, Dataset 2) was used for the next step of data analysis. In Dataset 2, sixty-five participants had peer review scores and a dataset (*n* = 65, Dataset 3) was extracted to examine the 3C SJTs' predictability of the individuals' actual multicultural team performance.

In both Study 1 and Study 2, the general life satisfaction data (*n* = 180, Dataset 4) and the oversea life satisfaction and sociocultural adaptation data (*n* = 136, Dataset 5) have

[5] Missing data were found in the face validity survey in Dataset 1, and the cases with missing data were removed.

been collected. The two types of data were extracted from Dataset 1 and Dataset 2, combined and formed Dataset 4 and 5 for the validity study. The demographic information of the participants in the four datasets was displayed in Table 8 in Section 7.1.

7.2.2 Procedures

The data were collected from the participants and their teammates respectively via Qualtrics.com. The same assessment survey in Study 1 was used to collect the data from the participants. Additionally, the participants were asked to provide their name and their teammates' names at the end of the survey. A team performance evaluation link was then sent to the teammates of each participant in the same email. In the email the teammates of each participant were invited to rate the participant's team performance in an objective manner, and the peer review evaluation was anonymous. The second and the third round of emails were sent out as reminders if the team performance of the participants wasn't rated. The reliability of both 3C SJTs were examined using the finalized versions of the two SJTs in Study 1. The reliability of the SME-driven SJT, and the adjusted reliability of the model-based SJT and its five subscales were compared with the Cocron analysis for testing Hypotheses 2 and 3. The convergent and divergent validity evidence of both 3C SJTs were accumulated by examining the relationship of the two 3C SJTs with CQs, social desirability and general life satisfaction.

To test Hypothesis 4, the correlations of the two SJTs and social desirability were compared. The Steiger *z* test was used for estimating whether these correlations differed in a significant manner. The face validity scores of the SJTs were calculated and the means were compared to examine Hypothesis 5. With regression analysis, criterion validity evidence of the SJTs CQs were accumulated by estimating their relationships with satisfaction with oversea life, sociocultural adaptation, and actual teamwork performance in a multicultural team. To test Hypotheses 1 and 5 the Steiger *z* test was also used for comparisons among the predictive coefficients.

Face validity of both SJTs were examined and the means were compared. The readability of each item of the two SJTs was assessed with Word Readability Statistic Tool, and then a *t*-test was used to investigate if the two SJTs differed in readability.

7.2.3 Measures

The same assessment survey was used in Study 2 as in Study 1 with additional questions about the participant’s name and their teammates’ names.

SME-driven 3C SJT. The finalized version of the SME-driven 3C SJT was used in Study 2, which consists of 20 scenarios and each scenario has 6-8 response options. Participants are instructed to choose their most and least likely behaviors in each scenario. The reliability of the finalized version was estimated at .77.

Model-based 3C SJT. The finalized version of the model-based 3C SJT was used in Study 2, which consists of 14 scenarios and each scenario has 5-8 response options. The participants are instructed to choose their most and least likely behaviors in each scenario. The reliability of the finalized version was estimated with .77.

Cultural Intelligence Scales (CQs). See Study 1. The reliability was estimated at .89.

Face validity measure. See Study 1. The reliability was estimated at .65 for the SME-driven 3C SJT and at .56 the model-based 3C SJT.

Social desirability scale (SD). See Study 1. The reliability was estimated as .73.

Satisfaction with Life Scale (SLS). See Study 1. The reliability was estimated as .74.

Satisfaction with Oversea Life Scale (SLS_Oversea). See Study 1. The reliability was estimated as .67.

Sociocultural Adaptation Scale (SCAS). See Study 1. The reliability was estimated as .89.

Comprehensive Assessment of Team Member Effectiveness (CATME). CATME was developed by Loughry and Ohland (2007) to assess how effectively a team member contributes to the team via five dimensions: contributing to the team's work, interacting with teammates, keeping the team on track, expecting quality and having relevant knowledge, skills and abilities. It is a behaviorally anchored scale, and each of the five dimensions are rated on a 5-point Likert scale. 1 indicates poor team performance while 5 indicates high team performance. The reliability was estimated with Cronbach α-value (.92).

7.2.4 Results and discussion

The acceptable reliabilities of the two SJTs (α_{SME} = .77 and α_{Model} = .77) contributed additional evidence to the internal consistency of the two SJTs (Table 15). Although the model-based SJT showed higher α-value than the SME-driven SJT, the difference was not significant with the Cocron test (t = 1.60, df = 88, p = .11), therefore, Hypothesis 2 was not supported. The model-based SJT subscales, except Self-efficacy scale, demonstrated higher internal consistency than the overall scale after corrections, but the Cocron test only found significant differences for the Self-efficacy, the Interpersonal Skill and the Emotional Stability scales (see Table 16). Therefore, Hypothesis 3 was only partially supported.

The correlations of the SJTs and the subscale SJTs with CQs and SDS were examined, which demonstrated the same tendency as in Study 1, although the magnitudes of correlation varied to some extent (Table 18). The two SJTs were highly correlated (r = .46, p < .01), and both of them were moderately related with CQs (r_S = .30, r_M = .35, p < .01). Both SJTs were significantly correlated with SDS while the correlation between the SME-driven SJT and SDS increased compared with Study 1. The SME-driven SJT (r_S = .38, p < .01) had a lower correlation with SDS than the model-based SJT (r_M = .45, p < .01), but the Steger z test showed the difference was not significant (z = -.71, p = .23). Therefore, Hypothesis 4 was not supported.

The correlations of the two 3C SJTs and CQS were also examined with the score of SLS, which were supposed to be slight. The SME-driven 3C SJT diverged from general life satisfaction (r = .10, *n.s.*), and the model-based 3C SJT and CQs also demonstrated adequate divergence from general life satisfaction with small correlations (rs = .19 and .18, respectively, p < .05). This evidence supports the discriminant validity of the two SJTs, indicating that they do not measure general life satisfaction. The data analysis results were shown in Table 17.

Table 15 — Means, standard deviation of the SME-driven SJT, the model-based SJT and its sub-scales, CQS and SDS, and the correlations (Dataset 2, n=90)

Measure[a]	# of Item	*Mean*	*SD*	α	Adjusted α[b]	2	3	4	5	6	7	8	9
1. S_f	20	13.87	7.57	.77	-	.46**	.21*	.27*	.35**	.36**	.40**	.30**	.38**
2. M_f	14	9.39	6.41	.77	.83		.68**	.59**	.62**	.76**	.81**	.35**	.45**
3. M_f_Inq	2	1.44	1.39	.41	.87			0.15	.42**	.44**	.48**	.24*	0.19
4. M_f_Md	3	.76	1.70	.43	.84				.22*	.30**	.37**	.23*	.42**
5. M_f_Se	2	2.44	1.27	.21	.73					.36**	.39**	.28**	.27**
6. M_f_Is	3	2.16	2.29	.53	.88						.43**	.28**	.34**
7. M_f_Es	4	2.61	2.40	.60	.88							.22*	.31**
8. CQS	20	96.46	16.05	.89	-								.22*
9. SDS	13	6.25	3.06	.73	-								

Notes. [a] S_f =the final version of SME-driven 3C SJT, M_f = the final version of the model-based 3C SJT, M_f_Inq=Inquisitiveness subscale of the model-based 3C SJT, M_f_Md= Mindfulness subscale of the model-based 3C SJT, M_f _Se=Self-efficacy subscale of the model-based 3C SJT, M_f _Is=Interpersonal Skills subscale of the model-based 3C SJT, M_f_Es=Emotional Stability subscale of the model-based 3C SJT, CQs=Cultural Intelligence Scales, and SDS=Social Desirability Scale.

