12-2018

The Development of a Situational Judgment Test to Assess Collegiate Judgment: A Pilot Study

Jared Stevens
University of Nebraska - Lincoln, jstevens0010@gmail.com

THE DEVELOPMENT OF A SITUATIONAL JUDGMENT TEST TO ASSESS COLLEGIATE JUDGMENT: A PILOT STUDY

by

Jared T. Stevens

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor James A. Bovaird

Lincoln, Nebraska

December, 2018

THE DEVELOPMENT OF A SITUATIONAL JUDGMENT TEST TO ASSESS COLLEGIATE JUDGMENT: A PILOT STUDY

Jared T. Stevens, M.A.

University of Nebraska, 2018

Advisor: James A. Bovaird

Traditionally, colleges and universities have focused primarily on cognitive predictors (e.g., ACT/SAT scores, high school GPA), and have struggled to find an accurate and objective way of measuring non-cognitive skills, often resorting to personality measures or interviews, or deciding not to measure them at all. Recently, there has been a push for alternative forms of student selection that result in less adverse impact and do not ignore important skills and traits that are necessary to be successful in college (Peeters & Lievens, 2005; Atkinson, 2001).

Growing evidence suggests Situational Judgment Tests (SJTs) may be one way to achieve this goal. SJTs are a type of psychological aptitude test in which respondents are presented with a short vignette/scenario about a particular situation, and are then asked to either rate the effectiveness of the responses (knowledge SJTs), or indicate what response the participant would choose if they were in that situation (behavioral tendency SJTs).

The current study collected pilot data from undergraduate students at a large Midwestern university. Students answered several SJT items, a measure of the Big 5 (Goldberg's Big 5 Markers), and a small number of demographic items, including students' GPA and standardized test scores (ACT/SAT). Students' responses on the SJT items were compared with the other data collected to determine if there was validity evidence for the use of SJTs in predicting college GPA. The results provide evidence that SJTs may be

useful for admissions departments to aid in the selection of students or for student retention and development.

Acknowledgements

First, I would like to thank my graduate advisor, *Dr. James Bovaird*, for his help and support with this master's thesis, and for his patience. Next, I would like to thank my unofficial co-advisor, *Dr. Leslie Hawley*, for her guidance, support, and commitment to this project, even though she was no longer affiliated with the University of Nebraska. I would also like to thank my fellow graduate students for their help, feedback, and willingness to be a part of the 'Subject Matter Expert' pool, as well as all of the educators for letting me collect data in their classes. Finally, I would like to thank my family and friends for their constant support along this journey. Most importantly, I would like to thank my parents, *Jeff and Nancy*, who have always been supportive of me in anything I try to accomplish, and my girlfriend, *Mariah Poore.* Your love and support was what I needed most throughout this process.

Table of Contents

List of Figures

List of Tables

Introduction

With selection methods, businesses, organizations, and universities are concerned with selecting the best applicant for the prospective job or task (Salgado & Moscaso, 2008). Traditionally, the focus has been on cognitive predictors or general mental ability (GMA); for example, IQ and job knowledge tests for businesses/organizations, and ACT/SAT scores and high school GPA for colleges and universities. While evidence for the validity of cognitive ability measures for predicting job/school performance is stronger than for any other method, and it has been considered the primary selection method for hiring decisions (Schmidt & Hunter, 1998; Schmidt, Shaffer, & Oh, 2008; Hough & Oswald, 2000), concerns have been raised that selection methods may be too heavily focused on cognitive predictors (Peeters & Lievens, 2005; Atkinson, 2001; Hough, Oswald, & Ployhart, 2001), almost to a fault. Cognitively-oriented measures have been shown to have an adverse impact on minorities and women (see Hough, Oswald & Ployhart, 2001, Whetzel et al. 2008), and such a heavy focus on cognitive predictors may result in entities ignoring important skills and traits (e.g., personality, non-cognitive attributes, interpersonal skills) that are necessary to be successful in the job/task. Additionally, an accurate and objective way of measuring non-cognitive skills has yet to become widespread, as businesses and universities often resort to personality measures or interviews, or decide not to measure them at all. As a result, there has been a considerable push within the last several years for alternative selection methods that result in less adverse impact and do not ignore important skills that are necessary to be successful in college (Atkinson, 2001; Peeters & Lievens, 2005).

Growing evidence suggests that Situational Judgment Tests (SJTs) may be one way to accomplish this goal. SJTs are a type of psychological aptitude test in which respondents are presented with a short vignette/scenario about a particular situation, and are then asked to rate the effectiveness of the responses, or indicate what response the participant would choose if they were in that situation (called behavioral tendency SJTs). SJTs are now becoming widely used in personnel selection, and are increasingly used in student selection for medical (Luschin-Ebengreuth, Dimai, Ithaler, Neges, & Reibnegger, 2015), dental (Buyse & Lievens, 2011), and graduate school (Koczwara et al., 2012; Patterson et al. 2016) selection. Research in these areas have demonstrated that SJTs may provide incremental validity over and above measures of cognitive ability and personality. Because SJTs are often designed around non-cognitive, interpersonal, and intrapersonal factors, they often capture domains that are not traditionally the focus of graduate admissions departments.

However, to date, limited researchers have attempted to develop situational judgment tests for the purposes of predicting *undergraduate* 'collegiate success' (see Oswald et al., 2004). Currently, college admissions departments focus on students' high school transcripts, grade point averages (GPA) and test scores (ACT/SAT), letters of recommendation, and essays/personal statements (Oswald et al., 2004; Peeters & Lievens, 2005; Sacks, 2000). Based on prior research and the relative successfulness of SJTs in other domains, college admission departments should consider using SJTs as a supplement to these measures for several reasons.

First, SJTs can be generalized to basically any construct that admissions departments want to measure. Research has also shown that there may be some useful

level of incremental validity for SJT measures, over and above the traditional measures of cognitive ability and personality that colleges use (see Buyse & Lievens, 2011; Lievens & Coatsier, 2002; O'Connel, Hartman, McDaniel, Grubb III, & Lawrence, 2007; Patterson et al. 2016). Additionally, SJTs are a more subjective measure than letters of recommendation, portfolios, and interviews, which admissions departments frequently utilize. Evidence also suggests SJTs *may* show less adverse impact or subgroup differences for minority populations than traditional methods colleges use (e.g. standardized tests) (Chan & Schmitt,1997; Lievens & Coatsier, 2002; Motowidlo et al.,1990; Weekley & Jones 1999; Whetzel, McDaniel, & Nguyen, 2008). Finally, SJTs have the potential to be used for formative assessment/evaluation of the students. For example, a student may be struggling in their first semester and placed on academic probation. It may be helpful for the college/university to go back and examine their SJT responses, to determine if their decision-making and judgment could be improved. This would make SJTs a useful measure for student retention.

Purpose of Proposed Research

The primary objective of this study is to create a Situational Judgment Test (SJT) for the purposes of predicting undergraduate collegiate success, as measured by college GPA. Students' responses on the SJT items will be compared to the other data collected to determine if there is validity evidence of SJTs in predicting college GPA. Particular focus will be spent on the evidence of criterion-related validity in predicting college GPA, construct validity (evidence of convergent and discriminant validity), and incremental validity in predicting college GPA over and above the traditional measures that college admissions departments use (i.e. ACT/SAT scores and personality measures).

This study also has several secondary aims. Prior research has demonstrated that the different scoring methods utilized for SJTs influence their validity in predicting the criterion (Bergman et al., 2006; Legree et al., 2010). Therefore, one secondary objective is to determine if the validity estimates of SJTs with SME-based scoring in predicting college GPA are different than the validity estimates of SJTs with consensus-based scoring. Prior research has also shown that SJTs result in less subgroup differences than measures of cognitive ability, as a result of the heterogeneous factor structure and the fact that they often encompass both cognitive and non-cognitive abilities (Whetzel, McDaniel, & Nguyen, 2008). Therefore, another secondary objective is to determine whether the SJTs designed for the purposes of this study had lower subgroup differences than the students' ACT/SAT scores.

Proposed Study and Research Questions

This study collected pilot data from undergraduate students at a large Midwestern university, in which students answered 14 SJT test items (7 knowledge SJTs as well as 7 behavioral tendency SJTs), a measure of the Big 5 (Goldberg's Big 5 Markers – short form), and a small number of demographic items, including students' GPA and standardized test scores (ACT/SAT). The data collected will then be analyzed to answer the following research questions:

1. Do the different scoring methods used in this study (consensus-based vs. SME-based) result in different validity estimates in predicting college GPA?
2. Is there evidence of criterion-related validity for SJTs in predicting college GPA?

3. Does construct validity evidence, in the form of convergent and discriminant validity, provide evidence that the different response instructions (knowledge vs. behavioral tendency) result in the measurement of different constructs?
4. Is there evidence of incremental validity for the prediction of college GPA, over and above the traditional measures/predictors college admissions departments use (i.e. ACT/SAT scores and personality tests)?
5. Do the SJTs designed for the purposes of this study result in less subgroup differences than the students' ACT/SAT scores?

Practical Implications

If there is significant validity evidence for SJTs in predicting college SJTs, there are several practical implications. First, college admissions departments may choose to use them as a supplement to ACT/SAT scores and/or high school GPA, as a way to capture the non-cognitive attributes students' possess. Second, findings from the research question focusing on the different response instructions should help provide evidence on whether changing the response instructions of SJTs, while leaving the content of the vignette/scenario constant, results in the assessment of different constructs.

Beyond the validity evidence of the SJTs, there are several practical implications to using SJTs as a form of student selection or retention. First, some students may struggle on standardized tests, but have some non-cognitive skills or attributes that would lead them to be successful in college. Because colleges and universities focus so heavily on cognitive predictors, these students may not get admitted into a college/university. SJTs may be one way to overcome this problem. They would allow colleges and

universities to capture the non-cognitive skills and attributes these students possess, which may help them overcome their poor standardized test results.

The final practical implication of using SJTs is their potential to be used for formative assessment/evaluation. For example, a student may be struggling in their first semester of college and may even be placed on academic probation. It might be worthwhile for the student to meet with an academic advisor and examine their SJT responses, to analyze their decision-making ability and see if their judgment in college could be improved. This would make SJTs a useful measure for student retention.

Literature Review

Selection methods have received considerable attention among researchers in the 21st century. Selection methods are one focus of Industrial/Organization psychology – the scientific study of human behavior in organizations and the work place (Aamodt, 2012) - and are concerned with selecting the best individual for the prospective job or task. Selection methods are often categorized into two major types – employee selection for a job or business, and student selection for college or graduate school.

Employee Selection

The traditional model of employee selection has remained stable for several years (Robertson & Smith, 2001). Employee selection is concerned with selecting the best applicant for the prospective job (Salgado & Moscaso, 2008). Often, the first step in employee selection is to perform a detailed job analysis. A job analysis, as defined by Cascio (1991, p. 188) is the "process of defining a job in terms of the behaviors necessary to perform it. Job analysis is comprised of two major elements: job descriptions and job specifications." A job analysis includes the characteristics of the job – procedures,

methods, and standards of performance – and helps identify the behaviors, knowledge, abilities, skills, and personality characteristics that are necessary to successfully perform the job (Cascio, 1991).

The overarching goal of a job analysis is to identify the attributes (physical and psychological) required by someone who will perform the job effectively (Robertson & Smith, 2001). Often the individual performing the job analysis will utilize a wide variety of methods to capture the full range of characteristics of the job. Questionnaires and surveys, interviews with job incumbents, direct observations of workers, diaries of workers, and documents like instructional materials and maintenance records are then used to capture these attributes and characteristics (Smit-Voskuijl, 2005). For instance, several questionnaires and surveys have been developed for the sole purpose of job analyses. Examples of these include the Position Analysis Questionnaire (PAQ, McCormick, Jeanneret, & Mecham, 1969), the Functional Job Analysis (FJA), which was developed by the Employment and Training Administration of the United States Department of Labor, and the Ability Requirements Scales (ARS; Fleishman & Mumford, 1988). Selection methods are then designed to enable those responsible to evaluate each candidate's capabilities established by the job analysis. A variety of selection methods have been used by workplaces, but the most common methods include interviews, cognitive ability tests, and personality measures (Robertson & Smith, 2001).

Interviews are the most common selection method used by employers (Moscaso, 2000; Robertson & Smith, 2001). In an interview, potential employees are asked questions designed to assess different attributes the employee should possess (Robertson & Smith, 2001). In a meta-analysis conducted by McDaniel, Whetzel, Schmidt, &

Maurer (1994), researchers found an average validity estimate of .37 (n = 25,444) for interviews, using job performance ratings as the criterion. A criterion is a principle or standard by which something is judged (Hooker, 1959). Additionally, research has shown structured interviews have much higher levels of predictive validity than unstructured interviews (r = 0.56 for structured vs. r = 0.20 for unstructured; Huffcut & Arthur, 1994; Salgado, 1999). In this study, interviews with some sort of embedded structure, like using a situational interview or behavior description interviewing, were much more predictive of future job success (as measured by supervisory ratings of performance) than interviews that were unstructured.

Cognitive ability tests are another widely used measure preferred by many employers for personnel selection. Cognitive ability tests are used frequently due in part to their strong predictive validity evidence compared to other selection methods (Robertson & Smith, 2001; Schmidt & Hunter, 1998). Examples of cognitive ability tests used frequently by employers include tests of general mental ability (e.g. IQ tests), tests of practical intelligence, and tests of emotional intelligence. Tests of general mental ability (GMA) have been shown to be most predictive of job performance, with validity estimates averaging r = .65 according to a meta-analysis conducted by Schmidt and Hunter (1998). GMA has been shown to be predictive of acquisition of job knowledge (Schmidt, Ones, & Hunter, 1992) and of performance in job training programs (Schmidt & Hunter, 1998), so their relation to work performance is no surprise. Furthermore, research has shown that the use of specific cognitive abilities, like practical intelligence or tacit knowledge, do not add incremental prediction of job performance, over and above the use of GMA alone (Olea & Ree, 1994; Ree, Earles, & Teachout, 1994).

Personality measures did not become popular methods for personnel selection until the early 1990s. However, researchers and employers have shown personality plays a significant role in job performance (see Barrick & Mount, 1991; Ones, Viswesvaran, & Schmidt, 1993; Robertson & Kinder, 1993; Salgado, 1998). One type of personality measure often used in personnel selection is the Big 5 personality traits. Personality measures based on the Big 5 personality traits are centered on the lexical hypothesis – that personality characteristics most important in people's lives will eventually become a part of their language (John & Srivastava, 1999; Saucier & Goldberg, 1996). They posit that five broad dimensions - openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism - are commonly used to describe human personality and are therefore the most prominent personality traits. Examples of Big 5 personality measures include the NEO PI-R (Costa & McCrae, 2008) and Goldberg's Big 5 (Goldberg, 1992). Other examples of personality measures used by employers include the Minnesota Multiphasic Personality Inventory (MMPI), and the 16PF Questionnaire (Cattell, Eber, & Tatsuoka, 1970).