[b] α values of the model-based SJT and its subscales were adjusted with the Spearman-Brown formula.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 16 — The results of the Cocron test of the reliability comparison between the model-based SJT and its subscales

Comparison	*t*	*df*	*p*
M_f *vs.* S_f	1.60	88	.11
M_f_Inq *vs.* M_f	1.72	88	.09
M_f_Md *vs.* M_f	.35	88	.73
M_f_Se *vs.* M_f	2.79**	88	.01
M_f_Is *vs.* M_f	2.53**	88	.01
M_f_Es *vs.* M_f	2.80**	88	.01

Table 17 — Means, standard deviation, the SME-driven SJT, the model-based SJT and its sub-scales, CQS and the SLS, and the correlations (Dataset 4, n=180)

	Mean	*SD*	α	1	2	3
1. S_f	14.84	6.67	.71			
2. M_f	9.46	5.76	.73	.50**		
3. CQ	89.29	14.41	.86	.32**	.27**	
4. SLS	24.84	5.44	.77	.10	.19*	.18*

Notes. ** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 18 — Means, standard deviation, the SME-driven SJT, the model-based SJT and its sub-scales, CQS and the LS_Overseas, and the correlations (Dataset 5, n=136)

	Mean	*SD*	α	1	2	3	4
1. S_f	15.00	7.24	.77				
2. M_f	9.93	6.01	.73	.51**			
3. CQ	101.03	13.71	.85	.28**	.27**		
4. SCAS	50.49	13.19	.89	-.20*	-.21*	-.24**	
5. LS_Overseas	25.69	5.07	.71	.24**	.18*	.26**	-.36**

Notes. ** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

The statistical information and the correlations of the variables in Dataset 5 were listed in Table 18. Regression analysis indicated that both 3C SJTs and CQs were predictive of individuals' satisfaction with oversea life ($\beta_{\text{SME}} = .24$, $p < .01$; $\beta_{\text{model}} = .18$, $p < .05$; $\beta_{\text{CQS}} = .26$, $p < .01$) and their sociocultural adaptation to foreign countries ($\beta_{\text{SME}} = -.20$, $p < .05$; $\beta_{\text{model}} = -.21$, $p < .05$; $\beta_{\text{CQS}} = -.24$, $p < .01$) (Table 19). The Steiger *z* tests indicated no significant differences among the three measures in predicting oversea life satisfaction ($z_{\text{S-M}} = .72$, $p = .23$; $z_{\text{S-C}} = -.20$, $p = .42$; $z_{\text{M-C}} = -.79$, $p = .21$). Results with sociocultural adaptation were similar ($z_{\text{S-M}} = .12$, $p = .45$; $z_{\text{S-C}} = .40$, $p = .34$; $z_{\text{M-C}} = .30$, $p = .38$).

Table 19 — Regression of the prediction of the SME-driven SJT, the model-based SJT and CQs

Model	DV	IV	*N*	R	Std. Error	F	*β*	*se*	*t*
1	SCAS	SME_SJT	136	.20	12.98	5.32*	-.20	.15	-2.31*
2	SCAS	Model_SJT	136	.21	12.94	6.14*	-.21	.19	-2.48*
3	SCAS	CQ	136	.24	12.84	8.50**	-.24	.08	-2.92**
4	LS_Overseas	SME_SJT	136	.24	4.94	8.21	.24	.06	2.87**
5	LS_Overseas	Model_SJT	136	.18	5.00	4.63	.18	.07	2.15*
6	LS_Overseas	CQ	136	.27	4.91	10.00	.26	.03	3.16**
7	CATME	SME_SJT	65	.26	2.90	4.65	.26	.05	2.16*
8	CATME	Model_SJT	65	.04	3.01	.11	-.04	.06	-.33
9	CATME	CQ	65	.01	3.01	.00	.01	.02	.06

Notes. [a] Dataset 5 was used for Model1-6 regression analysis (n=136), and Dataset 3 was used for Model 7-9 regression analysis (n=65). ** p<.01; * p<.05.

Only the SME-driven 3C SJT significantly predicted the actual multicultural team performance rated by the team members ($\beta_{\text{SME}} = .26$, $p < .05$; $\beta_{\text{model}} = -.04$, *n.s.*; $\beta_{\text{CQs}} = .01$, *n.s.*). The statistical information and the correlations of the variables of Dataset 3 were shown in Table 20. The Steiger *z* tests confirmed that the SME-driven SJT outperformed the other two measures in predicting the peer reviewed team performance ($z_{\text{S-M}} = 2.04$, $p < .05$; $z_{\text{S-C}} = 1.57$, $p = .058$). In sum, the SME-driven SJTs displayed a stronger criterion validity than the model-based SJTs, which supports Hypothesis 6. The SME-driven SJT also outperformed CQs in predicting peer-reviewed performance in the multicultural teamwork while the model-based SJT displayed a similar criterion-validity with CQs ($z_{\text{M-C}} = -.32$, $p = .37$), hence, Hypothesis 1 was partially supported.

Table 20 — Means, standard deviation, the SME-driven SJT, the model-based SJT and its sub-scales, CQS and the performance in multicultural team, and the correlations (Dataset 3, n=65)

	Mean	*SD*	α	1	2	3
1. S_f	15.00	7.24	.77			
2. M_f	9.93	6.01	.73	.30*		
3. CQS	101.03	13.71	.85	0.14	.34**	
4. CATME	21.53	2.98	.92	.26*	-0.04	0.01

Notes. ** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

In addition, the two SJTs were compared in terms of face validity and readability. The SME-driven SJT displayed a higher mean than the model-based SJT, although the difference was not significant (Table 21). The readability of each SJT item was examined in terms of Flesch reading ease, and the readability scores were displayed in Appendix D. The 3C SME-driven SJT appeared more difficult to read than the model-based 3C SJT ($t = -2.19$, $p < .05$) (Table 22).

Table 21 — Means, standard deviation, and the t-test results of face validity differences of the SME-driven SJT and the model-based SJT (Dataset 2, n=90)

SJTs	N[a]	*Mean*	*SD*	Difference	*SE*	*t*	*df*	*p*	95%CI
SME-driven SJT	89	12.08	1.95						
Model-based SJT	89	11.81	1.81	-0.27	0.28	-.96	176	*n.s.*	-.83, .29

Note. [a] The cases with missing data were removed for face validity analysis.

Table 22 — T-test statistics of the readability of the SME-driven SJT and the model-based SJT

SJTs	N of items	*Mean*	*SD*	*df*	*t*	*p*
SME-driven SJT	20	72.49	6.89	19		
Model-based SJT	16	78.12	8.10	15	-2.19	.02

7.3 Summary of Study 1 and Study 2

The results of Study 1 were largely replicated in Study 2 (Table 24). Hypothesis 2 was not supported by the two studies but they both partially supported Hypothesis 3. Hypothesis 5 was fully supported by Study 1 while partially supported by Study 2. The only inconsistency happened in Hypothesis 4 which was only supported by Study 1 although the result of Study 2 showed the same direction but the difference was not statistically significant. Overall, both SJTs displayed acceptable psychometric properties. The model-based SJT consistently showed higher internal consistency than the SME-driven SJTs. The SME-driven SJT was less susceptible to social desirability than the model-based SJT. The face validity scores of the SME-driven SJT were higher than those of the model-based SJT.

Table 23 — Summary of hypotheses testing results of Study 1 and Study 2

Hypotheses	Study 1	Study 2
1	N.A.	Partially supported
2	Not supported	Not supported
3	Partially supported	Partially supported
4	Supported	Not supported
5	Supported	Partially supported
6	N.A.	Supported

The criterion validity of the two SJTs were only examined in Study 2 (Hypotheses 1 and 6). The two 3C SJTs shown similar criterion validity with CQS when the criteria were assessed by self-report measures, that is, self-reported satisfaction with oversea life, and adaptation to foreign cultures and societies. However, the SME-driven 3C SJT shown higher predictability than the model-based 3C SJT and CQS when the criterion was assessed by others, the multicultural teamwork performance. There is no significant difference of criterion validity between the model-based SJT and CQs.

Chapter 8
General Discussion

This chapter discusses about the findings of the dissertation research and summarizes the answers to the dissertation research questions.

8.1 The Findings of the Dissertation Research

This dissertation research developed and evaluated alternative assessment tools to replace the self-report Likert scales which have been dominant in 3C assessment despite their flaws. Considering that 3C is a capability specific to context, that is, cross cultural communication and interaction, I believe that contextualized measurement methods are more proper to assess 3C rather than the decontextualized Likert scale. The decision of developing paper-pen SJTs rather than other contextualized methods, like assessment center (AC) and high-tech simulations, came from the trade-offs between the psychological properties and the development and implement costs. The cost of developing and implementing a paper-pen SJT is much lower than AC, video SJTs, and other types of high-tech simulations; meanwhile, no evidence has indicated that AC, video SJTs, or high-tech simulations outperform paper-pen SJTs. Also, considering a paper-pen SJT is relatively easy to adapt, I developed two paper-pen 3C SJTs. I expected these SJTs would perform better than the traditional Likert scales in assessing individuals' 3C and predicting their actual performance in cross cultural situations.

The results of this dissertation research have indicated that the SME-driven 3C SJT performed better in predicting individuals' multicultural teamwork performance than CQS, a renowned 3C self-report Likert scale. Although CQS has demonstrated good criterion validity across a large number of studies, most of those studies have used other self-report Likert scales when assessing CQS criterion validity. In this dissertation CQS failed to display any correlation with performance rated by team-mates. It may be that the CQS's criterion validity is elevated due to the common method bias. In contrast, the SME-driven 3C SJT showed higher criterion validity in evaluating 3C regardless of the rating entities. Specifically, stable and similar correlations of the SME-driven SJT were found with both self-report and other-rating performance.