Predictive validity coefficients for personality measures, like the Big 5 factors, hover around .40, which is not far behind cognitive ability tests, with validity coefficients of .65 (Schmidt & Hunter 1998). There has also been much debate on whether entire personality measures (e.g. a measure of the Big 5 personality factors; Goldberg, 1992) or measures of single personality constructs (e.g. conscientiousness or agreeableness) are more appropriate for assessment. For instance, Mount and Barrick (1998), and Schneider, Hough, and Dunnete (1996) believe that narrow measures of specific personality factors, like conscientiousness, are most appropriate. Ones and Visweveran (1996), on the other

hand, believe that broad measures like the Big 5 should be utilized, because there is too much invalid variance in a homogeneous measure of a specific personality dimension, and reliability and criterion-related validity suffer when narrow traits are used (Jenkins & Griffith, 2004).

Employers often use some combination of these methods – interviews, cognitive ability tests, and personality measures - in order to select the best applicant for the job. Recently, some employers have started to shift away from the use of solely cognitive ability measures because of the stigma associated with them, and are starting to supplement cognitive ability with other measures, like personality tests, integrity tests, or biodata items (Hough & Oswald, 2000; Schmidt & Hunter, 1998). Employers want to avoid methods that may lead to adverse impact, and are starting to see the value of assessing important non-cognitive attributes necessary to be successful at the job.

Student Selection

Student selection procedures are concerned with selecting students that will be successful in college. As Nayer (1992) explains, "the purpose of admission [selection] procedures is to select students who will complete the educational program and go into professional careers" (p. 41). Student selection procedures often utilize similar predictors that employee selection methods use, including cognitive ability, personality measures, and to a lesser extent, interviews (Salvatori, 2001). Currently, the majority of colleges/universities use students' high school transcripts, grade point averages (GPA), and standardized test scores (e.g. ACT/SAT) to measure cognitive ability (Clinedinst & Koranteng, 2015; Peeters & Lievens, 2005). This is because research has demonstrated these indicators of cognitive ability are the most powerful predictors of academic

performance (Deary, Strand, Smith, & Fernandes, 2007; Hezlett et al., 2001; Richardson, Abraham, & Bond, 2012). For example, in studies conducted by College Board, SAT scores correlated in the mid .50s with first year college GPA (Shaw, 2015), and in studies conducted by Allen and Robbins (2010), ACT composite scores correlated .49 with first year college GPA. And, in a meta-analysis, Westrick, Le, Robbins, Radunzel, and Schmidt (2015) found the average correlation between high school GPA and cumulative college GPA was .58. Moreover, College Board has longitudinal data showing that SAT scores correlate (r = .33) with graduation rates (Burton & Ramist, 2001). However, such a heavy focus on cognitively-oriented measures has resulted in schools ignoring important skills and traits (e.g., personality, non-cognitive attributes) important for being successful in college. As a result, many schools are going away from the use of solely cognitive ability measures, and are starting to utilize personality measures and non-cognitive attributes as a supplement to cognitive ability.

The validity evidence of personality measures and interviews and their relation to college success is also well-established. For example, in a meta-analysis conducted by O'Connor and Paunonen (2007), researchers found several Big 5 personality factors were positively correlated with academic achievement (i.e. college GPA, exam grades, course grades). Specifically, the mean correlation between conscientiousness and academic achievement was r = 0.24, with a 95 percent confidence interval of 0.12 to 0.36; and the mean correlation between agreeableness and academic achievement was r = .06, with a 95 percent confidence interval of r = .01 to r = .11. While the magnitude of these effects is not large, these are examples that help provide evidence that personality factors play a role in academic performance and achievement (see also Komarraju, Karau, & Schmeck,

2009; Poropat, 2009; Trapman, Hell, Hirn, & Schuler, 2007; Van der Linden, Te Nijenhuis, & Bakker, 2010;).

In another meta-analysis by Goho and Blackman (2006), researchers found the mean correlation between academic admissions interviews and academic achievement (i.e., grade point average, exam scores) was r = .06, with effect sizes ranging from -.14 to .18 in the 19 studies included in the meta-analysis. Other studies and meta-analyses provide evidence that selection interviews weakly predict academic achievement (Eva, Rosenfeld, Reiter, & Norman, 2004; Goho & Blackman, 2006; Salvatori, 2001). While the effect sizes are not large and the validity evidence is mixed, many schools have traditionally used interviews as a selection method, and will continue to do so.

Some schools may also utilize resumes, essays and personal statements, letters of recommendation, or biodata items (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Peeters & Lievens, 2005; Sacks, 2000) but their relation to academic achievement in college is less established. However, concerns have been raised that academic admissions departments may be too heavily focused on cognitive measures like GPA and standardized test scores, and less focused on non-cognitive measures or predictors (Atkinson, 2001; Hough, Oswald, & Ployhart, 2001; Peeters & Lievens, 2005). Cognitively-oriented measures have been shown to have an adverse impact on minorities (see Hough et al., 2001; Peeters & Lievens, 2005; Whetzel, McDaniel, & Nguyen, 2008;), as research has demonstrated there are large mean differences in performance on cognitive ability tests for Whites, Blacks, Hispanics, and other races. For example, meta-analytic studies have shown that Blacks generally perform approximately one standard deviation lower than Whites on measures of cognitive ability while Hispanics generally

score about .6 to .8 standard deviations less than Whites (Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Schmitt, Clause, & Pulakos, 1996; Sackett and Wilk, 1994; Hunter & Hunter, 1984). Standardized tests like the ACT and SAT are also expensive, which may further marginalize minorities and their ability to enroll in prep courses or take the test multiple times to improve their scores (Atkinson, 2001). In addition, standardized tests like the ACT and SAT have a profound effect on how students regard themselves and their self-efficacy (Atkinson, 2001; Simpson, 2016; Dutro & Selland, 2012).

Such a heavy focus on cognitively-oriented measures has resulted in schools ignoring important skills and non-cognitive traits (i.e. personality, leadership, interpersonal skills) necessary to be successful in college (Peeters & Lievens, 2005). Non-cognitive measures, on the other hand, not only would add information over and above what a cognitive measure could provide (Lievens & Coatsier, 2002; O'Connel, Hartman, McDaniel, Grubb III, & Lawrence, 2007; Buyse & Lievens, 2011; Patterson et al. 2016), but they often do not have the negative stigma associated with them. There seems to be less evidence of the adverse impact of non-cognitive measures, and students enjoy taking them, rather than them being detrimental to their self-efficacy and confidence (Klassen, Durksen, Rowett, and Patterson, 2014).

Situational Judgment Tests

As a result of this movement, there has been a considerable push for alternative forms of student selection and admission that result in less adverse impact and do not ignore important skills and traits that are necessary to be successful in college (Peeters & Lievens, 2005; Atkinson, 2001). Growing evidence suggests Situational Judgment Tests (SJTs) may be one way to achieve this goal (see Lievens & Coatsier, 2002; Peeters &

Lievens, 2005; Oswald et al. 2004). SJTs are a type of psychological aptitude test in which respondents are presented with a short vignette/scenario about a particular situation, and are then asked to either: a) rate the effectiveness of the responses (i.e., knowledge SJTs), or b) indicate what response the participant would choose if they were in that situation (i.e., behavioral tendency SJTs). SJTs are widely used in personnel selection, and are increasingly used in student selection for graduate schools (see Buyse & Lievens, 2011; Koczwara et al. 2012; Patterson et al. 2016; Luschin-Egenbreuth, Dimai, Ithaler, Neges, & Reibnegger, 2015).

SJTs are designed to assess one's judgment in a specific situation, most commonly work or school-specific settings. SJTs were first used by the United States' military during World War II, as they needed an assessment tool that could help select competent soldiers to join the army (Northrop, 1989). Psychologists developed a 'job test' which consisted of realistic situations those in the armed forces would likely encounter while on the job. Each situation had several potential 'reactions' according to the specific threat or challenge, and potential recruits were asked to select the choice which they considered to be the most effective response (Northrop, 1989). The use of SJTs in this domain turned out to be a worthwhile investment for the U.S. armed forces, as it gave potential recruits the chance to see what kind of situations they would be likely to face, while also measuring recruits' judgment skills in important situations. This enabled military branches to select competent individuals with sound judgment skills.

There is sparse documented use of SJT-type instruments from the mid 1940's to the early 1990's. In 1990, Stephen Motowidlo and colleagues (1990) began to develop SJTs to predict job performance. Researchers developed "low-fidelity simulations" for

selecting entry-level managers in the telecommunications industry, which were defined as simulations that presented only a written description of the task stimulus. This is contrasted with "high fidelity simulations," which were simulations that presented a realistic representation of the task stimulus and elicit actual responses for performing the task (Motowidlo, Dunnette, & Carter, 1990). Researchers presented applicants with the low-fidelity work situations and five response options, and applicants were instructed to select the option they would be most likely and least likely to do. In a sample of managerial incumbents (n = 120), correlations ranged from r = .28 *to* r = .37 between the scores on the low-fidelity simulation and supervisory ratings of performance. These results provided evidence even low-fidelity simulations could predict job performance to a degree (Motowidlo, Dunnette, & Carter, 1990). Since it was the first SJT article published in a major personnel selection journal, this study revived interest and investigation into SJTs among researchers and practitioners (Whetzel & McDaniel, 2009). Use of SJTs as personnel selection instruments greatly expanded as organizations and businesses saw the potential of SJT use.

A typical SJT begins with a short vignette about a specific contextual situation. Sometimes, the vignette deals with a specific situation the respondent would likely face in the workplace or school, while other times it is a generic scenario that can be generalized to any job or school setting. An example of a SJT, with the vignette and several response options, is presented in figure 1.

You are facing a project deadline and are concerned that you may not complete the project by the time it is due. It is very important to your supervisor that you complete the project by the deadline. It is not possible to get anyone to help you with the work.

A. Ask for an extension of the deadline.
B. Let your supervisor know that you may not meet the deadline.
C. Work as many hours as it takes to get the job done by the deadline.
D. Explore different ways to do the work so it can be completed by the deadline.
E. On the day it is due, hand in what you have done so far.
F. Do the most critical parts of the project by the deadline and complete the remaining parts after the deadline.
G. Tell your supervisor that the deadline is unreasonable.
H. Give your supervisor an update and express your concern about your ability to complete the project by the deadline.
I. Quit your job.

Figure 1: An example of a Situational Judgment Test (SJT) item
Source: Whetzel & McDaniel (2009, pg. 188)

This SJT item above, from Whetzel and McDaniel (2009), includes a short vignette about a general situation likely to take place in many job or school settings. The item includes 9 response options with a variety of different reactions to the situation. In this example, the instructions were removed from the item, because there are actually two main types of instructions common among SJT items, which may have an effect on the construct being measured (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Nguyen, 2001), or on the validity and reliability of the SJT measure (Ployhart & Ehrhart, 2003). The two types of SJT response instructions are ‘knowledge’ instructions and ‘behavioral tendency’ instructions.

McDaniel and Nguyen’s (2001) SJT meta-analysis found that two different types of response instructions are most often used. The first type asks participants to choose the best response (i.e. *What is the best response?*), while the second type requires participants

to choose what they would do in that situation (i.e. *What would you do?*). McDaniel and Nguyen (2001), as well as McDaniel, Hartman, Whetzel, and Grubb (2007) decided to categorize the two differing types of response instructions as 'knowledge' and 'behavioral tendency' SJTs in order to provide a better response instruction taxonomy for researchers.

Knowledge SJTs. Knowledge SJTs are structured so respondents either choose the best response to the situation, or rank the response options from best to worst. (McDaniel & Nguyen, 2001; McDaniel et al., 2007). An example of a knowledge SJT, with the specific response instructions bolded and italicized for emphasis, is presented in figure 2.

> Your team is writing a business case for creating a new flavor of soda. You have a tight deadline, and everyone is really busy. You have been assigned to write the section of the report that describes the results of taste-test research. You are a new employee, and you are not sure if your section of the report is clear.
> ***What would be the best course of action to take?***
>
> A. Work as hard as you can on your section until the deadline.
> B. Write several versions of your section and submit them all to the team.
> C. E-mail a draft of your section to the entire team for their comments.
> D. Ask an experienced team member for advice on your draft.

Figure 2: An example of an SJT item with knowledge instructions
Source: Kyllonen, P. C., In Invitational Research Symposium on Technology Enhanced Assessments (2012)

Behavioral tendency SJTs. On the other hand, 'behavioral tendency' SJT items ask respondents what they would do in a particular situation. An example of a behavioral tendency SJT, with the specific instruction 'What would you do?' bolded and italicized for emphasis, is presented in figure 3.

Your team is writing a business case for creating a new flavor of soda. You have a tight deadline, and everyone is really busy. You have been assigned to write the section of the report that describes the results of taste-test research. You are a new employee, and you are not sure if your section of the report is clear.
What would you do?

A. Work as hard as you can on your section until the deadline.
B. Write several versions of your section and submit them all to the team.
C. E-mail a draft of your section to the entire team for their comments.
D. Ask an experienced team member for advice on your draft.

Figure 3: An example of an SJT item with behavioral tendency instructions Source: Kyllonen, P. C., In Invitational Research Symposium on Technology Enhanced Assessments (2012)

Comparing Knowledge and Behavioral Tendency SJTs. The difference between knowledge SJTs and behavioral tendency SJTs is akin to the maximal performance vs. typical performance argument first addressed by Cronbach (1960) and later by Sackett, Zedeck, and Fogli, (1988) and Barnes and Morgeson, (2007). Maximal performance is concerned with how someone performs when exerting as much effort as possible (i.e. when they are doing their best) and is therefore heavily related to cognitive ability (Cronbach, 1960; Sackett, Zedeck, & Fogli, 1988). Examples of maximal performance measures include high-stakes tests (defined as a test that is used to make important decisions; Plake, 2011) like the ACT/SAT and employment tests (Plake, 2011), and are generally used to make inferences about ability (Oostrom, De Soete, & Lievens, 2015). Knowledge SJTs, because they ask respondents to identify the best responses and are concerned with how the respondents are doing when exerting maximum effort, are considered maximal performance measures (McDaniel et al., 2007). As a result of the underlying cognitive nature of the response instructions (e.g. what is the best response) and their relation to maximal performance measures, knowledge SJTs have been shown

to be more highly correlated with measures of cognitive ability or intelligence (McDaniel et al., 2007; McDaniel & Nguyen, 2001, Whetzel & McDaniel, 2009). For example, a meta-analysis by McDaniel and colleagues (2007), found the estimated population correlation with cognitive ability was $r = .35$ for knowledge compared to $r = .19$ for behavioral tendency SJTs. Lievens, Sackkett, and Buyse (2009) found correlations of $r = .19$ between knowledge SJTs and cognitive ability, which was significantly higher ($z = 1.91$; $p < .05$) than the correlation between the behavioral tendency SJT and cognitive ability ($r = .11$).

On the other hand, typical performance is concerned with how one performs on a regular basis or how one typically behaves, and is therefore more dependent upon personality traits (Cronbach, 1960; Sackett, Zedeck, & Fogli, 1988). Typical performance measures, like personality tests, are generally used to make inferences about someone's personality, attitudes, or other non-cognitive attributes (Oostrom, De Soete, & Lievens, 2015). Behavioral tendency SJTs, because they ask what the respondent would do in that situation or what the respondent would typically do, are considered typical performance measures (McDaniel et al., 2007). As a result, they are shown to be more highly correlated with measures of personality (McDaniel et al., 2007; Nguyen, Biderman, & McDaniel, 2005). Specifically, McDaniel et al.'s meta-analysis (2007) found three correlations between the SJT scale score and specific personality traits were higher for the behavioral tendency SJT than knowledge SJT – agreeableness ($r = .37$ vs. $r = .19$), conscientiousness ($r = .34$ vs. $r = .24$) and emotional stability ($r = .35$ vs. $r = .12$). Similarly, Whetzel and McDaniel (2009) performed a meta-analysis and found that SJTs with behavioral tendency instructions had larger correlations with four of the Big 5

factors; agreeableness (r = .20 vs. r = .14), conscientiousness (r = .33 vs. r = .21), emotional stability (r = .13 vs. r = .02) and extraversion (r = .07 vs. r = .02).

The result of the type of instruction used – knowledge or behavioral tendency - not only has an effect on the construct being measured, it also influences the implications for how the SJTs are used and the possible adverse impact of use (Whetzel & McDaniel, 2009; McDaniel & Nguyen, 2001; McDaniel et al., 2007; Lievens, Sackett, & Buyse, 2009). For instance, in high-stake situations, behavioral tendency SJTs may be more susceptible to faking in that participants might perceive one of the options is better than the others, so they then choose that item as the best response in that situation (Peeters & Lievens, 2005; Nguyen, Biderman, & McDaniel, 2005), when they were supposed to be choosing what they would do. More discussion on the faking of SJT measures is presented in the 'considerations for SJTs' section.

Use of SJTs. Today, SJTs are most widely used in employee selection to predict job performance. Several individual studies and collective meta-analyses have shown that SJTs are effective by providing evidence for predicting future job performance (McDaniel et al. 2007; Whetzel & McDaniel, 2009). SJTs designed to predict job performance are usually based on a job analysis, and capture aspects of job/contextual knowledge, practical intelligence, and/or non-cognitive factors (Christian, Edwards, & Bradley, 2010; McDaniel et al., 2007, McDaniel & Nguyen, 2001). For example, SJTs have been developed around job performance constructs such as managerial and supervisory performance (Weekley & Ployhart, 2005; Hanson, 1994; Howard & Choi, 2000), task performance (Chan & Schmitt, 2002), contextual performance (Bergman, Donovan, Drasgow, & Overton, 2001; Clevenger, Pereira, Wiechmann, Schmitt, &

Schmidt-Harvey, 2001), and teamwork (Elias & Shoenfelt, 2001; McClough & Rogelberg, 2003; Mumford, Van Iddekinge, Morgeson, & Campion, 2008).

SJTs are now being used for medical (Luschin-Ebengreuth, Dimai, Ithaler, Neges, & Reibnegger, 2015), dental (Buyse & Lievens, 2011), and graduate school (Koczwara et al., 2012; Patterson et al. 2016) selection. Research in these areas have shown SJTs may provide incremental validity over and above measures of cognitive ability and personality. Because SJTs are often designed around non-cognitive, interpersonal, and intrapersonal factors, they often capture domains not traditionally covered by graduate admissions departments, which focus heavily on cognitive measures (Lievens & Coatsier, 2002; Peeters & Lievens, 2005; Oswald et al., 2004). For example, Lievens and Coatsier (2002) found SJTs offered incremental validity of 3.1% over and above cognitive ability measures in predicting students' final grades at the end of their first year of medical/dental school. With the large sample sizes colleges utilize, even small changes in R^2 values (like 3.1%) can result in better prediction. More discussion on the incremental validity of SJT measures – and other sources of validity evidence - is provided in the following section

Validity Evidence of SJTs

To date, limited research has attempted to develop SJTs for the purpose of predicting *undergraduate* 'collegiate success' (see Oswald et al., 2004). However, validity evidence has shown that SJTs do predict elements of work or school performance, and may be a worthwhile investment for college admissions departments to use as a student selection method. A summary of the validity evidence of SJT measures is provided below.

Construct validity. Construct validity is defined as the degree to which a test measures what it claims, or purports, to be measuring, (Cronbach & Meehl, 1955). According to the 2014 American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), & Joint Committee on Standards for Educational and Psychological Testing Standards for Educational and Psychological Testing (Hereafter referred to as the *2014 Standards*), it is one of the most researched sources of validity evidence. Evidence of construct validity of SJTs has been widely explored through a variety of studies and meta-analyses. Many studies claim SJTs are successful at predicting job/school performance because they measure one of three general constructs important for adequate job/school performance including job knowledge (Motowildo, Borman, & Schmit, 1997), practical intelligence (Sternberg, Wagner, & Okagaki, 1993), or general cognitive ability (McDaniel et al. 2001).

Convergent validity, the degree to which two measures are related, and discriminant validity, the degree to which two measures are not related, are two methods researchers use to provide evidence of construct validity (*2014 Standards*). Research has found evidence of both convergent and divergent sources with SJTs. For example, McDaniel et al. (2001) showed SJTs had average correlations of $r = .31$ with general cognitive ability, indicating SJTs may measure some aspects of general cognitive ability, but there is some variance unrelated related to cognitive ability. Evidence has also shown SJTs are correlated with measures of personality; specifically, three of the Big 5 measures: conscientiousness ($r = .27$), emotional stability ($r = .22$), and agreeableness ($r = .25$) (McDaniel et al., 2007). However, the specific type of response instruction has

been shown to have a clear moderating effect on the construct validity of SJTs (McDaniel et al. 2007).

McDaniel and Nguyen (2001) and McDaniel and colleagues (2007) investigated the moderating effects of response instructions on SJTs. Researchers compared two different response instructions of SJTs - knowledge and behavioral tendency - and their relation to the construct being measured. McDaniel et al. (2007) found SJTs with knowledge instructions correlated more highly with cognitive ability than SJTs with behavioral tendency instructions (average correlations of r = .32 for knowledge SJTs as compared to r = .17 for behavioral tendency SJTs), while SJTs with behavioral tendency instructions correlated more highly with measures of the Big 5; specifically conscientiousness (average correlations of r = .30 for behavioral tendency SJTs vs. r = .21 for knowledge SJTs), emotional stability (average correlations of r = .31 for behavioral tendency SJTs vs. r = .10 for knowledge SJTs), and agreeableness (average correlations of r = .33 for behavioral tendency SJTs vs. r = .17 for knowledge SJTs). This has clear effects for the construct validity of SJTs, because simply changing the response instructions seems to have an effect on the construct being measured.

However, one drawback from these two studies was that construct differences may have been due to differences in the specific content of the SJTs. More specifically, the scenario/vignette of the SJT, as well as the response choices, were not held constant when the instructions changed. Therefore, McDaniel and colleagues (2007) performed a second meta-analyses looking at studies in which the SJT content was held constant (the scenario and response choices) but the response instructions were changed. In these studies (k = 8) the same exact SJT items were administered twice, once with knowledge

instructions, and a second time with behavioral tendency instructions. Authors then correlated the results with measures of cognitive ability and measures of personality, and found SJTs with knowledge instructions had significantly larger correlations with cognitive ability (r = .28) than the same SJTs administered with behavioral tendency instructions (r = .17). Additionally, SJTs with behavioral tendency instructions had higher correlations with several of the Big 5 factors than the same SJT items with knowledge instructions. For agreeableness, the average correlation was r = .17 for behavioral tendency SJTs compared to r = .12 for knowledge SJTs; for conscientiousness the average correlation was r = .29 for behavioral tendency SJTs compared to r = .19 for knowledge SJTs; and for emotional stability the average correlation was r = .11 for behavioral tendency SJTs, compared to r = .02 for knowledge SJTs. Therefore, it seems the response instructions of SJTs has a clear moderating effect on the construct being measured, in that SJTs with knowledge instructions seem to capture cognitive abilities, while SJTs with behavioral tendency instructions seem to capture more aspects of personality.

Criterion-related validity. The extent to which a measure is related to a criterion is known as criterion-related validity, and is an area which has been highly explored among SJT research. According to the *2014 Standards*, the fundamental question of criterion validity is "how accurately do test scores predict criterion performance?" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Educations, 2014, p. 17). Several empirical studies have explored the criterion-related validity evidence of SJTs related to job performance (Chan & Schmitt, 1997; Motowildo et al. 1990). In addition, several comprehensive

meta-analyses (McDaniel et al. 2001; McDaniel et al. 2007; Christian et al. 2010) have summarized the criterion-related validity evidence of SJTs. The estimated population validity for SJTs in the McDaniel et al. (2001) meta-analysis was $r = .34$ across all the measures and samples; however, they also found that SJTs based on a job analysis had significantly more validity evidence ($r = .38$) than SJTs not based on a job analysis ($r = .29$). McDaniel et al. (2007) meta-analysis showed the overall validity coefficient relating the SJT with a measure of job performance was $r = .26$ ($n = 24{,}756$), while Christian et al. (2010) obtained criterion-related validity estimates ranging from $r = .19$ to $r = .43$, based on the construct being measured (teamwork, leadership skills, interpersonal skills, job knowledge and skills).

Evidence has shown SJT scores have a high degree of validity support for predicting certain criterions, especially when SJTs are based on a job analysis. However, there are a few caveats to the McDaniel et al. meta-analyses. First, neither study evaluated the response instructions of the SJTs. Additionally, the validity studies included in the meta-analyses were almost entirely concurrent validity studies in which the participants were job or school incumbents; very little research to date has focused on job or school applicants, which would capture evidence of predictive validity (McDaniel et al. 2001; McDaniel et al. 2007; Christian et al. 2010).

Limited research has examined the predictive validity of SJTs. One study by Livens and Coatsier (2002) examined the predictive validity of an SJT measure for student selection at a medical and dental school in Flanders, Belgium. Researchers found SJT scores predicted final scores of students at the end of the first year of medical and dental studies ($r = .23$) similar to cognitive ability tests ($r = .27$). Another study by Chan

and Schmitt (2002), among civil service employees, showed there is validity evidence for SJTs in predicting technical proficiency (r = .30), interpersonal facilitation (r = .27), and overall job performance (r = .30). While these few studies have demonstrated SJTs may be useful predictors for job performance and student success, more studies establishing the predictive validity of SJTs, and not just concurrent validity, need to be conducted.

Incremental validity. Another important source of validity evidence, incremental validity, is concerned with whether or not a measure increases the predictive validity beyond what is provided by an existing measure (Sackett & Lievens, 2008). Incremental validity has been a popular topic among SJT researchers, because businesses, organizations, and universities want to know whether administering SJTs for selection is a worthwhile investment. Weekly and Jones (1997, 1999) were first to perform incremental validity studies of SJTs, and they found significant incremental validity of SJTs over measures of cognitive ability and job experience, with change in R^2 values of R^2 = .021 and R^2 = .033 respectively(Weekley & Jones, 1997). This indicates the additional variable (in this case SJTs) significantly improved the prediction of the dependent variable (Miles, 2014). Chan and Shmitt (2002), also found incremental validity evidence for predicting job performance. In a sample of 160 civil service employees, seven predictors used to predict job performance (cognitive ability, Big Five factors, and job experience) were entered first, then followed by an SJT outcome. Researchers found that adding the SJT to the set of seven predictors resulted in a significant increase in criterion variance, in each of the four performance criteria. The change in R^2 values were R^2 = .05, R^2 = .08, R^2 = .03, and R^2 = .04 for task performance, motivational contextual performance, interpersonal contextual performance, and overall

job performance, respectfully. Furthermore, Clevenger et al. (2001), in three different samples, found evidence of incremental validity for SJTs over and above measures of cognitive ability, conscientiousness, job experience, and job knowledge, with change in R^2 values ranging from $R^2 = .016$ to $R^2 = .026$.

In a meta-analysis, McDaniel and colleagues (2007) summarized the results of several incremental validity studies that used hierarchical linear regressions with a correlation matrix of all variables (knowledge SJTs, behavioral tendency SJTs, a measure of cognitive ability, and a measure of the Big 5). They found evidence of incremental validity over and above measures of cognitive ability with both knowledge SJTs and behavioral tendency SJTs, although larger incremental validity evidence was found for behavioral tendency SJTs ($r = .05$ for behavioral tendency SJTs vs. $r = .03$ for knowledge SJTs). Evidence of incremental validity over a composite measure of the Big 5 was also found. Knowledge SJTs had a slightly higher level of incremental validity over measures of the Big 5 ($r = .07$) than behavioral SJTs ($r = .06$). This is likely a result of behavioral tendency SJT items being slightly more correlated with personality measures than knowledge SJT items (as demonstrated by McDaniel et al., 2007; Nguyen, Biderman, & McDaniel, 2005; McDaniel & Nguyen, 2001).

Researchers in this study also found evidence of incremental validity over a composite of cognitive ability and a measure of the Big 5, although validity estimates only ranged from $R^2 = .01$ to $R^2 = .02$. While the incremental validity estimates are relatively small, McDaniel et al. (2007) noted very few predictors would provide incremental validity evidence over a composite measure of *both* cognitive ability and the Big 5. Additionally, as Hunter, Schmidt, and Judiesch (1992), as well as Schmidt and

Hunter (1998) explained, even small increases in validity estimates can produce large increases in hiring efficiency for organizations, especially when they are summed across multiple hiring decisions. The overall results from these studies suggest administering a SJT may provide some useful level of prediction over and above measures of cognitive ability and/or personality, and therefore may be a worthwhile investment.

As a result, McDaniel et al. (2007) recommends a cognitive ability test should always be administered in selection contexts, and that if one wants an additional test to supplement cognitive ability, they suggest utilizing a Big 5 measure or a SJT to provide approximately the same amount of incremental prediction. Because SJTs are generally shorter and less time-consuming than many measures of the Big 5, it may be a worthwhile investment to administer a SJT for many organizations or universities.

Content validity. Evidence of content validity from SJTs is an area of research less explored than other aspects of validity. This is most likely because, as Anastasi notes, "the use of content validity in the evaluation of aptitude or personality tests has little to commend it" (Anastasi, 1980). However, according to the *Standards*, "important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure (AERA, APA, & NCME, 2014, p.14)." The current recommendation according to the *Standards* is that scales should be checked over by subject matter experts (SMEs) or by those who currently have the job/role, to ensure that all the situations and responses are realistic, and that the SJT items sample a wide domain of relevant situations. A subject matter expert is defined as an individual who displays a high level of expertise in the domain of interest (Osterlind, 1989). This process enables SMEs to provide evidence of content validity of the SJTs,

because it provides supporting evidence that the situations/scenarios and responses included in the SJTs are realistic and applicable to the criterion. It would also be advisable to have current job incumbents (those who currently have the job) serve as SMEs to look over scenarios and response options in order to provide evidence of the content validity of the SJT.