In addition, this dissertation examined the psychometric properties of SJTs developed by the SME-driven approach and the model-based approach, and compared the utility of the two development approaches. The utility of the development approach was assessed by the development costs indexed with development time, momentary cost, and human resources. The utility index values and the psychometric properties of the two SJTs were listed in Table 24. A detailed time log, the momentary cost and the human resource information were presented in Appendix E, Appendix F and Appendix G respectively.

Table 24 — The utility comparison between the two SJT development approaches

	SME-Driven	Model-based
Development Cost		
Time (minute)	9405	2200[6]
Momentary cost (dollar)	285	60
Human resource (person)	99	76
Psychometric properties		
Reliability[7]	.72 ~ .77	.77 ~ .83
Internal consistency	Acceptable	Higher, but not statistically significant, than the SME-driven SJT
Convergent validity	Moderately correlated with CQS	Moderately correlated with CQS
Discriminant validity	Not correlated with SLS; Minimally discriminant against SD	Slightly correlated with SLS; Not discriminant against SD
Predictive validity	Predictive of SCAS, SLS_overseas, and peer-review multicultural team performance	Predictive of SCAS and SLS_overseas; Not predictive of peer-review multicultural team performance
Utility	The SME-driven approach outperforms the model-based approach.	

Compared with the model-based approach, the SME-driven approach explicitly requires more time, momentary cost, and SMEs. A large amount of time was consumed to validate the content of the two SJTs in a different way. For the SME-driven approach, most time was used to interview SMEs, transcribe interviews, code and recode interview content, categorize the content into themes, write scenarios, test scenario typicality, and generate

[6] The model used the model-based SJT was adapted from my master's 3C model. The time cost didn't include the time spent on the original model development.

[7] The reliability is indexed with the adjusted α with the Spearman-Brown formula.

group response options. For the model-based approach, a high proportion of time was used for writing and editing the scenarios and response options. A similarly high proportion of time was consumed in examining whether each SJT item taps the specific construct of interest, and notably the ad-hoc statistic test for item reduction consumed more time than with the SME-driven approach, which are beyond my expectations.

Noteworthily, the time spent by the model-based approach in this dissertation is largely discounted because I adapted, instead of building, a 3C model from my previous master's thesis research (Chen, 2017). The development of the original model consumed a large amount of time, but I didn't count in this amount of time due to two reasons. First, it is not easy to make an accurate estimate, or even a close one, on a literature review endeavor conducted two years ago. Any inaccurate estimate may communicate wrong information to readers, so I prefer to avoid such a miscommunication in this scientific research. Second, the time of building a model for SJT largely depends on the level of the expertise and familiarity the developers have on the specific constructs or competencies. If the developers are the experts on the specific constructs/competencies and have accumulated adequate knowledge in the relevant field, like I have on 3C, they may not need a lot of time to conduct a literature review in order to build a sound model. However, if the developers lacks adequate knowledge, they will need large amount of time to do a comprehensive literature review and other information collection before building a model. Apparently, the actual amount of time for model building varies across the SJT developers. When the developers decide to use the model-based approach, they are recommended to reflect beforehand on how well they know the specific constructs/competencies and on how much they have already accumulated the knowledge and information. Without earlier knowledge accumulation, it will likely consume more time to develop SJTs with the model-based approach than the SME-driven approach.

Very limited amounts of money were invested in developing both 3C SJTs mainly because of my student status. As a student I received voluntary help from my professors, program peers, and colleagues in the I/O psychology program. They volunteered as SMEs in different phases throughout the whole SJT development procedure. The voluntary work from a total of sixteen SMEs saved me a large amount of money, and their undoubted expertise in I/O psychology and long-term experience in cross cultural interactions were indispensable sources for my SJT scenarios and response options development. Most momentary cost was for the payment for outside SMEs that I recruited to generate the SME-driven SJT scenarios, and the other part of cost was for the rewards for the validation study participation (Study 2). It is not hard to predict if the SME-driven approach will cost much more money to recruit SMEs for scenarios and response option generalization than the model-based approach with which both scenarios and response options are supposed to be written by the developers.

In the current study more SMEs were required for the SME-driven approach; however, the difference was not as big as previously assumed. Unlike the name implies, the model-based approach actually requires SME involvement in most development phases except scenario generation and response option writing. Groups of SMEs are in needed to examine the content validity of SJT items, that is, to check if each item measures the construct/competency of interest. SMEs are also needed to investigate whether the model-based response options fall into the scope of possible responses in reality, and whether those response options cover the range of competency levels from incompetent to competent. Differentially, a group of SMEs is required to generate the SME-driven SJT scenarios. The number of the required SMEs is not fixed, which depends on information saturation. In this dissertation, the information got saturated in the eighteenth interviewee, but the interview proceeded until the twenty-third to guarantee this saturation. Different groups of SMEs are invited to rate the typicality of each SME-driven scenario and to generate the response options respectively.

Both SJTs demonstrated acceptable reliabilities in terms of Cronbach α. Although the model-based SJT showed higher α-values than the SME-driven SJT, these differences were not significant. Similarly, higher internal consistence was found in the model-based SJT with a small effect size. Both SJTs converged to cultural intelligence and diverged from life satisfaction in a similar way, but the SME-driven SJT outperformed the model-based SJT in discriminating against social desirability. Both SJTs showed similar predictivity of foreign life satisfaction and sociocultural adaptation to a foreign country, while only SME-driven SJT predicted the actual performance in teamwork rated by others. Overall, the SJT developed with the SME-driven approach manifests a higher validity than the one with the model-based approach.

8.2 Answers to the Research Questions

The findings of this dissertation research have indicated that the SME-driven SJTs has better psychometric properties than the model-based SJTs, which is highlighted with its enhanced criterion validity (Research Question 1). It is not proper to make an assertation on which SJT development approach generally has a higher utility than the other because the development cost of the model-based SJT partially depends on the developers' expertise and experience while the cost of the SME-driven SJT partially depends on how many SMEs are needed to get information saturation. The developers' expertise and experience can reduce the model-based development cost considerably, and the fewer number of SMEs needed for achieving information saturation can also lower the cost of the SME-driven approach. Worthy to mention, this dissertation uncovered two potential costs which are ignored by those who advocate for the model-based approach. One potential cost is that at least two groups of SMEs are needed in developing a qualified model-based SJTs. One group of SMEs is to judge the construct or the competency the SJT measures, and another group is to rate the competence level that each response option targets. The second potential cost is that the model-based SJT would likely consume much more time during the item reduction phase than the SME-driven SJT. The item removal decision is much easier to make for the SME-driven SJT, which is based only on the corrected item-total correlations, while for the model-based SJT, each item removal decision requires not only examining the corrected item-total correlation, but also need reassess the content validity of each subscale as well as intercorrelations among subscale items.

When it comes to this dissertation, no big discrepancy was found in the development costs or psychometric properties between the two approaches except time consumption and predictive validity. I believe that predictability should be prioritized when evaluating assessment tools when the cost is comparable. Also, considering the fact that it is not the approach per se saving the development time, instead, much time was saved due to adapting an existent model, therefore, I believe that the SME-driven approach has a higher utility than the model-based approach (Research Question 2).

Chapter 9
Implications, Future Research and Limitations

In the last chapter, the SJT development issues and the implications are discussed based on my reflections when using the two approaches for SJT development. Followed is the future research discussion. The limitations of this dissertation research are discussed at the end.

9.1 Implications and Future Research

9.1.1 About the SME-driven approach

Most existent SJTs were developed by the traditional SME-driven approach, and were repeatedly criticized for their low internal consistency and problematic reliability (Lievens et al., 2008; Whetzel & McDaniel, 2009; Patterson et al., 2012). These researchers consensually attributed the two issues to the heterogeneous nature of SJTs and improperly using Cronbach α as the estimate index. Although those statements make sense, I found that a structure interview with well-designed questions would solve the issues to large extent. The scenario interview should be structured, the scope of the situations of interest should be settled down beforehand, and the interview questions should directly target, and only target, the situations within the scope. The restricted questions would delimit the critical incidents provided by the interviewees in the specific scope, which may reduce, although not remove, heterogeneity for the SJT. With this approach the SME-driven 3C SJT displayed acceptable reliability and internal consistency similar with the model-based SJT, which implies that well-designed interview questions and the interview procedure may decrease the variance of internal consistency among SJTs, ranging from .49 to .98 (Lievens et al., 2008).