However, there is one defining feature of SJTs that may limit the evidence of content validity, according to Mosier (1947). SJTs are generally completed via paper and pencil measures or on a computer screen, and the participant must read a short vignette and then answer the question. As a result, the assumption that SJTs measure the criterion of job performance may not be appropriate. While the scenarios presented in the SJTs are related to job performance, it can't necessarily be assumed that performing well on *written* SJTs would result in successful job performance. However, Motowidlo and colleague's (1990) work on low-fidelity simulations, in which they found correlations ranging from $r = .28$ to $r = .37$ between the scores on the low-fidelity situation and supervisory ratings of job performance, provide evidence that that even low-fidelity simulations, without all of the equipment and role players required of the high-fidelity situations, can still be valid predictors of job performance.

Perhaps one way to improve the content validity evidence of SJT measures is to utilize video-based SJTs, in which actors are hired to act out the scenario, and then the video freezes at an important point and respondents have a limited amount of time to answer the question related to the scene presented. Video-based SJTs have shown to provide additional content validity evidence relative to written SJTs (Chan & Schmitt, 1997), and similar levels of concurrent and predictive validity (Lievens & Sackett, 2006;

Richman-Hirsch, Olson-Buchanan, and Drasgow, 2000). For example, Lievens and Sackett (2006) performed a study examining video-based versus written SJTs. Because of cost and technological concerns, a medical admissions exam at a University in Belgium had to transform their video-based SJT (which was being as a selection method) to a written format. This provided two samples to test video-based SJTs versus written SJTs, in which the content was held constant. Researchers found video-based SJTs had higher correlations (r = .35) with the criterion (college GPA) than correlations between the written SJTs and the criterion (r = .09).

Considerations of SJTs

While validity evidence exists for the use of SJT scores, there are several considerations for using SJTs in practice. One thing to consider is test-taker reactions. For example, in the admissions process, research has shown that individuals enjoy taking SJTs, and they may be helpful for students to help them prepare and think through situations that might be difficult in college. Klassen et al., 2014 found support for this idea, as 76.7% of applicants to a teacher training program who took SJTs as a selection assessment rated the content and format of the SJT measure favorably. Furthermore, Truxillo, Donahue, and Kuang (2003) hypothesized that applicant reactions to SJTs are favorable because they contain characteristics applicants want, including face validity and samples of job/school behavior.

Another reason colleges/universities could consider using SJTs is because they could be used for formative assessment/evaluation. For example, a student may be struggling in their first semester and placed on academic probation. It may be helpful for the student to meet with an academic advisor and go back and examine their SJT

answers, to analyze their decision-making ability and see if their judgment could be improved. This would make SJTs a useful measure for student retention. However, there are several caveats to SJTs that may make them difficult for college admissions departments to utilize. A summary of these limitations is provided below.

Subgroup differences. First, there is a chance SJTs may contain some level of subgroup differences among test participants, which may lead to adverse impact for minorities. Adverse impact is when members of one subgroup (e.g., members of minority groups, women) are selected disproportionately more or less often than members of other subgroups. According to the Uniform Guidelines on Employee Selection Criteria, the "threshold for adverse impact is established if one group is selected for the job less than 80% of that for the group with the highest selection rate" (Section 60-3, Uniform Guidelines on Employee Selection Procedure, 1978). While the majority of studies found little evidence of gender differences among SJT scores (Lievens & Coatsier, 2002), other studies have found evidence that SJTs tend to favor women (Mullins & Schmitt, 1998; Nguyen, 2004), but only slightly, with Cohen's *d* values ranging from *d* = .19 to *d* = .31 (Weekley & Jones, 1999). Whetzel and colleagues (2008) performed the most comprehensive review of studies over gender differences in SJT performance, and found that, in general, SJTs do tend to favor women, although the advantage was small (Cohen's *d* = -.11).

The meta-analysis performed by Whetzel and colleagues (2008) also explored racial differences in SJT scores. They found, on average, White respondents perform better on SJTs than Black (Cohen's *d* = .38), Hispanic (Cohen's *d* = .24), and Asian (Cohen's *d* = .29) respondents. As Cohen (1992) described, *d* = .10 is 'small' and *d* = .30

is ‘medium’, so the differences in SJT scores among different groups are significant. This means that White respondents performed slightly better on the SJTs than other races (Whetzel et al., 2008). However, further exploration in the meta-analysis revealed that mean differences in SJT scores may be the result of mean cognitive ability differences between races and not necessarily due to the SJT measure itself, but this area could use further exploration to determine if SJTs can address the limitations of cognitive measures.

Additionally, further subgroup differences in race were found when comparing knowledge SJTs and behavioral tendency SJTs. SJTs with knowledge instructions had slightly higher values of Cohen’s d than SJTs with behavioral tendency instructions. Cohen’s *d* values were $d = .39$ and $d = .34$ for knowledge and behavioral tendency SJTs, respectively, for Black–White, $d = .28$ and $d = .16$ for knowledge and behavioral tendency SJTs, respectively, for Hispanic–White, and $d = .30$ and $d = .27$ for knowledge and behavioral tendency SJTs, respectively, for Asian–White (Whetzel et al., 2008). The authors hypothesized this was a result of the mean cognitive ability differences between races, as knowledge SJTs seem to capture cognitive ability more than behavioral tendency SJTs.

Although differences in SJT scores seem to exist for race and gender, many researchers argue greater subgroup differences are found in traditional measures of cognitive ability. In fact, researchers and practitioners have continually searched for measures that have lower sub-group differences than measures of general cognitive ability, and research has repeatedly shown SJTs do, at minimum, have less race-based and gender-based group differences than cognitive ability measures (Chan &

Schmitt,1997; Motowidlo et al.,1990; Whetzel, McDaniel, & Nguyen, 2008; Weekley & Jones 1999). Therefore, many argue SJTs as a whole may be more appropriate as a selection instrument because there tends to be less subgroup differences for minorities.

Faking of SJTs. Another possible issue with using SJTs in high-stakes testing situations is there still may be some level of faking possible. Fakability is the idea of respondents choosing responses that misrepresent his or her self, and is often done so respondents can appear better than they actually are (Ziegler, Schmidt-Atzert, Buhner, & Krumm, 2007). With selection assessments in high-stakes testing situations, applicants have a strong motivation to get selected/hired, so they may respond in socially desirable ways on personality measures or SJTs. While SJTs are designed so there is no true correct answer, there are often responses that seem better than others (Bergman, Donovan, Drasgow, Henning, & Juraska, 2006). In particular, behavioral tendency SJTs are more susceptible to faking, as the instructions ask participants to indicate what they would do in that situation. (Lievens, Sackett, & Buyse, 2009; Nguyen, Biderman, & McDaniel, 2005). For example, if an individual is completing an SJT to be considered for a job, and behavioral tendency instructions (*What would you do?*) are given, the individual may try to fake their response and instead select the best possible response and ignore the directions all together, leading to an over-inflation of their score on the behavioral tendency SJT.

On the other hand, knowledge SJTs ask individuals to choose the best possible response or to rank the response from best to worst. As a result, knowledge SJTs are inherently difficult to fake, as the respondent is not choosing what *they* would do, they are instead choosing what they think is the best possible response. McDaniel et al.

(2007), Patterson et al. (2013), and Nguyen et al. (2005), suggest that in high-stakes situations, SJTs with knowledge instructions should be used so respondents are unable to choose socially desirable responses.

Furthermore, researchers and practitioners argue SJTs are less susceptible to faking than many other measures, especially those that use a Likert-type response or an agreement scale. With those measures, the 'good' responses are transparent to respondents, so in high-stakes situations, they are extremely susceptible to faking. FFollowing the advice of several researchers (McDaniel et al., 2007; Nguyen, Biderman, & McDaniel, 2005) using SJT items with knowledge instructions for high-stakes testing situations can reduce the ability of respondents to fake responses.

Psychometrics. Another issue with SJTs is that many traditional psychometric indications of quality are inappropriate for SJTs. Reliability of SJTs has received sparse exploration in research, because an appropriate reliability estimate for SJTs has yet to be agreed upon among researchers. Additionally, individual SJT items lack clear factor loadings, so homogeneous SJTs scales are difficult to create, and SJTs are often multi-dimensional (Whetzel & McDaniel, 2009; Chan & Schmitt, 1997, 2002). As a result, the heterogeneity of SJT measures makes Cronbach's alpha, the traditional measure of reliability for scales, a relatively inappropriate reliability index (Cronbach, 1951). Because SJTs generally measure a variety of constructs, and Cronbach's alpha is driven by the number of items in the scale and their inter-item correlations, reliability estimates are generally quite low. For example, Ployhart and Ehrhart (2003) found internal consistency coefficients of SJTs ranging from $\alpha = .23$ to $\alpha = .73$, while McDaniel et al. (2001) found in their meta-analysis internal consistency coefficients ranging from $\alpha = .43$

to α = .73. However, many authors performing research on SJTs continue to utilize Cronbach's alpha to measure the reliability of SJT scores, because it is the most utilized measure of reliability.

Whetzel and McDaniel (2009), however, argue test-retest reliability and parallel forms reliability are a more appropriate reliability estimate for SJTs. The few studies that have performed reliability studies on SJTs using these methods have found evidence supporting the reliability of these methods. For instance, Chan and Schmitt (2002) estimated parallel form reliability at r = .76 for an SJT measure of job performance, while Catano and colleagues (2012) found high test-retest reliability estimates (r = .82) compared to low estimates of Cronbach's alpha (α = .46).

Scoring of SJTs. Another possible issue with SJTs is scoring. As Bergman et al. (2006) noted, the difficulty with scoring arises because there is not a single, objectively correct answer. Several scoring methods have been proposed and utilized by researchers. Additionally, the response instructions of the SJTs effect how these scoring methods are enacted. For example, some response instructions require respondents to select the single best answer, to select the best and worst answers, to rank-order the response options, or to rate the response options on a continuous/Likert scale (St-Saveuer, Girouard, & Goyette, 2014). A discussion of different scoring methods for SJTs is provided below.

Empirically-based. One type of method used to score SJTs are empirically-based scoring methods, in which response options are scored according to their relationship with a criterion measure. These methods often require choosing a criterion measure (e.g. college GPA or supervisory ratings of job performance) and developing decision rules based on the criterion (which differ based on the response instructions; Bergman et al.,

2006). For example, researchers may require a certain relationship with the criterion measure (i.e. a correlation of r = .2 or greater) for the response option to be scored as correct (Mumford & Owens, 1987). If the threshold is met, then the response option is scored as correct. Or, researchers may develop decision rules based on the response options' ability to differentiate between people who score at different levels on a criterion variable (Lievens, Peeters, & Schollaert, 2008). In this method, response options often selected by individuals who perform highly on the criterion are then scored as correct, while response options often selected by low performing individuals are scored as incorrect.

After choosing the criterion and developing decision rules, SJTs are then administered to a large pilot sample. The final steps in the empirically-based scoring method are weighting each of the individual items (according to their relationship with the criterion) and cross-validating results (Bergman et al. 2006; Devlin et al. 1992; Hogan, 1994) with the sample collected. Therefore, these methods rely on actual data collected and the item responses and criterion scores of a sample, in order to select and weight items based on their ability to differentiate higher and lower performing criterion groups (Hogan, 1994), which is the strength of this scoring method.

However, several limits of this method exist. First, is that decision rules often require subjectivity (Whelpley, 2014). For example, researchers may decide that a response option needs to correlate .20 with the criterion in order for it to be scored as correct. If the item correlates at .19, it is really not that different from an item that correlates at .20, but the decision rule would still score it as incorrect. Additionally, the correlations between an individual item and the criterion measure are often small, so

large sample sizes are required in order to produce accurate estimations. As a result, researchers may reject items due to lack of statistical power, simply because their sample size is too small (Whelpley, 2014). Another limitation is the idea that empirically-based scoring is sample specific; the scores for the SJTs only apply to the specific sample in which the data was collected. This limits the generalizability of the SJTs and the ability to use them outside of the sample in which they were collected. A final limitation of this method is that empirical scoring methods are heavily dependent on the quality of the criterion variable (Campbell, 1990; Mumford & Owens, 1987). So, a bad criterion measure could result in poor estimates of validity.

Expert-based. The expert-based scoring method is where SMEs are consulted to make judgments about the response options. Responses are then scored based on the specific response instructions of the SJTs. In some expert-based scoring methods, consensus is reached among the SME pool for each of the response options. With this method, all SMEs agree that an option is correct or incorrect, and disagreements are resolved *a priori* (Legree, Psotka, Tremble, & Bourne, 2005). In other methods, determining the effectiveness of each response option is derived statistically, through empirical methods applied to the group of SMEs' decisions. This is often accomplished with measures of central tendency (e.g. mean of the respondent's ratings for response instructions that require rating multiple choice options, or mode of the respondents' selection for response instructions that require choosing one response option; Legree et al., 2005).

For example, in SJTs that are multiple choice format in which respondents are required to choose the best answer, SMEs examine each of the response options. If

researchers are using the consensus of SMEs, then the group of SMEs would get together and make a decision on the correct answer choice, prior to administering the SJT items. If the researchers are utilizing empirical methods to determine the best responses on the SJTs, the SJTs are administered to the group of SMEs, and researchers would then use the mode of the SMEs' selections in order to derive the 'correct' answer. In both cases, the 'best choice' is scored as correct (+1), while all other response options receive a 0. Bergman and colleagues (2006) assert the expert-based scoring method often yields higher validity coefficients as compared to other methods.

While the expert-based scoring method is among the most utilized scoring methods for SJTs, there are several difficulties associated with it. First, SMEs may not fully understand the situations in which the SJT is being applied. Therefore, it is recommended to select job or school incumbents (or at least someone with knowledge of the situations included on the SJTs) as the SMEs for the purposes of scoring the responses. Another difficulty is that SMEs may have difficulty reaching a consensus on the 'correct' response option, and no objective method exists to find a correct method when they disagree (Whelpley, 2014). Finally, SME responses may be unstable across a group of test takers, especially considering the small sample sizes often used for SMEs (Motowidlo, 1990; Whelpley, 2014). By chance or by systematic differences, a group of SMEs may be qualitatively different than another group of SMEs, leading to instability of the scoring key and poor generalization across samples.

Theoretical scoring. Theoretical scoring uses existing theory to determine which of the response options are effective and ineffective (Bergman et al. 2006). Response options that reflect the theory are scored as correct (+1), while response options that

contradict the theory are scored as incorrect (-1); all other response options receive a zero. Because of its relationship to theory, this scoring method has been shown to be more likely to generalize (Bergman et al., 2006). However, the main limitation of this method is that it may make the SJT items more susceptible to faking, as the response options are often transparent to the respondent (Hough & Paullin, 1994; Bergman et al., 2006).