The impact of SME sources on SJT validity was addressed by researchers (Weekly et al., 2006; Whetzel & McDaniel, 2009), but no studies have specified such an impact so far. SMEs refers to the individuals with experience and KSAOs[8] in a specific job or field. The SJT development requires the diversity of SMEs from low performers with no or very limited experience and KSAOs to high performers with plenty of experience and KSAOs. To be noted, the diverse SMEs can't be used indiscriminately during the SJT development; instead, differently competent performers should be used rationally in accordance with the requirements specific to each development phase. For the 3C SMEs-driven scenario development, adequate cross cultural experience is the key criterion to select SMEs. In this phase the task of SMEs is to provide critical incidents, and involvement in cross cultural communication and interactions ensures the incidents are critical. The level of cross cultural performance is not the major concern at this stage. The same type, but a different group, of SMEs is needed for testing the typicality of each SME-driven scenario. In this phase of the response option development, the ideal SME group should include high, medium, and low performers with the purpose that the responses options cover the full range of good, bad, and either good or bad reactions.

[8] KSAO refers to knowledge, skills, abilities/attitudes and others.

For the scoring key development, the qualified SMEs should be high performers with plenty of cross cultural experience, and their expertise and experience enable them to make a good judgement on the best and the worst options. Using the wrong SMEs may be detrimental to the quality and validity of the final SJT. For instance, if the inexperienced SMEs were recruited for the scenario interview, the developers are less likely to obtain appropriate critical incidents for SJT scenarios, and the content validity will be attenuated. It is also less likely for an SME group with a certain level of performance to generate the responses options to cover the full spectrum of reactions. If the low or medium performers are less likely to make correct judgment on the best and worst response options, the scoring keys developed by less competent performers will not function and the criterion validity could be impaired. These implications of SME selection come from my experience in developing the SME-driven 3C SJT, and empirical studies are in need to justify this practice of selecting SMEs in the future.

An interesting phenomenon appeared when I collected the responses to SME-driven scenarios via the open-end survey. The SMEs were required to write down their reactions to the situations described by the scenarios and describe the reason for their reactions. The same or similar responding actions were found triggered by various reasons, and those reasons seemed to indicate differential cross cultural competence levels. For instance, the target action to a scenario was to find another person to help communication with a foreign student. Two SMEs provided different reasons for this action: one SME tried to complete the communication as soon as possible with extra help while the other was to find extra help for better understanding and learning purpose. Obviously, the first reason reveals impatience and withdrawal attitude when facing cross cultural challenges while the second reason indicates the opposite, more proactive and inquisitive attitude to such challenges. This finding hints at an alternative way to frame the response options by including the underlying reason. The future research is suggested to explore the nature and characteristics of the response options with the reason included.

9.1.2 About the model-based approach

The developers own more autonomy in writing and designing the model-based SJT scenarios and response options. How to write the scenario question and response options can change the construct the SJT item taps. The developers should be mindful of what an SJT item is expected to measure -- behavioral, cognitive or affective reactions -- before writing scenario questions and response options. If the developers decide to measure behavioral reactions, they need to place stress on behaviors and phrase scenario questions like "*what are you most likely to do in the situation*". If they want to measure cognitive reaction, the scenario question should be phrased like "*what are you most likely to think about the situation*". And if they intend to assess affective reaction, the scenario question would be "*what are you most feel like in the situation*". The response options should be written as "*I would do*", "*I would think*", "*I would feel*", correspondingly. The wording of scenario questions and response options can direct the test-takers to recall their response in the lines of the expected measure dimensions, which is confirmed in this dissertation. Some model-based SJT items used the same scenario while adopting different scenario questions and response options, and these items succeeded in measuring different constructs according to study results. An example is illustrated in Appendix H. Wording adjustment is minor in the scenario questions; however, such a minor adjustment combined with different response options would effectively prime test-takers to provide their responses in the targeted behavioral, cognitive or affective dimension. Therefore, caution is specifically needed when the developers decide to use the model-based approach to create the scenarios and response options on their own. Any word change can potentially make an SJT item deviate from its originally target construct.

Rockstuhl et al. (2015) distinguished response judgement and situation judgement, and proposed the additional value of only using scenarios in the assessment. They suggested that test-takers had made a judgement on the situation described in the scenario before arriving at the response options, and such a beforehand judgment provided of incremental information for the assessment results. However, my studies show that the same scenario can tap different constructs when paired with different questions and response options, which implies the unreliability of situation judgement because test-takers' beforehand judgement can be easily reshaped by the different scenario-question-option combinations. However, my finding on the effect of scenario-question-option combination is a by-product of the dissertation research, and future research could make a systematic investigation on the combination effect on the situation judgement. What's more, my research manifests two merits of using the combination design: fewer scenarios for use lowers the development cost and reduces the cognitive load of SJTs. Future research is suggested to empirically examine the two merits and explore the psychometric strengths and weaknesses of SJTs using the combination strategy.

9.1.3 About the reliability estimation

This dissertation research utilized Cronbach's alpha values to estimate SJTs reliability. With the careful structure and design during SJT scenario and response option development, both SJTs demonstrate acceptable α-indexed reliability; however, Cronbach's alpha is not the best method to estimate the SJTs because Cronbach's alpha is only suitable to assess the reliability of unidimensional measures (Cronbach, 1949) while neither of the SJTs are unidimensional. The SME-driven SJT measures an individual's overall performance in cross cultural interactions, so it is subject to heterogeneity like other SJTs developed in the same way. The model-based SJT is designed as a measure of five dimensions. Therefore, along with other researchers (Lievens et al., 2008; McDonald & Whetzel, 2007; Whetzel & McDonald, 2009), I agree that test-retest should be the more appropriate estimation method for the two SJTs. In the future research, I will examine the test-retest reliability of the two SJTs.

9.1.4 About the common method bias issue

Common method bias is a typical issue for social, psychological and psychometric studies. Measurement method, other than the construct of interest, contributes to the variance, which lead to an inflated or deflated relationship of interests (Podsakoff et al., 2003). The dissertation research design and the characteristics of SJTs fundamentally controlled the common method bias. Common raters and item characteristic effects are two main sources of common method bias. The validation study of this dissertation utilized the different sources for the predictor (3C) and one of the criteria (team performance), and controlled a potential common method bias stemmed from common rater design. Before conducting the surveys, I personally presented in each class to introduce the research purposes and the objectives of the surveys. I stressed on the usage of the data for evaluating the quality of SJTs rather than the test-taker's performance. I emphasized that there was no right or wrong answers, the results of the surveys would be kept highly confidential, and used only for the validation purpose. This intervention could reduce the evaluation apprehension of test-takers and hence control the common method bias (Podsakoff et al., 2003).

The distinctions of the form and characteristics between SJTs and the self-report Likert scales also help to reduce common method bias. Self-report Likert scales were a set of one-sentence statement, and the test-takers were instructed to rate the extent to which each statement described themselves in a five- or seven-point Likert scale. In contrast, the SJTs presented test-takers a situation described with a short paragraph, and each scenario was followed with responding actions. The test-takers were instructed to indicate their most and least likely reactions to the situations. SJTs direct test-takers to reflect their actual or possible behavioral tendency in specific contexts while the Likert scales tap test-takers' self-perception regardless of contexts. The distinctive forms and characteristics reduced the common method bias. Since I kept aware of common method bias from the very beginning of this dissertation study design, several design remedies were adopted to avoid the common method issues. No statistical remedy was utilized for controlling common method bias in the dissertation considering the facts that the superiority of design remedies over the statistical ones, the problematic inflated relationship assumption of statistical remedies and the impossibility of completely eliminating common method bias (Podsakoff et al., 2003; Brannick et al., 2010).

9.2 Limitations

No psychometric research is flawless, and this dissertation is not an exception in spite of its careful design and solid literature foundation. One limitation is its relatively small sample size (*N* = 65) for the criterion validity study using peer-reviewed performance as a criterion. Although the sample size met the minimal requirement for predictive validation, I have to admit the power might be impaired, post hoc power analysis indicate the power is .68 for the SME correlation. A larger sample size is favored for robust validation results. Another limitation of the sample is uneven numbers of peer scores each participant obtained. On average each participant got 2.6 peer review scores, but it ranged from one peer review score per participant to five peer review per participants. No literature or studies systematically discussed about the impact of the number of peer review scores on criterion validation results or the potential issues caused by the various number of peer review scores per participant; however, the data cases with only one peer review score may be more vulnerable to bias than those with several peer review scores.

The third limitation is about the utility calculation for the two SJT development approaches. There lacks a formula specific to calculating the utility of an assessment development approach. Therefore, I referred the concepts of the utility estimate commonly used in business world, the ratio of cost and quality. My comparison of utility between the two SJT approach stopped at presenting the information of each utility criterion, and any attempt to quantify the utility is beyond my capability in this dissertation research. Hopefully, a joint effort from academia and practitioners could yield a proper formula to calculate the utility of assessment development approaches in the near future.