Consensus-based scoring. The final scoring method, in which responses are scored as deviations from the consensus, defined by response distributions of the sample (Legree et al., 2005), is called consensus-based scoring. In this method, pilot data are collected with a sample; measures of central tendency (i.e. mean, mode) on the sample's SJT scores are then used to analyze responses and determine the 'correct' answer(s) for the SJTs. For example, the mean of the respondent's ratings may be used for response instructions that require rating multiple choice options, while the mode of the respondents' selection may be used for response instructions that require choosing one response option.

Legree et al. (2005) highlighted several benefits of the consensus-based scoring method. First, it allows scales to be scored for knowledge domains in which experts do not exist or are hard to find. As a result, consensus-based scoring may allow for the assessment of knowledge domains that have not been traditionally examined in psychological or educational research. Another benefit is that it allows for a shorter development cycle of the SJT scales, because expert responses are not required to conduct scoring. In turn, this may allow for lower costs associated with the development of the SJT, as expert judgments can be expensive to collect.

There are also a few limitations to this method. First, the validity estimates provided by this method are entirely sample specific, similar to the empirically based scoring method, meaning it is not applicable or generalizable across samples. Another limitation is the high amount of variance that may result within the sample. As Motowidlo and colleagues (1990) demonstrated by dropping items in which a high level of disagreement occurred, a high amount of variance can prevent researchers from identifying the 'best' response.

Research has shown the validity of SJT scores depend in part on the scoring method utilized; poor choices of the scoring method used could result in the conclusion that SJT's are not valid, when the only problem is the scoring method utilized (Bergman et al., 2006; Whetzel & McDaniel, 2009).

Proposed Study

Purpose of Proposed Study

The primary objective of this study is to create an SJT for the purposes of predicting undergraduate collegiate success, as measured by college GPA. Students' responses on the SJT items will be compared to the other data collected to determine if there is validity evidence of SJTs in predicting college GPA. Particular focus will be paid to the evidence of criterion-related validity in predicting college GPA, construct validity (in particular, evidence of convergent and discriminant validity), and incremental validity in predicting college GPA over and above the traditional measures that college admissions departments use (i.e. ACT/SAT scores and personality measures).

This study has several secondary aims, in addition to the primary aim of determining if there is validity evidence of SJTs in predicting college GPA. Prior research has shown that the various scoring methods used for SJTs influence their validity in predicting the criterion (Bergman et al., 2006; Legree et al., 2010). Therefore, one secondary objective is to determine if the validity estimates of SJTs with SME-based scoring in predicting college GPA are different than the validity estimates of SJTs with consensus-based scoring in predicting college GPA. Previous research has also demonstrated that SJTs result in less subgroup differences than measures of cognitive ability, as a result of the heterogeneous factor structure and the fact that they often encompass both cognitive and non-cognitive abilities (Whetzel, McDaniel, & Nguyen, 2008). Thus, another secondary objective is to determine whether the SJT measure had lower subgroup differences than the students' ACT/SAT scores.

Proposed Study and Research Questions

This study collected pilot data from undergraduate students at a large Midwestern university, in which students answered 14 SJT test items (7 knowledge SJTs as well as 7 behavioral tendency SJTs), a measure of the Big 5 (Goldberg's Big 5 Markers – short form), and demographic items, including students' self-report of their GPA and standardized test scores (ACT/SAT). The data collected will then be analyzed to answer the following research questions:

1. Do the different scoring methods used in this study (consensus-based vs. SME-based) result in different validity estimates in predicting college GPA?
2. Is there evidence of criterion-related validity for SJTs in predicting college GPA?

3. Does construct validity evidence, in the form of convergent and discriminant validity, provide evidence that the different response instructions (knowledge vs. behavioral tendency) result in the measurement of different constructs?
4. Is there evidence of incremental validity for the prediction of college GPA, over and above the traditional measures/predictors college admissions departments use (i.e. ACT/SAT scores and personality tests)?
5. Do the SJTs designed for the purposes of this study result in less subgroup differences than the students' ACT/SAT scores?

These research questions will be the focus of the study, and data collected will be analyzed to answer each. Thus, there are several hypotheses for this study, which are presented below.

- *Hypothesis 1:* The scoring method utilized (consensus-based vs. SME-based) will influence the reliability of the SJTs in the prediction of college GPA.

The scoring methods that were used for this analysis were the SME-based and the consensus-based scoring methods. The other two methods often utilized by researchers to score SJT items were not appropriate for this study. Theoretical-based scoring would have been difficult to utilize, because of the way the items were developed. While some theory was used to build the SJT items and response options, they were not specifically built with theoretical scoring in mind, making theoretical scoring inherently difficult. Additionally, as Bergman et al. (2006) showed, theoretical scoring can be quite difficult because the response options often have more face validity, which means they may be more susceptible to faking. The other scoring method often used in SJT research, empirically-based scoring, could also not be used, because cross-validation would not be

possible with the way data were collected. Participants did not receive all of the SJT items; instead, each participant received 14 SJT items, half of them with knowledge instructions (n = 7), and half of them with behavioral-tendency instructions (n = 7). In cross-validation, the full dataset is partitioned randomly into a number of subsets; as a result of the missing data (because the random assignment of SJT items), model estimations cannot be computed. Additionally, correlations with the criterion measure – college GPA – were quite low, so using the empirically-based scoring method, and the response options' correlation with the criterion measure, would not be appropriate.

Research has demonstrated that the different scoring methods influence the validity of SJTs in predicting the criterion (Bergman et al., 2006; Legree et al., 2010). To test the first hypothesis that the different scoring methods have an effect on the validities of the SJTs in the prediction of college GPA, the validity coefficients of SJTs with SME-based scoring will be compared to the validity coefficients of SJTs with consensus-based scoring. It is expected that the scoring methods will produce different validity estimates.

- *Hypothesis 2:* SJT scores (knowledge and behavioral tendency) will be positively correlated with the criterion measure, college GPA.
- *Hypothesis 3:* Behavioral tendency SJT scores will be more strongly correlated with Big 5 factors (extraversion, agreeableness, conscientiousness, emotional stability, intellect) compared to knowledge SJT scores
- *Hypothesis 4:* Knowledge SJT scores will be more strongly correlated with students' self-reported standardized test scores compared to behavioral tendency SJT scores.

The next three hypotheses focus on the validity evidence of SJTs. Hypothesis 2 is concerned with the evidence of criterion-related validity, and whether SJT scores positively correlate with the criterion measure, student's self-reported college GPA. If SJT scores are positively related to the student's college GPA, it will provide criterion-related validity evidence for the specific SJT (knowledge or behavioral tendency). Hypotheses 3 and 4 are related to sources of convergent and divergent validity evidence. Therefore, correlations will be used to provide evidence of convergent validity (knowledge SJTs being more strongly related to the students' ACT/SAT scores, and behavioral tendency SJTs being more strongly related to Big 5 factors) and discriminant validity (knowledge SJTs being less related to Big 5 factors, and behavioral tendency SJTs being less related with ACT/SAT scores).

- *Hypothesis 5:* Knowledge SJTs will provide significant incremental evidence for predicting of college GPA, over and above what is predicted by the Big 5 factors alone.
- *Hypothesis 6:* Behavioral tendency SJTs will provide significant incremental evidence for predicting college GPA, over and above what is predicted by the students' standardized test scores alone.

Hypotheses 5 and 6 focus on the potential incremental validity evidence provided by SJTs. Hierarchical regressions were used, with the primary measure (either students' ACT/SAT score or the Big 5 factors) entered as step 1, and the students' scores on the SJTs entered as step 2. Students' college GPA is the primary outcome measure of interest, thus change in R^2 values were used to determine if SJT scores significantly improve prediction of students' college GPA (Miles, 2014). As discussed, because

knowledge SJTs have been shown to be more highly related to cognitive ability, it is expected that knowledge SJTs would add a significant level of prediction over and above a personality measure. This is because there is less 'crossover' of constructs (i.e. the knowledge SJTs get at cognitive ability, so there is less overlap with personality measures). Moreover, because behavioral tendency SJTs have been shown to be more strongly related to personality measures, it is expected that behavioral tendency SJTs will add a significant level of prediction over and above the measure of cognitive ability.

- *Hypothesis 7:* SJTs will be have lower subgroup differences than traditional measures of cognitive ability (ACT/SAT scores).

Finally, hypothesis 7 is concerned with the performance of various subgroups on the SJTs. As discussed in the 'Considerations of SJTs' section, SJTs have been shown to result in less subgroup differences than measures of cognitive ability. Because it has been demonstrated that SJTs have a heterogeneous factor structure and often deal with both cognitive abilities and non-cognitive abilities, they should have lower subgroup differences (on race and gender) than measures of solely cognitive ability, like the students' composite ACT/SAT score (Whetzel, McDaniel, & Nguyen, 2008).

Method

Participants

The study sample consisted of 254 participants, with a mean age of 19.6 (range, 17 – 27 years). The majority of participants in the sample (56.3 %,) were female (*n* = 143). Participants' self-identified race/ethnicity was Caucasian (n = 179; 70.5%), Asian (*n* = 41; 16.1%), African American (*n* = 12; 4.7%), Hispanic (*n* = 10; 3.9%), and other (*n* = 8; 3.1%). The majority of the sample consisted of freshman (*n* = 106; 41.7%),

compared to sophomores (*n* = 80; 31.5%), juniors (*n* = 33; 13.0%), and seniors (*n* = 31; 12.2%).

A total of 24 classes were visited to recruit participants for this study. Courses and number of students participating from each course included 10 Agriculture and Leadership in Education (ALEC) courses (*n*=40), 4 sections in an introductory statistics course taught within Educational Psychology (*n*=35), and 10 sections of Skills for Academic Success taught within Educational Psychology (*n*=163). A total of 16 students did not indicate the course they were completing the study for. Students in the Educational Psychology (EDPS) courses received research credit for completing the study.

A total of 65 different majors were represented by the students completing the study, with the most popular majors business administration (*n* = 18), finance (*n* = 21), hospitality, restaurant, and tourism management (*n* = 23), and accounting (*n* = 18); 9 students (3.5%) who completed the study were double majors. Additionally, 23 (9.0%) students in the sample indicated they play a collegiate sport.

Design

This study used a mixed-factorial design. All participants were asked to make judgments of 14 various situations they are likely to face in college, with two different types of response instructions. Participants were randomly assigned 14 SJT items, half (*n* = 7) with knowledge instructions, and half (*n* = 7) with behavioral-tendency instructions. The decision was made to randomly assign seven SJTs of each type of response instruction to each participant so that comparisons could be made between the types of response instructions. Since the number of participants in the study was anticipated to be

fairly low, randomizing participants to either knowledge instructions or behavioral tendency instructions would have reduced the sample size of each group, limiting the power to determine an effect. Therefore, the between subjects factor was the type of instruction (knowledge or behavioral tendency), and the within subjects factor was the SJT items. Additionally, participants completed three questionnaires: a short, 50-item version of the Big 5 (Goldberg,), the Balanced Inventory of Desirable Responding, Short Form (BIDR-16, Hart et al. 2015), and demographics form.

Materials/Procedures

Students were recruited via presentations during EDPS and ALEC classes. A total of 24 classes were visited, with an average of 25-30 students per class. After in-person presentations, professors were provided a link to distribute to their students, inviting students to participate in the study. Some students received extra credit or research credit for their participation, but no one was required to participate. Participants completed all of the measures on Qualtrics, an online service for collecting and analyzing data (Qualtrics, 2005). The entire study took participants approximately 20-25 minutes to complete, and was completed online, at their own convenience.

Participants were informed the intent of the study was to assess college student's judgment in various situations they are likely to face in college, to develop and validate a predictive measure of collegiate success. Upon clicking the link to participate in the study, students were asked to complete an informed consent form and a demographics questionnaire. Students were asked to indicate their age, classification in school, expected graduation date, sex, race, major, hometown, and a few questions about their experiences in college (intent to drop out, satisfaction with college, etc.). Students were asked to

provide their ACT or SAT scores (self-report), as well as their cumulative college and high school GPAs. While self-reporting ACT/SAT scores and GPA are not ideal, research has shown this is an acceptable way to collect these data when other options are limited (Cole & Gonyea, 2009; Sanchez & Buddin, 2015). For example, Cole & Gonyea (2009), found high correlations ranging from .86 to .95 for self-reported test scores (SAT scale scores and ACT composite scores) with their actual test score.

Situational Judgment Tests. SJTs for this study were designed around the construct 'collegiate success,' using an iterative instrument development process. The approach used to develop the SJTs for this study is described in detail below.

Literature review. To build the SJT measures, first, a literature review and content analysis of predictors of college student success was performed. Resources like Pelligrino and Hilton (2012), Farrington et al. 2012, and a variety of studies related to the dimensions of college student performance (Oswald et al. 2014; Le, Casillas, Robbins, & Langley, 2005) were evaluated. Additionally, mission statements and educational objectives of various colleges and universities were also reviewed.

A total of seven interviews were also conducted with undergraduate students at various universities. Interview questions probed topics such as, "what kind of skills lead to success in college?" and "what situations have been difficult in college?" Students provided valuable information on situations that have been difficult in college, the skills they need to be successful, and anecdotal stories about their own personal experiences. Although students were obtained via a convenience sample, students were chosen because they are current undergraduate students and had perspective on the skills/abilities necessary to be successful in college. They provided an initial pool of potential

situations/scenarios as well as a list of skills and attributes needed to be successful in college.

After organizing the various pieces of information (i.e., literature review and interview data), a thematic analysis was conducted during the document review to identify salient constructs mentioned across all of the data sources. Constructs were then sorted into intrapersonal and interpersonal skills based on initial evidence in the document review and theoretical evidence discussed in Pelligrino and Hilton (2012). In their literature review, Pelligrino and Hilton (2012, p. 21) identified "three domains of competence" including cognitive, intrapersonal, and interpersonal. They posited that these three domains encompass all the differing dimensions of human behavior. Based on document and interview review, the constructs that emerged all seemed to be either intrapersonally-oriented (internal qualities like integrity, motivation, perseverance) or interpersonally-oriented (externalizing qualities like teamwork and social responsibility). So, the decision was made to sort them into these two domains; the third domain, cognitive, identified by Pelligrino & Hilton (2012), is generally captured by the students' standardized tests scores and GPA.

Within intrapersonal skills, three broad categories of constructs emerged. The first was adaptability and perseverance, and included planning, overcoming challenges, not giving up, and being flexible and adaptable. The next construct that emerged was ethics, and included integrity, character, values, honesty, not cheating, and having respect for others. The final interpersonal construct that emerged was motivation and work ethic. This construct included working hard, being goal-oriented, time management, and prioritizing.

Within interpersonal skills, two broad categories of constructs were chosen. The first was teamwork and working well with others, which included conflict resolution, communication, planning, organizing, and leadership. The other broad construct that emerged was social responsibility and citizenship/involvement. This construct included cultural responsibility, dealing with adversity, civic responsibility, and respect for others. While it was difficult to sort all of the information and constructs into categories, it was important for the purpose of designing and writing the SJT items. However, it is important to note that these 'domains' and constructs are not perfect, and there is some overlap between them.