References

Abbe, A., Gulick, L. M., & Herman, J. L. (2007). *Cross-cultural competence in Army leaders: A conceptual and empirical foundation*. Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences.

Abbe, A., & Halpin, S. M. (2010). *The cultural imperative for professional military education and leader development*. Army War College, Carlisle Barracks, PA.

Adler, N. J., & Bartholomew, S. (1992). Managing globally competent people. *The Executive, 6*(3), 52-65.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social psychology*, *68*(5), 804.

Akgün, A. E., Lynn, G. S., Keskin, H., & Dogan, D. (2014). Team learning in IT implementation projects: Antecedents and consequences. *International Journal of Information Management*, *34*(1), 37-47.

Ang, S., Dyne, L. Van, Koh, C., Ng, K.-Y., Templer, J. K., Tay, C., & Chandrasekar, N. A. (2007). Cultural Intelligence: Its Measurement and Effects on Cultural Judgment and Decision Making, Cultural Adaptation, and Task Performance. *Management and Organization Review*, *3*(3), 335–371.

Arasaratnam, L. A. (2009). The development of a new instrument of intercultural communication competence. *Journal of Intercultural Communication, 20*, 1-11.

Arasaratnam, L. A., & Doerfel, M. L. (2005). Intercultural communication competence: Identifying key components from multicultural perspectives. *International Journal of Intercultural Relations*, *29*(2), 137–163.

Arthur, W., & Bennett, W. (1995). the International Assignee: the Relative Importance of Factors Perceived To Contribute To Success. *Personnel Psychology*, *48*(1), 99–114.

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

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

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents*, *5*(1), 307-337.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development*, *67*(3), 1206-1222.

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, *41*(3), 586.

Basow, S. A., & Gaugler, T. (2017). Predicting adjustment of US college students studying abroad: Beyond the multicultural personality. *International Journal of Intercultural Relations*, *56*, 39-51.

Bass, B. M. (1985). *Leadership and performance beyond expectations*. City, State: Collier Macmillan

Bass, B. M., & Avolio, B. J. (Eds.). (1994). *Improving organizational effectiveness through transformational leadership*. City, State: Sage

Bauer, T. N., Truxillo, D. M., Tucker, J. S., Weathers, V., Bertolino, M., Erdogan, B., & Campion, M. A. (2006). Selection in the information age: The impact of privacy concerns and computer experience on applicant reactions. *Journal of Management*, *32*(5), 601-621.

Bell, M. P., & Harrison, D. A. (1996). Using intra-national diversity for international assignments: A model of bicultural competence and expatriate adjustment. *Human Resource Management Review*, *6*(1), 47-74.

Benet-Martinez, V. (2006). Biculturalism and Cognitive Complexity: Expertise in Cultural Representations. *Journal of Cross-Cultural Psychology*, *37*(4), 386–407.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*(3), 223-235.

Bernard, L. C., Mills, M., Swenson, L., & Walsh, R. P. (2005). An evolutionary theory of human motivation. *Genetic, Social, and General Psychology Monographs*, *131*(2), 129-184.

Bhaskar-Shrinivas, P., Harrison, D. A., Shaffer, M. A., & Luk, D. M. (2005). Input-based and time-based models of international adjustment: Meta-analytic evidence and theoretical extensions. *Academy of Management Journal*, *48*(2), 257-281.

Bird, A., Mendenhall, M., Stevens, M. J., & Oddou, G. (2010). Defining the content domain of intercultural competence for global leaders. *Journal of Managerial Psychology, 25*(8), 810-828.

Black, J. S., Mendenhall, M., & Oddou, G. (1991). Toward a comprehensive model of international adjustment: An integration of multiple theoretical perspectives. *Academy of Management Review*, *16*(2), 291-317.

Black, J. S. & Gregersen, H., 1999. The right way to manage expats. *Harvard Business Review*, https://hbr.org/1999/03/the-right-way-to-manage-expats

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, *62*(2), 229-258.

Brandl, J., & Neyer, A. K. (2009). Applying cognitive adjustment theory to cross-cultural training for global virtual teams. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, *48*(3), 341-353.

Brislin, R. W., Cushner, K., Cherrie, C., & Yong, M. (1986). *Intercultural interactions*. Beverly Hills.

Bruce, M. M., & Learner, D. B. (1958). A supervisory practices test. *Personnel Psychology*, *11*(1), 207-216.

Buckley P.J., Clegg J., Tan H. (2010). Cultural Awareness in Knowledge Transfer to China — The Role of Guanxi and Mianzi. In: *Foreign Direct Investment, China and the World Economy*. Palgrave Macmillan, London

Byram, M. (1995). Acquiring intercultural competence. A review of learning theories. In L. Sercu (Ed.), *Intercultural competence, I: The secondary school* (pp. 53-69). Aalborg, Denmark: Aalborg University Press.

Byram, M. (1997). *Teaching and assessing intercultural communicative competence*. UK: Multilingual Matters Ltd.

Caligiuri, P. M. (2000). The big five personality characteristics as predictors of expatriate's desire to terminate the assignment and supervisor-rated performance. *Personnel Psychology*, *53*(1), 67-88.

Poelmans, S. A., & Caligiuri, P. (2008). *Harmonizing work, family, and personal life: From policy to practice*. Cambridge University Press.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, *27*(4), 283-310.

Cardall, A. J. (1942). *Test of practical judgment*. Science Research Associates.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *Journal of applied psychology*, *82*(1), 143.

Chan, D., & Schmitt, N. (2017). Situational judgment tests. In Evers, Anderson, Voskuijl (Ed.), *The Blackwell Handbook of Personnel Selection*, 219-242.Wiley.

Chao, M. M., Takeuchi, R., & Farh, J. L. (2017). Enhancing cultural intelligence: The roles of implicit culture beliefs and adjustment. *Personnel Psychology*, *70*(1), 257-292.

Chen, G. M., & Starosta, W. J. (1997). Chinese conflict management and resolution: Overview and implications.

Chen, L., Fan, J., Zheng, L., & Hack, E. (2016). Clearly Defined Constructs and Specific Situations Are the Currency of SJTs. *Industrial and Organizational Psychology*, *9*(1), 34-38.

Chi, S. C. S., & Liang, S. G. (2013). When do subordinates' emotion-regulation strategies matter? Abusive supervision, subordinates' emotional exhaustion, and work withdrawal. *The Leadership Quarterly*, *24*(1), 125-137.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83-117.

Christopher, J. C., Wendt, D. C., Marecek, J., & Goodman, D. M. (2014). Critical Cultural Awareness. *American Psychologist*, *69*(7), 645–655.

Cleveland, H., Mangone, G. J., & Adams, J. C. (1960). The Overseas Americans. *The International Executive (pre-1986)*, *2*(3), 5.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, *13*(6), 653-665.

Cronbach, L. J. (1949). *Essentials of psychological testing*. Harper.

Deardorff, D. K. (2006). Identification and Assessment of Intercultural Competence as a Student Outcome of Internationalization. *Journal of Studies in International Education*, *10*(3), 241–266.

Deardorff, D. K. (2016). How to assess intercultural competence. *Research methods in intercultural communication: A practical guide*, 120-135.

De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P., & Themmen, A. P. N. (2017). Scoring method of a Situational Judgment Test: influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Sciences Education*, *22*(2), 243-265.

de Meijer, L. A., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity. *European Psychologist*, 15(3):229–236.

Diedenhofen, B., & Musch, J. (2016). Cocron: A Web Interface and R Package for the Statistical Comparison of Cronbach's Alpha Coefficients. *International Journal of Internet Science*, 11(1), 51-60.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, *49*(1), 71-75.

Doutrich, D., & Storey, M. (2004). Education and practice: Dynamic partners for improving cultural competence in public health. *Family & Community Health*, *27*(4), 298-307.

Duan, W., & Bu, H. (2017). Development and initial validation of a short three-dimensional inventory of character strengths. *Quality of Life Research*, *26*(9), 2519-2531.

Earley, P. C., & Ang, S. (2003). *Cultural intelligence: Individual interactions across cultures*. Stanford University Press.

Evelina, A., M., Schleicher, D. J., & Born, M. P. (2008). Cross-cultural social intelligence: An assessment for employees working in cross-national contexts. *Cross Cultural Management: An International Journal*, *15*(2), 109-130.

Fan, C., & Mak, A. S. (1998). Measuring social self-efficacy in a culturally diverse student population. *Social Behavior and Personality: an international journal*, *26*(2), 131-144.