Development of SJTs. Next, scenarios and response options related to the identified constructs were developed, with the help of several graduate students. In the first round, 21 SJT scenarios were designed, each with three to five response options. In this study, cognitive interviews were performed with a group of graduate students in the student affairs graduate program, who were defined as SMEs for the purposes of this study. This group was chosen as SMEs because they have sufficient knowledge of what it takes to be successful in college (i.e., they are successful undergraduate students due to their obtaining a bachelor's degree) and they are participating in a graduate program dedicated to supporting the academic and personal development of individuals who are attending college (i.e., student affairs).

Cognitive interviews, a technique often used by survey methodologists (Collins, 2003; Schwarz, 2007), were then used to gather feedback about the SJT items and response options. In cognitive interviews, the survey administrator walks step-by-step through the survey/measure with a participant, gaining feedback on the participant's

ability to comprehend the question being asked, their ability to retrieve from memory relevant information related to the question, their decision process in trying to answer the question, and the response process to see how well the respondent can match his/her decision into a response category given by the survey question (Tourangeau, 1984; Sudman, Bradburn, & Schwarz, 1996).

After performing several cognitive interviews (n = 10; 5 males and 5 females), updates were made to the SJT items and response instructions. For example, participants provided feedback on whether the situations and response options were relevant and appropriate for undergraduate students, recommendations for wording changes and more appropriate response options, and suggestions for new scenarios. Participants also provided feedback on the response instructions and formatting of the SJT items.

Prior to the first round of cognitive interviews, there were 21 SJT scenarios with 3-5 response options per scenario. At the conclusion of the first round, four scenarios were dropped. Graduate students felt these four scenarios were not representative of situations all students would encounter in college. For example, one scenario dealt with fraternity/sorority issues, and graduate students consistently commented not all students are involved in a fraternity or sorority, and several students do not even know what they entail. Additionally, several response options were edited or changed. Participants also explained it might be a good idea to standardize the number of response options for each scenario, so, the decision was made to have 4 response options for each scenario. A second round of cognitive interviews (n= 8) was performed with another group of student affairs graduate students, which led to additional changes and updates for the SJT

scenarios and response options. At the conclusion of all cognitive interviews, 15 situational judgment scenarios were retained with four response options each.

Operational SJTs. At administration, each participant received 14 SJTs, with seven of them randomly selected to have knowledge instructions, and the remaining seven to have behavioral-tendency instructions. The 'behavioral tendency' and 'knowledge' SJT items were identical. For each SJT item, the only difference between conditions was response instructions (i.e., between-subjects variable). For the 'behavioral tendency' items, the directions were: "What would you do? Rank order the responses from 1 (what you are most likely to do) to 4 (what you are least likely to do)." For the 'knowledge' SJT items, the directions were: "What is the best response? Rank order the responses from 1 (best response) to 4 (worst response)." An example of a behavioral tendency and knowledge SJT item used in this study are presented in figure 4.

Behavioral tendency SJT:

Each week, your math professor assigns several practice problems for the material that is being covered, but they do not count towards your grade. ***What would you do?*** *Rank order the responses from 1 (what you are most likely to do) to 4 (what you are least likely to do).*

- Complete all of the practice problems, as they will help you learn the material and pass the tests
- Do enough of the practice problems so that you know how to do each type of problem
- Do not do the practice problems, but if you struggle on the first test, then start doing them
- Get together with a group of friends each week, and complete the practice problems

Knowledge SJT:

Your grade for a class is based on only three (non-comprehensive) exams and five homework assignments. The professor for the class does not record attendance, and posts all of the slides, notes, and homework assignments online. The professor also provides review days, the class before each test. ***What is the best response?*** *Rank order the responses from 1 (best response) to 4 (worst response).*

- Attend the class anyways. You will learn more simply by listening to the professor's lectures.
- Do not attend class, but if you fail an exam or a homework assignment, then start attending class.
- Do not attend the class and instead spend the time you would be in class studying the material and working on homework assignments.
- Only attend classes that are review days, so that you can hear the professor review the material and you can ask any questions you might have.

Figure 4. Examples of behavioral tendency and knowledge SJTs used in the study

Scoring of SJTs. Several considerations for the scoring methods utilized in this study were discussed in the 'considerations of SJTs' section, and further discussion is provided below.

SME-based scoring. The SME-based scoring method was chosen because it is the most common scoring method utilized for SJTs, there was a readily available pool of SMEs that could be used for the purposes of the study, and as Bergman et al. (2006)

found, the expert-based scoring method often yields higher validity coefficients as compared to other methods. For this study, SMEs were defined as graduate students who were enrolled in a student affairs graduate program at the University the data were collected.

A total of 19 subject matter experts filled out the SJT items and were used for the SME-based scoring method. Each of the 19 graduate students completed the SJT items independently, and were not included in the other rounds of feedback or cognitive interviews. The response instructions that were provided were knowledge instructions [i.e. "*what is the best response*? Rank order the responses from 1 (the best response) to 4 (the worst response)]. These instructions were chosen because the goal is to achieve the 'best' response, not what they would do in that situation. After each of the subject matter experts completed the measure, frequencies of their selections for each of the response options were computed, and the 'best' order was decided based on the mode (highest frequency) of responses. For the most part, subject matter experts agreed on the 'best' and 'worst' answer choices. There were some disagreements on the middle response options; theory/face validity was used in these instances (n=3) to determine the better and worse response.

Consensus-based scoring. The consensus-based scoring method was also used to determine if it is as effective as the SME-based scoring method. For this method, descriptive statistics and frequency of responses were computed for each SJT item. The response option that the most amount of participants chose as the 'best' response was given a 1, the response option that was the second most frequent response was given a 2,

and so on. So, there was a subject-matter expert-based order and a consensus-based order for each of the 15 SJT items (for both knowledge and behavioral tendency SJTs). Participants were then scored based on each of the scoring methods; they received a 1 for each response option that matched the scoring method used. Therefore, for each of the items, scoring could range from 0 to 4 (all four matched the order of the scoring method utilized). Scale scores were then computed for both knowledge SJTs and behavioral tendency SJTs. Since participants were randomly assigned seven knowledge SJT items and seven behavioral tendency SJT items, the scale score could range from 0-28 for knowledge SJTs and 0 - 28 for behavioral tendency SJTs.

Additional measures. Participants were also asked to fill out a 50-item measure of the Big 5 personality domains from Goldberg et al., (1992) after completing the SJT items. In this measure, participants were given 50 statements and asked to rate how accurate statements were of themselves, on a scale from 1 to 5, ranging from very inaccurate to very accurate. Examples of items included, "I feel comfortable around people," and "I pay attention to details" (Goldberg et al., 1992). Participants also filled out a short form of the Balanced Inventory of Desirable Responding (BIDR-16, Hart et al. 2015). This questionnaire was a 16-item short-form measure of the 40-item Paulhaus Balanced Inventory of Desirable Responding (BIDR, Paulhaus, 1984), which incorporates self-deceptive enhancement, defined as honest but overly positive responding, and impression management, defined as bias towards pleasing others. If a participant's score on the 16-item BIDR measure is high, this means that the participant may be providing socially-desirable responses, and the validity of that participant's survey scores could be compromised (Hart et al. 2015).

Results

Test of Hypotheses

Hypothesis 1. The first hypothesis posited that the scoring method utilized (consensus-based vs. SME-based) would have an effect on the validity evidence of the SJTs in the prediction of college GPA. The different scoring methods – SME-based versus consensus-based – produced *some* different validity estimates, as seen in Table 1. The results showed that knowledge SJTs with SME-based scoring did not significantly predict college GPA, $\beta = .12$, $t(210) = 1.78$, $p > .05$. Therefore, knowledge SJTs with SME-based scoring did not explain a significant proportion of variance in college GPA, $R^2=.015$, $F(1, 210) =3.15$, $p > .05$. It was also found that knowledge SJTs with consensus-based scoring did not significantly predict college GPA, $\beta = .12$, $t(210) = 1.77$, $p > .05$. Knowledge SJTs with consensus-based scoring did not explain a significant proportion of variance in college GPA, $R^2=.015$, $F(1, 210) =3.12$, $p > .05$.

Table 1.

Standardized Regression Coefficients Predicting College GPA

Predictor	B	SE	β
SJT – Knowledge (SME-based scoring)	.015	.008	.121
SJT – Knowledge (Consensus-based scoring)	.013	.008	.121
SJT-Behavioral Tendency (SME-based scoring)	.012	.009	.086
SJT-Behavioral Tendency (Consensus-based scoring)	.016	.008	.135*

*Significant at the p < 0.05 level.

For behavioral tendency SJTs, differing estimates were found. The results showed that behavioral tendency SJTs with SME-based scoring did not significantly predict college GPA, $\beta = .086$, $t(210) = 1.257$, $p > .05$. As a result,behavioral-tendency SJTs with SME-based scoring did not explain a significant proportion of variance in college GPA, $R^2 = .007$, $F(1, 210) = 1.580$, $p > .05$. However, behavioral tendency SJTs with consensus-based scoring significantly predicted college GPA, $\beta = .135$, $t(210) = 1.982$, $p < .05$. Therefore, behavioral-tendency SJTs with consensus-based scoring explained a significant proportion of variance in college GPA, $R^2 = .018$, $F(1, 210) = 3.93$, $p > .05$.

Hypothesis 2. Means, standard deviations, and zero-order correlations are provided in Table 2 for both knowledge and behavioral tendency SJTs, which were used for hypotheses 2, 3, and 4.

Table 2.

Correlations between variables

Variables	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Extraversion	254	36.23	7.96	-										
2. Agreeableness	254	38.18	6.27	.26	-									
3. Conscientiousness	254	35.49	6.02	.12	.23**	-								
4. Emotional Stability	254	29.86	6.37	.07	.02	.05	-							
5. Intellect	254	34.45	5.60	.12	.36**	.28**	.02	-						
6. ACTSAT Composite Score	216	24.35	3.91	.04	.03	.14	.20*	.30**	-					
7. CollegeGPA	215	3.23	.55	.05	.19*	.22**	.09	.08	.34**	-				
8. SJT – Knowledge (Subject Matter Expert-based scoring)	251	12.46	4.74	.19*	.13	.24**	.10	-.04	.23**	.10	-			
9. SJT-Knowledge (Consensus-based scoring)	251	13.31	5.097	.22**	.09	.23**	.08	-.04	.09	.08	.81**	-		
10. SJT – Behavioral Tendency (Subject Matter Expert-based scoring)	251	12.55	4.0	.07	.17*	.29**	.01	.03	.06	.11	.50*	.46*	-	
11. SJT-Behavioral Tendency (Consensus-based scoring)	251	13.02	4.67	.09	.20**	.42**	.01	.08	.04	.16*	.68**	.56*	.73*	-

** Correlation is significant at the .01 level (2-tailed)
* Correlation is significant at the .05 level (2-tailed)

The second hypothesis stated that scores on the SJT measures (both knowledge and behavioral tendency) would be positively related to the students' self-reported GPA. This hypothesis was tested by reviewing the correlation between SJT scores and students' GPA, for both of the scoring methods used in the study. For knowledge SJTs, the correlation between students' scores on the SJTs with SME-based scoring and college GPA was non-significant, $r(213) = .12$, $p > .05$. Similarly, the correlation between students' scores on knowledge SJTs with consensus-based scoring and college GPA was also non-significant, $r(213) = .12$, $p > .05$. For behavioral tendency SJTs, the correlation between students' scores on the SJTs with SME-based scoring and their college GPA was non-significant, $r(213) = .09$, $p > .05$. However, the correlation between students' scores on behavioral tendency SJTs with consensus-based scoring and college GPA was significant, $r(213) = .14$, $p < .05$. Thus, there is limited support for the hypothesis providing evidence of criterion-related validity of the SJTs.

Hypotheses 3 and 4. These hypotheses were concerned with the construct validity of the SJTs. Specifically, the hypotheses stated that scores on the behavioral tendency SJT would be more highly correlated with Big 5 factors (extraversion, agreeableness, conscientiousness, emotional stability, and intellect) than knowledge SJT scores (hypothesis 3), and scores on the knowledge SJT would be more highly correlated with students' standardized test scores than behavioral tendency SJT scores (hypothesis 4). This would provide evidence of convergent and discriminant validity and also provide evidence that the two response instructions often utilized in SJTs are, in fact, measuring different constructs.

For behavioral tendency SJTs, hypothesis 3 was partly supported. As seen in Table 2, significance was found for two correlations between Big 5 factors and the students' scores on the behavioral tendency SJTs with SME-based scoring. Specifically, conscientiousness, $r(249) = .28$, $p < .01$) and agreeableness, $r(249) = .29$, $p < .05$. The pattern was similar for consensus-based scoring, as the same two Big 5 factors were significant; conscientiousness, $r(249) = .38$, $p < .01$ and agreeableness $r(249) = .27$, $p < .01$. For knowledge SJTs, on the other hand, significance was reached for three of the Big 5 factors with SME-based scoring; extraversion, $r(249) = .15$, $p < .05$, agreeableness, $r(249) = .22$, $p < .01$, and conscientiousness, $r(249) = .27$, $p < .01$. Similarly, significance was reached for the same three Big 5 factors with consensus-based scoring; extraversion, $r(249) = .15$, $p < .05$, agreeableness, $r(249) = .19$, $p < .01$ and conscientiousness, $r(249) = .27$, $p < .01$. Thus, there is mixed evidence of the convergent/discriminant validity of behavioral tendency SJT items.

For knowledge SJTs, hypothesis 4 was supported. As seen in Table 2, there was a significant correlation between the students' scores on the knowledge SJTs with SME-based scoring and the students' ACT/SAT composite score, $r(214) = .25$, $p < .01$. The correlation between the students' scores on the knowledge SJTs with consensus-based scoring and the students' ACT/SAT composite score was also significant, $r(214) = .14$, $p < .01$. For behavioral tendency SJTs, neither of the correlations between SJT scores and the students' ACT/SAT composite score were significant; $r(214) = .08$, $p > .05$ for SME-based scoring and $r(214) = .03$, $p > .05$ for consensus-based scoring. Thus, the hypothesis is supported for the evidence of convergent and discriminant validity of knowledge SJTs.

Hypotheses 5 and 6. These two hypotheses were concerned with establishing the incremental validity of the SJTs. Hypothesis 5 was tested using hierarchical regression analysis, with the Big 5 factors (extraversion, agreeableness, conscientiousness, emotional stability, and intellect) entered as step 1 of the regression, and scores on the knowledge SJT entered as step 2. As seen in Table 3, knowledge SJTs with SME-based scoring did not add to the prediction of the students' college GPA over and above the Big 5 factors ($\Delta R^2 = .01$, $p > .05$). Similarly, knowledge SJTs with consensus-based scoring, as seen in Table 4, did not add to the prediction of the students' college GPA over and above Big 5 factors ($\Delta R^2 = .01$, $p > .05$). As a result, hypothesis 5 was not supported.

Table 3.
Hierarchical regression results for knowledge SJTs with SME-based scoring predicting college GPA over and above Big 5 factors.

	b	**S.E.**	**β**	**t**	***p***	**R^2**	**ΔR^2**	**F**
DV = College GPA								
Model 1:						.091	.091**	4.142**
Extraversion	-.003	.005	-.042	-.616	.538			
Agreeableness	.012	.007	.132	1.779	.077			
Conscientiousness	.024	.006	.255	3.679	.000			
Emotional Stability	.006	.006	.068	1.013	.312			
Intellect	-.005	.007	-.052	-.709	.479			
Model 2:						.092	.001	3.495**
Extraversion	-.003	.005	-.046	-.660	.510			
Agreeableness	.011	.007	.125	1.665	.097			
Conscientiousness	.023	.007	.244	3.409	.001			
Emotional Stability	.006	.006	.070	1.042	.299			
Intellect	-.005	.008	-.048	-.647	.518			
SJT – Knowledge – SME-based scoring	.005	.008	.040	.573	.568			

** Correlation is significant at the .01 level
* Correlation is significant at the .05 level

Table 4.
Hierarchical regression results for knowledge SJTs with consensus-based scoring predicting college GPA over and above Big 5 factors

	b	**S.E.**	**β**	**t**	***p***	**R^2**	**ΔR^2**	**F**
DV = College GPA								
Model 1:						.091	.091**	4.142**
Extraversion	-.003	.005	-.042	-.616	.538			
Agreeableness	.012	.007	.132	1.779	.077			
Conscientiousness	.024	.006	.255	3.679	.000			
Emotional Stability	.006	.006	.068	1.013	.312			
Intellect	-.005	.007	-.052	-.709	.479			
Model 2:						.092	.001	3.495**
Extraversion	-.003	.005	-.046	-.659	.511			
Agreeableness	.011	.007	.125	1.657	.099			
Conscientiousness	.023	.007	.244	3.499	.001			
Emotional Stability	.006	.006	.070	1.038	.301			
Intellect	-.005	.008	-.045	-.608	.544			
SJT – Knowledge – consensus-based scoring	.004	.008	.041	.575	.566			

** Correlation is significant at the .01 level
* Correlation is significant at the .05 level

Hypothesis 6 was also tested using hierarchical regression analysis, with the students' ACT/SAT composite scores entered as step 1 of the regression, and scores on the behavioral tendency SJT entered as step 2. As seen in Table 5, behavioral tendency SJTs with SME-based scoring did not add to the prediction of students' college GPA over and above the students' ACT/SAT composite scores ($\Delta R^2 = .006$, $p > .05$). However, behavioral tendency SJTs with consensus-based scoring, as seen in Table 6, added to the prediction of the students' college GPA over and above the students' ACT/SAT scores ($\Delta R^2 = .025$, $p < .05$). As a result, there is partial support for hypothesis 6.

Table 5.
Hierarchical regression results for behavioral tendency SJTs with SME-based scoring predicting college GPA over and above cognitive ability

	b	**S.E.**	**β**	**t**	***p***	**R^2**	**ΔR^2**	**F**
DV = College GPA								
Model 1:						.068	.068**	13.507**
ACT/SAT score	.038	.010	.262	3.675	.000			
Model 2:						.074	.006	7.326**
ACT/SAT score	.037	.010	.256	3.583	.000			
SJT Behavioral Tendency – SME-based scoring	.011	.010	.076	1.065	.288			

** Correlation is significant at the .01 level
* Correlation is significant at the .05 level

Table 6.
Hierarchical regression results for behavioral tendency SJTs with consensus-based scoring predicting college GPA over and above cognitive ability

	b	**S.E.**	**β**	**t**	***p***	**R^2**	**ΔR^2**	**F**
DV = College GPA								
Model 1:						.068	.068**	13.507**
ACT/SAT score	.038	.010	.262	3.675	.000			
Model 2:						.094	.025*	9.457**
ACT/SAT score	.037	.010	.254	3.601	.000			
SJT Behavioral Tendency – consensus-based scoring	.019	.008	.159	2.259	.025			

** Correlation is significant at the .01 level
* Correlation is significant at the .05 level

Hypothesis 7. This hypothesis stated that SJT scores would have lower subgroup differences than traditional measures of cognitive ability. In the sample, Cohen's *d* was calculated to determine the differences in the scores on the SJT measure and the students' ACT/SAT scores between racial subgroups. In this sample, sample sizes were very small for minority groups (*n* = 12 for African American, *n* = 41 for Asian, and *n* = 10 for Hispanic, and *n* = 8 for other). As a result, minority groups were collapsed into one group; thus, there were two groups for comparison – White (*n* = 179) and a minority group (*n* = 71). Although this is not ideal, many studies have shown this is an acceptable form of comparison, as much of the literature in the selection and assessment industry compares the scores of White respondents to the scores of minority respondents (Roth et al., 2001).

In the sample of data collected, Cohen's *d* was calculated to determine if there were any score differences between white respondents and minority respondents on the SJTs. As shown in Table 7, both knowledge SJTs and the behavioral tendency SJTs had

Table 7.
Racial group comparison (White vs. minority group) of SJTs and cognitive ability

Measure	Mean Score Difference	Cohen's *d*
Knowledge SJT – SME-based scoring	2.689	.523
Knowledge SJT – Consensus-based scoring	2.061	.371
Behavioral Tendency SJT – SME-based scoring	1.453	.330
Behavioral Tendency SJT – consensus-based scoring	1.453	.314
ACT/SAT Score	1.413	.290
College GPA	0.215	.361

larger scoring differences based on race (White vs. minority group) as compared to measures of cognitive ability. In general, White participants performed better on all the measures included in the study. The value for Cohen's d was largest for knowledge SJTs, with values of d = .523 for SME-based scoring and d = .371 for consensus-based scoring. According to Cohen (1992), these values fall in the medium to large range. Behavioral tendency SJTs had slightly lower values of Cohen's d, with values of d = .330 for SME-based scoring and d = .314 for consensus-based scoring. Finally, mean differences were computed for the students' ACT/SAT composite score, where the lowest mean difference was found (d = .290).

Discussion

Non-cognitive skills and attributes – like adaptability, motivation, working well with others, and social responsibility – have been shown to be important skills necessary for success in college (see Pelligrino & Hilton, 2001; Farrington et al. 2012, Oswald et al. 2014; Le, Casillas, Robbins, & Langley, 2005; Atkinson, 2001). However, colleges and universities have traditionally focused on cognitive predictors, and have struggled to find an accurate and objective way of measuring these non-cognitive skills, often resorting to personality measures or interviews, or deciding not to measure them at all. This study detailed the development of an SJT measure that may be useful for college admissions departments to aid in the selection of students or for student retention and development.

Validity Evidence of SJTs

The first several hypotheses were concerned with the validity evidence of the SJTs. A discussion on the validity evidence found from the SJTs created for this study is provided below.

Evidence of criterion-related validity. Hypothesis 1 stated that scores on the SJTs would be positively related to the students' GPA. Partial support of this hypothesis was found, as the correlation between students' scores on behavioral tendency SJTs with consensus-based scoring and college GPA was significant. Furthermore, the regression model predicting college GPA with behavioral tendency SJTs with consensus-based scoring as the predictor was also significant. This provides evidence that at least one of the SJT measures (with a particular scoring method and response instructions) aided in the prediction of college GPA in the sample. Previous research and meta-analyses have found that the estimated validity coefficient for SJTs in predicting the criterion hover around $\beta = .30$ (McDaniel et al., 2001); while validity estimates obtained in this study were lower than the estimates found in the McDaniel et al. (2001) meta-analysis, it may have been a result of the way data was collected for the study. Students self-reported their own college GPA, which means some students may not have accurately reported their true GPA. Additionally, a majority (approximately 42 percent) of the sample were freshman in college. Since data collection took place during these students' second semester, their GPA is calculated from only a few classes. The students may have struggled initially in their transition to college or may have had some difficult classes to start their college careers, leading to lower GPAs than what may eventually be obtained by them.

Evidence of construct validity. Hypotheses 3 and 4 were concerned with the evidence of convergent and discriminant validity. Prior research has shown SJTs with behavioral tendency instructions are more strongly related to personality measures than cognitive ability, and SJTs with knowledge instructions are more strongly related to

cognitive ability than personality (McDaniel et al., 2007; McDaniel & Nguyen, 2001), even when the content of the SJTs is held constant. Therefore, it was expected that behavioral tendency SJT scores would be more strongly correlated with Big 5 factors (extraversion, agreeableness, conscientiousness, emotional stability, intellect) and knowledge SJT scores would be more strongly correlated with the students' standardized test scores.

Partial support of the convergent validity evidence of *behavioral tendency* SJTs was found. Specifically, the correlation between SJT scores with behavioral tendency instructions and college GPA was significant for two of the Big 5 factors, conscientiousness (r =.28 with SME-based scoring and r = .38 with consensus-based scoring) and agreeableness (r = .29 with SME-based scoring and r = .27 with consensus-based scoring). However, the correlation between SJT scores with knowledge instructions and college GPA was significant for three of the Big 5 factors – extraversion, agreeableness, and conscientiousness, although the magnitude was much smaller for knowledge SJTs (correlations ranging from .15 - .27) than it was for behavioral tendency SJTs (correlations ranging from .27 - .38).

Support for the convergent validity evidence of *knowledge* SJTs was found. Significant correlations between the students' scores on knowledge SJTs and the students' ACT/SAT composite scores were found (r = .25 for SME-based scoring, and r = .14 for consensus-based scoring), while significance was not reached for the correlations between the students' scores on behavioral tendency SJTs and ACT/SAT composite scores. These results align with prior research conducted by McDaniel et al. (2007), as they found correlations between Big 5 factors ranging from .30 - .33 for

behavioral tendency SJTs, and from .10 - .21 for knowledge SJTs, which is similar to the correlations found in this study. However, this study went a step further, and held the content of the SJTs constant, only changing the response instructions. The results provide evidence that even when the content of the SJTs is held constant, changing response instructions can have an effect on the construct being measured. Specifically, using knowledge response instructions (*what is the best response?*) will likely result in the SJTs being more highly related to cognitive ability, while using behavioral tendency response instructions (*what would you do?*) will likely result in the SJTs being more highly related to personality.

Evidence of incremental validity. The focus of hypotheses 5 and 6 was the incremental validity evidence of the SJTs. It was expected that knowledge SJTs would provide significant incremental evidence for predicting college GPA, over and above what is predicted by the Big 5 factors alone. It was also expected that behavioral tendency SJTs would provide significant incremental evidence for predicting college GPA, over and above what is predicted by the students' standardized test scores alone. However, it was found that knowledge SJTs did not meaningfully add to the prediction of college GPA, over and above what was predicted by the Big 5 factors. On the other hand, behavioral tendency SJTs with consensus-based scoring did meaningfully add to the prediction of the students' college GPA, over and above what was predicted by the students' standardized test scores alone. Therefore, it appears that behavioral tendency SJTs, combined with a measure of cognitive ability, can be used to aid in the prediction of college GPA. This aligns with previous research, as McDaniel et al. (2007) found greater evidence of incremental validity for behavioral tendency SJTs than knowledge

SJTs. Furthermore, their recommendation to first administer a cognitive ability test, and then if one wants an additional test to supplement cognitive ability, to use an SJT measure with behavioral tendency instructions, seems to align with the findings of this study. This is likely because there is less 'crossover' of constructs - behavioral tendency SJTs are more strongly related to personality measures, so there is less overlap with cognitive ability measures.

The significant incremental change in R^2 value obtained in this study was .025. While this value does not seem large or even meaningful on the surface, even small increases (e.g. .01 or .02) in validity estimates can produce large increases in hiring/selecting efficiency for organizations when they are summed across multiple hiring/selection decisions (Hunter et al., 1992; Schmidt & Hunter, 1998). Because colleges and universities are generally dealing with a very large number of applicants, significant increases in selecting efficiency can occur even with slightly improving selection measures, resulting in better applicants being selected into the school.

Subgroup Differences

Another focus of the study was to determine if subgroup differences could be minimized by using SJTs. Previous research has shown that, in general, SJTs have less race-based and gender-based subgroup differences than traditional measures cognitive ability (Chan & Schmitt,1997; Motowidlo et al.,1990; Whetzel, McDaniel, & Nguyen, 2008; Weekley & Jones 1999). However, on average, White respondents still perform better on SJTs than Black (Cohen's d = .38), Hispanic (Cohen's d = .24), and Asian (Cohen's d = .29) respondents (Whetzel et al., 2008). Prior research has also shown that SJTs with knowledge instructions had slightly higher values of Cohen's d than SJTs with

behavioral tendency instructions, likely because of their stronger relation to cognitive ability measures.

In this sample, sample sizes were very small for minority groups (n = 12 for African American, n = 41 for Asian, and n = 10 for Hispanic, and n = 8 for other). As a result, minority groups were collapsed into one group, resulting in two groups for comparison – White (n = 179) and a minority group (n = 71). Medium to large values of Cohen's d (range .314 to .523) were found for the SJTs with the various scoring options; the value for Cohen's d was smaller (d = .330 and d = .314) for behavioral tendency SJTs than it was for knowledge SJTs (d = .523 and d = .371) which aligns with previous research. However, the values were all higher than the Cohen's d value that was found for the students' ACT/SAT scores (d = .290). This indicates that race-based differences may actually have been lower in the students' ACT/SAT scores. However, when looking at the composition of the sample, this may have been a result of the high proportion of study abroad students in which English was not their first language, included in the sample. Therefore, the SJT items may not have been appropriate for this group of respondents because of the language barrier and a variety of cultural differences. While time was spent pre-testing the items and performing cognitive interviews in order to shed light on issues similar to this, it appears that some of the SJT items may have been inherently difficult for some respondents. This is an area of future research for this study – to determine which SJT items were troublesome for minority respondents.

Use of SJTs

Beyond the validity evidence of the SJTs, there are several advantages to using SJTs as a form of student selection or retention. First, some students may struggle on

standardized tests, but have some non-cognitive skills or attributes that would lead them to be successful in college. Because colleges and universities focus so heavily on cognitive predictors, these students may not get admitted into a college/university. SJTs may be one way to overcome this problem. They would allow colleges and universities to capture the non-cognitive skills and attributes these students possess, which may help them overcome their poor standardized test results. For example, several students included in the sample had very low ACT/SAT scores. However, their college GPA was relatively high, at least when predicting what their college GPA should be from their standardized test results. After going back to examine their SJT scores, the students may have a higher GPA than was expected because of their high scores on the SJTs. A potential explanation is those students may be high on some of the non-cognitive skills and attributes that are leading them to be successful in college, but evidence of this would not have been captured if schools were only concerned with the traditional cognitive predictors of college success.

Another benefit of using SJTs is their ability to be used for formative assessment/evaluation. For example, a student may be struggling in their first semester of college and may even be placed on academic probation. It might be worthwhile for the student to meet with an academic advisor and examine their SJT responses, to analyze their decision-making ability and see if their judgment in college could be improved. This would make SJTs a useful measure for student retention. Up to this point, there are limited instances of SJTs being used in this way, making this a possible direction for future research.