Fiedler, F. E., Mitchell, T., & Triandis, H. C. (1971). The culture assimilator: An approach to cross-cultural training. *Journal of applied psychology*, *55*(2), 95.

File, Q. W. (1945). The measurement of supervisory quality in industry. *Journal of Applied Psychology*, *29*(5), 323.

Fritzsche, B. A., Stagl, K. C., Salas, E., & Burke, C. S. (2006). Enhancing the Design, Delivery, and Evaluation of Scenario-Based Training: Can Situational Judgment Tests Contribute? In J. A. Weekley & R. E. Ployhart (Eds.), *SIOP organizational series. Situational judgment tests: Theory, measurement, and application* (p. 301–318). Lawrence Erlbaum Associates Publishers.

Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. *The Qualitative Report, 20*(9), 1408-1416.

Gabrenya Jr, W. K., Moukarzel, R. G., Pomerance, M. H., Griffith, H., & Deaton, J. (2012). *A validation study of the Defense Language Office Framework for Cross-cultural Competence*. Technical report, Defense Equal Opportunity Management Institute.

Gabrenya, W. K. Jr., & Chen, X-W (2019). Examining measures of cross-cultural competence: What do they really tell us? In *the International Academy of Intercultural Research, Shanghai, China.*

Garman, A. N., & Johnson, M. P. (2006). Leadership competencies: An introduction. *Journal of Healthcare Management*, *51*(1), 13.

Gertsen, M. C. (1990). Intercultural competence and expatriates. *The International Journal of Human Resource Management*, *1*(3), 341–362.

Grant, K.L. (2009). The Validation of a Situational Judgment Test to Measure Leadership Behavior. *Masters Theses & Specialist Projects. Paper 64*.

Griffith, R. L., Frei, R. L., Snell, A. F., Hamill, L. S., & Wheeler, J. K. (1997). Warnings versus no-warnings: Differential effect of method bias. In *12th annual conference of the Society for Industrial and Organizational Psychology, St. Louis, MO*.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, *36*(3), 341-355.

Golubovich, J., Seybert, J., Martin-Raugh, M., Naemi, B., Vega, R. P., & Roberts, R. D. (2017). Assessing perceptions of interpersonal behavior with a video-based situational judgment test. *International Journal of Testing*, *17*(3), 191-209.

Gong, Y., & Fan, J. (2006). Longitudinal examination of the role of goal orientation in cross-cultural adjustment. *The Journal of Applied Psychology*, *91*(1), 176–84

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59-82.

Hammer, M. R., Bennett, M. J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The Intercultural Development Inventory. *International Journal of Intercultural Relations*, *27*, 421-443.

Hammer, M. R., Gudykunst, W. B., & Wiseman, R. L. (1978). Dimensions of intercultural effectiveness: An exploratory study. *International Journal of Intercultural Relations*, *2*(4), 382-393.

Hanson, M. A., Horgen, K. E., & Borman, W. C. (1998). *Situational judgment: An alternative approach to selection test development* (No. AB-34D-SYMPOSIUM). Navy Advancement Center Persacola, FL.

Hauenstein, N. M., Findlay, R. A., & McDonald, D. P. (2010). Using situational judgment tests to assess training effectiveness: Lessons learned evaluating military equal opportunity advisor trainees. *Military Psychology*, *22*(3), 262-281.

Hedge, J. W., Borman, W. c., & Hanson, M. A. (1996). Videotaped crew resource management scenarios for selection and training applications. Paper presented at the 38th annual conference of the International Military Testing Association.

Hogan, R. & Hogan, J. (1992). *Hogan Personality Inventory Manual*. Tulsa, OK: Hogan Assessment Systems.

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational threats to the use of SJTs: Faking, coaching, and retesting issues. In Weekly, J.A., & Polyhart, R. E. (Eds.), *Situational judgment tests: Theory, measurement, and application*, 205-232.

Howard-Hamilton, M., Richardson, B. and Shuford, B. 1998. Promoting multicultural education: A holistic approach. *College Student Affairs Journal*, *18*(1), 5–17.

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology*, *12*(3), 317.

Hunter, B. (2006). What Does It Mean to Be Globally Competent? *Journal of Studies in International Education*, *10*(3), 267–285.

Imahori, T.T. and Lanigan, M.L. (1989). Relational model of intercultural communication competence. *International Journal of Intercultural Relations 13(3)*, 269–286.

Jensen, A. R. (1998). The suppressed relationship between IQ and the reaction time slope parameter of the Hick function. *Intelligence*, *26*(1), 43-52.

Johnson, J. P., Lenartowicz, T., & Apud, S. (2006). Cross-cultural competence in international business: toward a definition and a model. *Journal of International Business Studies*, *37*(4), 525–543.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the Subjects' point of view. *European Journal of Psychological Assessment*, *22*(3), 168-176.

Kaplan, S., Bradley, J. C., Luchman, J. N., & Haynes, D. (2009). On the role of positive and negative affectivity in job performance: A meta-analytic investigation. *Journal of Applied Psychology, 94(1),* 162–176.

Kawashima, A. (2008). *Study on cultural competency of Japanese nurses* (Doctoral dissertation).

Kealey, D. J. (1996). The challenge of international personnel selection. In D. Landis & R. S. Baghat (Eds.), *Handbook of intercultural training* (pp. 81–105). Thousand Oaks, CA: Sage.

Kim, Y. Y. (1988). *Communication and cross-cultural adaptation: An integrative theory.* Philadelphia: Multilingual Matters.

Kirkpatrick, D. L., & Planty, E. (1960). Supervisory inventory on human relations. *Chicago: Science Research Associates*.

Koo Moon, H., Kwon Choi, B., & Shik Jung, J. (2012). Previous international experience, cross-cultural training, and expatriates' cross-cultural adjustment: Effects of cultural intelligence and goal orientation. *Human Resource Development Quarterly*, *23*(3), 285-330.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests?. *Journal of Applied Psychology*, *100*(2), 399.

Kupka, B. (2008). *Creation of an instrument to assess intercultural communication competence for strategic international human resource management* (Doctoral dissertation, University of Otago).

Langer, E. J. (1997). *The power of mindful learning*. Reading, MA: Addison-Wesley.

Langer, E. J. (2000). Mindful learning. *Current directions in psychological science*, *9*(6), 220-223.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research*, *39*(2), 329-358.

Legree, P. J., & Psotka, J. (2006). *Refining situational judgment test methods*. Army Research Institute for Behavioral and Social Sciences Arlington, VA.

Leiba-O'Sullivan, S. (1999). The distinction between stable and dynamic cross-cultural competencies: implications for expatriate training. *Journal of International Business Studies 30*(4): 709–725.

Leung, K., Ang, S., & Tan, M. L. (2014). Intercultural competence. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 489-519.

Libbrecht, N., & Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. *International Journal of Psychology*, *47*(6), 438-447.

Lievens, F. (2006). International situational judgment tests.

Lievens, F. (2013). Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. *Medical education*, *47*(2), 182-189.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*(4), 981-1007.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90*(3), 442.

Lievens, F., Corstjens, J., Sorrel, M. A., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The Cross-cultural Transportability of Situational Judgment Tests: How does a US-based integrity situational judgment test fare in S pain?. *International Journal of Selection and Assessment*, *23*(4), 361-372.

Lievens, F., De Soete, B. (2015). Situational Judgment Test. In: James D. Wright (editor-in-chief), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13–19), 2nd edition, Vol 22. Oxford: Elsevier.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, *9*(1), 3-22.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426-441.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of applied psychology*, *91*(5), 1181.

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, *94*(4), 1095.

Li-Yueh, L., & Alfiyatul, Q. (2015). The Effects of Work-Role Demands on Cross-cultural Adjustment and Expatriate Effectiveness: A Meta-analysis. *Anthropologist*, *22*(3), 636-649.

Long, J. H., Yan, W. H., Yang, H. D., & Van Oudenhoven, J. P. (2009). Cross-cultural adaptation of Chinese students in the Netherlands. *US-China Education Review*, *6*(9), 1-9.

MacCann, C., Fogarty, G. J., Zeidner, M., & Roberts, R. D. (2011). Coping mediates the relationship between emotional intelligence (EI) and academic achievement. *Contemporary Educational Psychology*, *36*(1), 60-70.

Matsumoto, D., & Hwang, H. C. (2013). Assessing Cross-Cultural Competence: A Review of Available Tests. *Journal of Cross-Cultural Psychology*, *44*(6), 849–873.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & GRUBB III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, *60*(1), 63-91.

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

McDaniel M. A. & Nguyen NT. (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*(5), 515-525.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. *Applied measurement: Industrial psychology in human resources management*, 235-257.

Mendenhall, M., & Oddou, G. (1985). The dimensions of expatriate acculturation: A review. *Academy of management review*, *10*(1), 39-47.

Messelink, A., & tenThije, J. D. (2012). Unity in Super-diversity: European capacity and intercultural inquisitiveness of the Erasmus generation 2.0. *Dutch Journal of Applied Linguistics*, *1*(1), 80–101.

Meydanlioglu, A., Arikan, F., & Gozum, S. (2015). Cultural sensitivity levels of university students receiving education in health disciplines. *Advances in Health Sciences Education*, *20*(5), 1195-1204.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology*, *60*(4), 1029-1049.

Moss, F. A. (1926). Do you know how to get along with people?. *Scientific American*, *135*, 26-27.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, *95*(2), 321.

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

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance*, *29*(4), 331-346.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests.

Motowidlo, S. J., & Peterson, N. G. (2008). Effects of organizational perspective on implicit trait policies about correctional officers' job performance. *Human Performance*, *21*(4), 396-413.

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*(4), 337-344.

Mowry, H. W. (1964). *Leadership evaluation and development scale: LEADS. Casebook (1964)*. Psycholog. Services.

Mumford, M.D. (1999) Construct validity and background data: Issues, abuses, and future directions. Human Resource Management Review, 9, 117–145

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The Team Role Test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology*, *93*(2), 250.

Murphy, K. R. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance?. *Human Performance*, *18*(4), 343-357.

Myers, D. G. (1998). *Psychology* (5th ed.). New York: Worth.

Northrop, L. C. (1989). *The psychometric history of selected ability constructs*. United State Office of Personnel Management.

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

Nguyen, N. T. & McDaniel, M. A. (2003). Response instructions and racial differences in a situational judgment test. *Applied HRM Research*, *8*(1), 33-44.

Ones, D. S., & Viswesvaran, C. (1999). Relative importance of personality dimensions for expatriate selection: A policy capturing study. *Human performance*, *12*(3-4), 275-294.

Ones, D. S., Viswesvaran, C., & Korbin, W. P. (1995). Meta-analyses of fakability estimates: Between-subjects versus within-subjects designs. In *10th annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL*.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2012). Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior?. *Human Performance*, *25*(4), 335-353.

Oostrom, J. K., De Soete, B., & Lievens, F. (2015). Situational judgment testing: A review and some new developments. In *Employee Recruitment, Selection, and Assessment* (pp. 184-201). Psychology Press.

Ormel, J., Oldehinkel, A. J., & Brilman, E. I. (2001). The interplay and etiological continuity of neuroticism, difficulties, and life events in the etiology of major and subsyndromal, first and recurrent depressive episodes in later life. *American Journal of Psychiatry*, *158*(6), 885-891.

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

Paige, R. M., Jacobs-Cassuto, M., Yershova, Y. A., & DeJaeghere, J. (2003). Assessing intercultural sensitivity: An empirical analysis of the Hammer and Bennett Intercultural Development Inventory. *International journal of intercultural relations*, *27*(4), 467-486.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical education*, *46*(9), 850-868.

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Medical education*, *43*(1), 50-57.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*(1), 70-89.

Peltokorpi, V., & Froese, F. (2014). Expatriate personality and cultural fit: The moderating role of host country context on job satisfaction. *International Business Review*, *23*(1), 293-302.

Peterson, M. H., Griffith, R. L., & Converse, P. D. (2009). Examining the role of applicant faking in hiring decisions: Percentage of fakers hired and hiring discrepancies in single-and multiple-predictor selection. *Journal of Business and Psychology*, *24*(4), 373.

Peus, C., Braun, S., & Frey, D. (2013). Situation-based measurement of the full range of leadership model—Development and validation of a situational judgment test. *The Leadership Quarterly*, *24*(5), 777-795.

Pinder, C. C. (2014). *Work motivation in organizational behavior*. Psychology Press.

Plint, S., & Patterson, F. (2010). Identifying critical success factors for designing selection processes into postgraduate specialty training: the case of UK general practice. *Postgraduate medical journal*, *86*(1016), 323-327.

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 tests. *International Journal of Selection and Assessment*, *11*(1), 1-16.

Ployhart, R. E., & MacKenzie Jr, W. I. (2011). Situational judgment tests: A critical review and agenda for the future.

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. *Situational judgment tests: Theory, measurement, and application*, 345-350.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of clinical psychology*, *38*(1), 119-125.

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology*, *100*(2), 464.

Rockstuhl, T., & Ng, K. Y. (2015). The effects of cultural intelligence on interpersonal trust in multicultural teams. In *Handbook of cultural intelligence* (pp. 224-238). Routledge.

Ruben, B. D. (1976). Assessing communication competency for intercultural adaptation. *Group & Organization Studies*, *1*(3), 334-354.

Ryan, A. M., & Huth, M. (2008). Not much more than platitudes? A critical look at the utility of applicant reactions research. *Human Resource Management Review*, *18*(3), 119-132.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of management*, *26*(3), 565-606.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs. *Handbook of industrial, word and organizational psychology*, 165-199.

Shaffer, M. A., Harrison, D. A., Gregersen, H., Black, J. S., & Ferzandi, L. A. (2006). You can take it with you: Individual differences and expatriate effectiveness. *Journal of Applied psychology*, *91*(1), 109.

Sharma, S., Gangopadhyay, M., Austin, E., & Mandal, M. K. (2013). Development and validation of a situational judgment test of emotional intelligence. *International Journal of Selection and Assessment*, *21*(1), 57-73.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. *Situational judgment tests: Theory, measurement, and application*, 135-155.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of applied psychology*, *47*(2), 149.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures (4th ed.)*. Bowling Green, OH: Author.

SPENCER, L. Y. S., & Spencer, S. (2011). SM (1993). Competence at work. Models for superior performance.

Spitzberg, B. H., & Changnon, G. (2009). Conceptualizing intercultural competence. *The SAGE handbook of intercultural competence*, 2-52.

Stagl, K. (2006). The construct validity of a situational judgment test in a maximum performance context.

Sternberg, R. J. (Ed.). (1990). *Wisdom: Its nature, origins, and development*. Cambridge University Press

Sternberg, R. J. (2000). Images of mindfulness. *Journal of Social Issues*, *56*(1), 11–26. https://doi.org/10.1111/0022-4537.00149

Sternberg, R. J. (2015). Successful intelligence as a framework for understanding cultural adaptation. *Handbook on Cultural Intelligence; Ang, S., van Dyne, L., Eds*, 306-317.

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), 207-228.

Storti, C. (2017). *Cross-cultural dialogues: 74 brief encounters with cultural difference*. Nicholas Brealey.

Strekalova, E. (2013). *Intercultural sensitivity of teachers working with refugeechildren*. State University of New York at Buffalo.

Tamam, E. (2010). Examining Chen and Starosta's model of intercultural sensitivity in a multiracial collectivistic country. *Journal of Intercultural Communication Research*, *39*(3), 173-183.

Tarique, I., & Caligiuri, P. (2009). The role of cross-cultural absorptive capacity in the effectiveness of in-country cross-cultural training. *International Journal of Training and Development*, *13*(3), 148-164.

Tarique, I. & Weisbord, E. (2013). Antecedents of dynamic cross-cultural competence in adult third culture kids (ATCKs). *Journal of Global Mobility: The Home of Expatriate Management Research*, *1*(2), 139–160.

Thibodeaux, H. F., & Kudisch, J. D. (2003). The relationship between applicant reactions, the likelihood of complaints, and organization attractiveness. *Journal of Business and Psychology*, *18*(2), 247-257.

Thomas, D. C., Elron, E., Stahl, G., Ekelund, B. Z., Ravlin, E. C., Cerdin, J. L., ... & Maznevski, M. (2008). Cultural intelligence: Domain and assessment. *International Journal of Cross Cultural Management*, *8*(2), 123-143.

Ting-Toomey, S. (1999). *Communicating across cultures*. New York: Guilford.

Ting-Toomey, S., & Kurogi, A. (1998). Facework competence in intercultural conflict: An updated face-negotiation theory. *International Journal of Intercultural Relations, 22*, 187-225.

Trejo, B. C., Richard, E. M., van Driel, M., & McDonald, D. P. (2015). Cross-cultural competence: The role of emotion regulation ability and optimism. *Military Psychology*, *27*(5), 276-286.

Tung, R. L. (1981). Selection and training of personnel for overseas assignments. *Columbia Journal of World Business, Spring,* 69–78.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, *61*(4), 871-925

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. *Canadian Journal of School Psychology*, *24*(2), 108-124.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. *Situational judgment tests: Theory, measurement, and application*, *26*, 157-182.

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

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*(3), 188-202.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*(3), 291-309

Whetzel, D. L., & Reeder, M. C. (2016). Why some situational judgment tests fail to predict job performance (and others succeed). *Industrial and Organizational Psychology*, *9*(1), 71-77.

Wildman, J. L., Griffith, R. L., & Armon, B. K. (Eds.). (2016). *Critical Issues in Cross Cultural Management*. Springer.

Wilson, J. K. (2013). Exploring the past, present, and future of cultural competency research: The revision and expansion of the sociocultural adaptation construct.

Wolf, M. M. (1978). Social validity: the case for subjective measurement or how applied behavior analysis is finding its heart 1. *Journal of applied behavior analysis*, *11*(2), 203-214.

Appendix A
Interview Questions

We know you worked, or you are working in a group project, with some persons from other countries or with different cultural background.

Describe the multicultural team you are recently involved briefly

- Please tell me about what the teamwork is.
- From what country your teammates come?
    - How about his/her personality?
    - Could he/she speak good English? Can he/she be easily understood?
- How is your experience when working with them?
    - Describe your unpleasant or unsuccessful experience in multicultural teamwork?
        - Describe the group
        - How are the teamwork outcomes? Project results?
        - What caused the failure except techniques?
        - If it comes again, what do you think is an effective way for more positive/favorable outcomes
        - Any other reasons lead to such a failure? (repeat till exhausted)
        - Any other unpleasant or unsuccessful experience in multicultural teamwork?

        (repeat it till exhausted)
    - Describe your favorable and successful experience in multicultural teamwork?
        - Describe the team
        - How are the teamwork outcomes? Project results?
        - Why do you like it?
        - What lead to success of the teamwork? (repeat it till exhausted)
        - Any major differences between the successful teamwork and the failed teamwork?
        - Any other successful multicultural teamwork experience

        (repeat till exhausted)
- Except techniques, what challenges you meet when working with those persons?
- As you observe, how often does the challenge happen in the multicultural teamwork?
- Do you think how important it is to solve the challenge (timely)?
    - What if the challenge is left alone?
- How difficult to solve the challenge? Why?

  (repeat till exhausted)
- People are very likely to react differently to the challengeable situations in the group project.
    - Could you recall some situations you work out very well?
        - What is the situation?

- What did you do?
- What were the outcomes?

- Could you recall more examples of excellent performance, either yours or other persons’, in dealing with the situations?
- Could you recall some situations you failed to work out or didn’t work well?
    - What is the situation?
    - What did you do?
    - What were the outcomes?
- Could you recall some poor performance, either yours or other persons’, in dealing with those situations?
    - If you had an opportunity to deal with it again, what would you do to effectively solve it?

- Could you recall a challengeable situation where high capable people respond distinctly from low capable people?
    - What do the highly capable people tend to respond?
    - What do the low capable people tend to respond?

Appendix B
Item Sample of the SME-driven 3C SJT

Imagine English is the official language of your university. When you talk to an international student in English, you find they can't understand what you say. What do you most/least likely to do in this situation?

a. I would try to communicate with them in a different way by using translation tools, simple words, writing, or drawing.
b. I would slow down and repeat what I said.
c. I would direct them to the academic center for more help.
d. I would find another person to help with our communication.
e. I would complete the conversation and leave.
f. I would try to speak their native language.

Imagine you study at a foreign university for your master's degree. One day you hear a Ph.D. student complain you are rude because you didn't greet him first when you met. You also learn that the country is hierarchical and status sensitive. What do you most/least likely to do in this situation?

a. I would greet him and explain that you don't mean to offend him.
b. I would greet him as if nothing happened.
c. I wouldn't greet him and mind my own business.
d. I would tell him it is not proper to talk behind my back.
e. I would speak to him about cultural differences.
f. I would avoid him by keeping your distance from him.

Appendix C
Item Sample of the Model-Based 3C SJT

Image you are an international student, and you and your local friends decide to watch a movie in the theater on the weekend. You visit the theater website and find a new movie with the strange name *Bohemian Rhapsody*. What do you most/least likely to do in this situation?

a. I would ignore it and go on reading through the movie list.
b. I would watch the trailers and reviews available online.
c. I would search for the movie information and what Bohemian Rhapsody means.
d. I would choose this movie to watch.
e. I would ask my friends to make a decision.
f. I wouldn't watch it until my friends recommend it.
g. I would ask my friends what the film is about.

You are invited to a weekend party hosted by one of your friends. When you arrive, you find you do not know anyone except the host. What do you most/least likely to do in this situation?

a. I would walk around and network with others.
b. I would leave the party shortly.
c. I would ask my friend to introduce me to other guests.
d. I would stay close with my friend.
e. I would spend most of my time by myself.
f. I would have a good time at the party with new people.

Appendix D
The Readability of SJTs and each item

SME-driven items	Readability	Model-based Items	Readability
S5	68.3	M30	73.8
S7	66.7	M31	80.9
S9	76.8	M32	69
S11	75	M33	91.6
S15	73.8	M34	92.3
S17	75.6	M35	67.4
S19	79.9	M36	80.8
S21	74.5	M37	83.3
S23	58	M38	74.2
S25	64.3	M39	79.4
S27	79.8	M40	63.1
S29	67.6	M41	77.4
S31	63.9	M42	79.1
S33	74.1	M43	87.8
S35	68.6	M44	69.9
S37	77.7	M45	79.9
S39	63.2		
S45	84.5		
S47	75.6		
S49	82		
Overall	**72.9**		**77.5**

Appendix E
The Time Log of the Two SJTs Development

	Model-based SJT	Time (*mins*)	SME-driven SJT	Time (*mins*)
1	INQ item writing	30	Pilot interview (1 person)	150
2	Materials review and search	180	Logistic preparation for informed consent, email, protocol, IRB and ads	240
3	Material review and scenarios writing I	210	Emails to potential interviewees	30
4	Material review and scenarios writing II	180	#2 interview + transcription, 12.21	300
5	Material review and scenarios writing III	180	#3 interview + transcription (1.3)	275
6	Material review for self-efficacy IV	150	Interview questions design and draft (12.5, 1.15)	90
7	Grammatical check scenario draft v1.0	280	Create time slots and sign-up link for interview (1.21)	30
8	Discussed about v1.0 with Rich and item-construct rating	60	Interview notification + reserve rooms	30
9	Mindfulness Write and edit v1.1	70	3 interviews (1.23)	240
10	Response collection for double check response options	180	3 interviews (2.8-2.10) and description	810
11	Content validity test and item adjustment	180	Recruitment, create time slots, sign-up link for interview, purchase (2.6)	60
12	Content validation 1	120	Interview (2) logistics	60
13	Content validation 2	120	17 interviews and transcription	4605
14	Item reduction 1 (based on item-total correlation)	60	Theme extraction and scenario writing	1000
15	Item reduction 2 (based on content validation and item-subtotal r)	200	Scenarios selection & SMEs communication	120
16			1st scenario selection processing (a total of 24 scenarios)	120

17	Response collection	180
18	Response pool building and response option establishing	950
19	Item reduction 1 (based on item-total correlation) – revised version 1	60
20	Item reduction 2 (trial)	55

Appendix F
The Momentary Cost of the Two SJTs Development

	Model-based SJT	**Cost** ($)	**SME-driven SJT**	**Cost** ($)
1	Scenario development	0	Interviews for scenario development	225
2	Reaction option development	0	Reaction option development	0
3	Content validation	0	Typicality Investigation	0
4	Finalization & Validation	60	Finalization & Validation	60
Total		60		285

Appendix G
The Numbers of SMEs Used for the Two SJTs Development

	Model-based SJT	# of SMEs	SME-driven SJT	# of SMEs
1	Scenario development	0	Scenario generation	23
2	Reaction option generation	0	Reaction option generation (I)	4
3	Reaction option finalization	66	Typicality Investigation	6
4	Content validation	10	Reaction option generalization (II)	66
Total		76		99

Appendix H
An Example of Constructs measured by the Same Scenario with Different Response Options

You are invited to a weekend party hosted by one of your friends. When you arrive, you find you do not know anyone except the host. What would you most/least likely do?

a. I would walk around and network with others.
b. I would leave the party shortly.
c. I would ask my friend to introduce me to other guests.
d. I would stay close with my friend.
e. I would spend most of my time by myself.
f. I would have a good time at the party with new people.

You are invited to a weekend party hosted by one of your friends. When you arrive, you find you do not know anyone except the host. What would you most/least likely feel?

a. I would be nervous.
b. I would be angry.
c. I would feel nothing special.
d. I would be excited.
e. I would feel awkward.
f. I would be upset.
g. I would stay calm.