Differences based on the scoring method and response instructions. Another area of interest in this study was to examine whether validity differences occurred as a result of the scoring method and response instructions utilized. In this study, two scoring methods were used—the SME-based scoring method, and the consensus-based scoring method. Different validity estimates were found based on the scoring method used. Interestingly, the consensus-based method actually produced higher validity estimates and was more likely to be significant during analyses. This is a meaningful finding for several reasons. First, the consensus-based scoring method allows SJTs to be scored for domains in which experts do not exist or are hard to find. The implication is consensus-based scoring may allow for the assessment of knowledge domains that have not been traditionally examined in psychological or educational research. Another benefit is it allows for a shorter development timeframe of the SJTs, because SMEs are not required to score the items. This may allow for lower costs and time associated with the development of the SJT, as expert judgments can be expensive and time-consuming to collect.

Differences in validity estimates were also found as a result of the response instructions used. In this study, the content of the SJT items was held constant, while only the response instructions varied (*what would you do?* vs. *what is the best response?).* The results provide evidence that even when the content of the SJTs is held constant, simply changing the wording of the response instructions can have an effect on the construct being measured. Specifically, requiring respondents to choose the best response will likely result in the SJTs being more highly related to cognitive ability, while requiring respondents to choose what they are most likely to do in that situation will likely result in

the SJTs being more highly related to personality traits. Additionally, it was found that knowledge SJTs did not meaningfully add to the prediction of college GPA, while behavioral tendency SJTs with consensus-based scoring *did* meaningfully add to the prediction of the students' college GPA. Therefore, it appears that behavioral tendency SJTs, when combined with the students standardized test scores, have more potential to aid in the prediction of college GPA.

These findings also have several implications for how the different types of SJTs are used in practice. First, if one wants to avoid socially desirable responding or reduce respondents' ability to fake on the SJTs, it is suggested to use knowledge response instructions in order to limit that ability. However, in some circumstances, SJTs with behavioral tendency instructions may be preferred. For instance, when there is already a cognitive test or predictor being used, it may be best to combine that with a behavioral tendency SJT, as there is less crossover of constructs as compared to SJTs with knowledge instructions.

Limitations

While there are several strengths of this study, there are also some limitations. First, students self-reported their ACT/SAT scores and college GPA. As a result, students may have not been fully honest in their self-report, or may not have remembered their ACT/SAT score or known their current GPA. While self-reporting ACT/SAT scores and GPA are not necessarily ideal, research has shown this is an acceptable way to collect these data when other options are limited (Cole & Gonyea, 2009; Sanchez & Buddin, 2015). Cole & Gonyea (2009), found high correlations ranging from .86 to .95 for self-

reported test scores (SAT scale scores and ACT composite scores) with their actual test score, and similar correlations have been found for self-reported GPA and actual GPA.

Another limitation of this study was using self-reported college GPA as the sole criterion measure. While college GPA is often used in the literature as a proxy for 'college success,' there were some limitations to its use in this study. For example, a high proportion of students in the sample were college freshman (41.7%), which means their GPA is calculated from only a few classes. Students may have struggled with their transition to college or may have some very difficult (or very easy) classes in the one semester they attended the school so far, which would result in their GPA being biased. If the study were to be conducted again, a variety of criterion measures should be used. Possibilities include absenteeism, using the grades of particular classes, or other established measures of 'college success' like the Collegiate Learning Assessment (CLA, Hardison & Vilamovska, 2009).

An additional limitation of the study is traditional psychometrics were unable to be used in this study. Because individual SJT items lack clear factor loadings, homogeneous SJTs scales are difficult to create, and SJTs are often multi-dimensional (Whetzel & McDaniel, 2009; Chan & Schmitt, 1997, 2002). As a result, the heterogeneity of SJT measures makes Cronbach's alpha a relatively inappropriate reliability index (Cronbach, 1951). Additionally, because participants were randomly assigned a set of knowledge and behavioral-tendency SJT items in order to make comparisons between the types of response instruction, very few participants received the same set of items. While Whetzel and McDaniel (2009) argued that test-retest reliability and parallel forms reliability are more appropriate reliability estimates for SJTs, these estimates could also

not be computed for this study. Participants only completed the SJT items at one time point, so test-retest reliability would not be possible. And, only one set of SJT items was created, whereas an 'alternate form' is needed to compute parallel forms reliability.

The final limitation of this study is that large differences in race/ethnicity were found for the SJT measures created in this study. This was likely a result of the large population of study abroad students in which English is not their first language included in the sample. These students have qualitatively different experiences as American college students, so the content included in the SJTs may not have been entirely applicable for them. This is likely why there were high estimates of subgroup differences in the sample.

References

Aamodt, M. (2012). *Industrial/organizational psychology: An applied approach*. Nelson Education.

Allen, J., & Robbins, S. (2010). Effects of interest–major congruence, motivation, and academic performance on timely degree attainment. *Journal of Counseling Psychology*, *57*(1), 23.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014*). Standards for educational and psychological testing.* Washington, DC: AERA.

Anastasi, A. (1980). Abilities and the measurement of achievement. *New Directions for Testing and Measurement*, 5, 1-10.

Atkinson, R. (2001). *Standardized tests and access to American universities.* The 2001 Robert H. Atwell Distinguished Lecture, presented at the annual meeting of the American Council on Education, Washington, DC.

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance*, *20*(3), 259-274.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, *44*(1), 1-26.

Bergman M. E., Donovan M. A., Drasgow F., Overton R. C. (2001). *Assessing contextual performance: Preliminary tests of a new framework.* Paper presented at the 16th

Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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.

Burton, N. W., & Ramist, L. (2001). Predicting Success in College: SAT® Studies of Classes Graduating since 1980. Research Report No. 2001-2. *College Entrance Examination Board*.

Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental student selection. *Journal of Dental Education*, *75*(6), 743-749.

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

Cascio, W. F. (1991). *Applied Psychology in Personal management*. London: Prentice Hall.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, *20*(3), 333-346.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). Handbook for the 16 personality factor questionnaire. *Champaign, IL: Institute for Personality and Ability Testing*.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of

assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143–159.

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

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

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, *86*(3), 410.

Clinedinst, M., & Koranteng, A. M. (2017). State of college admission. Retrieved from the National Association of College Admission Counseling website: https://www.nacacnet.org/globalassets/documents/publications/research/2015soca.pdf

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98-101.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229-238.

Costa, P. T., & McCrae, R. R. (2008). The revised neo personality inventory NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*, *2*(2), 179-198.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297−334.

Cronbach, L. J. (1960). *Essentials of Psychological Testing: 2d Ed*. Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13-21.

Devlin, S.E., Abrahams, N.M. and Edwards, J.E. (1992) Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. *Military Psychology,* 4, 119–136.

Dutro, E., & Selland, M. (2012). “I like to read, but I know I'm not good at it”: Children's perspectives on high-stakes testing in a high-poverty school. Curriculum Inquiry, 42(3), 340- 367. doi:10.1111/j.1467-873X.2012.00597.x

Edwards, W. R, Schleicher D. J. (2004). On selecting psychology graduate students: Validity evidence for a test of tacit knowledge. *Journal of Educational Psychology, 96*, 592–602.

Elias, D. A., & Shoenfelt E. L. (2001, April). *Use of a situational judgment test to measure teamwork components and their relationship to overall teamwork performance*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: the multiple mini-interview. *Medical education*, *38*(3), 314-326.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance--A Critical Literature Review*. Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.

Fleishman, E. A., & Mumford, M. D. (1988). Ability requirement scales. In S. Gael (Ed.) *The job analysis handbook for business, industry, and government* (pp. 917-935). New York, NY: John Wiley & Sons.

Goho, J., & Blackman, A. (2006). The effectiveness of academic admission interviews: An exploratory meta-analysis. *Medical Teacher, 28*(4), 335-340.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26-42.

Hanson MA. *(1994*). *Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army.* (Doctoral dissertation). Retrieved from Dissertation Abstracts International, 56(2-B), 1138.

Hardison, C. M., & Vilamovska, A. M. (2009). *The Collegiate Learning Assessment: Setting standards for performance at a college or university* (Vol. 663). Rand Corporation.

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The balanced inventory of desirable responding short form (BIDR-16). *Sage Open*, *5*(4), 2158244015621113.

Hezlett, S.A., Kuncel, N. R., Vey, M. A., Ahart, A. M., Ones, D. S., Campbell, J. P., et al. (2001, April). *The predictive validity of the SAT: A meta-analysis.* Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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

Hooker, E. (1959). What is a criterion? *Journal of Projective Techniques, 23*(3), 278-281.

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

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future--Remembering the past. *Annual Review of Psychology*, *51*(1), 631-664.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1-2), 152-194.

Howard A, Choi M. (2000). How do you assess a manager's decision-making abilities? The use of situational inventories. *International Journal of Selection and Assessment, 8,* 85–88.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*(2), 184-190.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72.

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*(1), 28.

Jenkins, M., & Griffith, R. (2004). Using personality constructs to predict performance: Narrow or broad bandwidth. *Journal of Business and Psychology, 19*(2), 255-269.

John, O. P., & Benet-Martinez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In C. M. Judd (Ed.) *Handbook of research methods in social and personality psychology* (pp. 339 – 367). Cambridge: Cambridge University Press.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.) *Handbook of personality: Theory and research* (pp. 102-138). New York, NY: Guildford Press.

Klassen, R., Durksen, T., Rowett, E., & Patterson, F. (2014). Applicant reactions to a situational judgment test used for selection into initial teacher training. *International Journal of Educational Psychology, 3*(2), 104-124.

Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education*, *46*(4), 399-408.

Komarraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the Big Five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, *19*(1), 47-52.

Kyllonen, P. C. (2012). *Measurement of 21st century skills within the common core state standards.* In Invitational Research Symposium on Technology Enhanced Assessments, 7-8.

Le, H., Casillas, A., Robbins, S. B., & Langley, R. (2005). Motivational and skills, social, and self-management predictors of college outcomes: Constructing the Student Readiness Inventory. *Educational and Psychological Measurement*, *65*(3), 482-508.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.) *Emotional intelligence: An international handbook* (pp. 155-179). Cambridge, MA: Hogrefe & Huber.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*(4), 245-257.

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

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.

Luschin-Ebengreuth, M., Dimai, H. P., Ithaler, D., Neges, H. M., & Reibnegger, G. (2015). Situational judgment test as an additional tool in a medical admission test: an observational investigation. *BMC Research Notes*, *8*(1), 81.

McClough, A. C, & Rogelberg S. G. (2003). Selection in teams: An exploration of the teamwork knowledge, skills, and ability test. *International Journal of Selection and Assessment, 11,* 56–66.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1969). *Position analysis questionnaire*. West Lafayette, IN: Purdue Research Foundation.

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, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*(1-2), 103-113.

McDaniel, M.A., Whetzel, D.L., Schmidt, F.L., and Maurer, S. (1994). The validity of employment interview: A comprehensive review and meta- analysis. *Journal of Applied Psychology, 79*, 599- 617.

Miles, J. (2014). R squared, adjusted R squared. *Wiley StatsRef: Statistics Reference Online*.

Moscoso, S. (2000). Selection interview: A review of validity evidence, adverse impact and applicant reactions. *International Journal of Selection and Assessment*, *8*(4), 237-247.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, *7*(2), 191-205.

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.

Motowildo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human performance*, *10*(2), 71-83.

Mullins, M. E., & Schmitt, N. (April, 1998). *Situational judgment testing: Will the real constructs please present themselves?* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology.

Mumford T., Van Iddekinge C., Morgeson F., Campion M. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93*, 250–267.

Mumford, M.D., & Owens, W.A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. Applied Psychological Measurement, 11, 1-31. doi: 10.1177/01466216870110010

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.

Nayer, M. (1992). Admission criteria for entrance to physiotherapy schools: How to choose among many applicants. *Physiotherapy Canada 44*: 41–46.

Nguyen, N. T. (2004). *Response instructions and construct validity of a Situational Judgment Test.* Proceedings of the 11th Annual Meeting of the American Society of Business and Behavioral Sciences, Las Vegas, NV.

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.

Northrop, L. C. (1989). *The Psychometric history of selected ability constructs*. Washington, DC: United States Office of Personnel Management

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual differences, 43*(5), 971-990.

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment*, *15*(1), 19-29.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology*, 79, 845–851.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17*(6), 609-626.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679.

Oostrom, J. K. & De Soete, B. & Lievens, F. (2015). Situational judgment testing: A review and some new developments. In J. K. Oostrom & I. Nikolaou (Eds.) *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice* (pp. 172-189) London, UK: Psychology Press.

Osterlind, S. J. (1989). *Constructing test items.* Norwell MA: Academic Press

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.

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., & Martin, S. (2016). The predictive validity of a situational judgement test and multiple-mini

interview for entry into postgraduate training in Australia. *BMC Medical Education*, *16*(1), 87.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598-609.

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.

Pellegrino, J., & Hilton, M. L. (2012). *Education for Life and Work. Transferable Knowledge and Skills for the 21st Century (*Rep. B8767.) Washington, DC.

Plake, B. S. (2011). Current state of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 11-26). Washington, DC: American Psychological Association.

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.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, *135*(2), 322.

Qualtrics, L. L. C. (2005). Qualtrics (version September, 2018) [computer software]. Provo, Utah, USA: Qualtrics.

Ree, M. J., Earles, J. A., & Teachout, M. S.(1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79*, 518–524.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological bulletin*, *138*(2), 353.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*(6), 880.

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology*, *66*(3), 225-244.

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of occupational and Organizational psychology*, *74*(4), 441-472.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, 297–330.

Sackett P. R, Wilk S. L. (1994). Within group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929-954.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419-450.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*(3), 482.

Sacks, P. (2000). Standardized minds: The high price of America's testing culture and what we can do to change it. *National Association of Secondary School Principals NASSP Bulletin*, *84*(616), 118.

Salgado, J. F. (1998). Big Five personality dimensions and job performance in army and civil occupations: A European perspective. *Human Performance, 11*(2-3), 271-288.

Salgado, J. F., & Cooper, C. L. (1999). Personnel selection methods. In C. L. Cooper, & I. T. Robertson (Eds.), *International review of industrial and organizational* (pp. 1-54). New York, NY: John Wiley & Sons.

Salgado, J. F., & Moscoso, S. (2008). Personnel selection in industry and public administration: From the traditional view to the strategic view. *Psychologist Papers*, *29*(1), 16–24.

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*, *6*(2), 159-175.

Sanchez, E., & Buddin, R. (2015). *How accurate are self-reported high school courses, course grades, and grade point average*. ACT Working Paper Series WP-2015-03). Iowa City, IA: ACT, Inc.

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.) *The five-factor model of personality: Theoretical perspectives* (pp. 21-50). New York, NY: Guilford Press.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of

research findings. *Psychological Bulletin, 124*(2), 262.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43*(1), 627-670.

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy of range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61*, 827-868.

Schmitt N., Clause C, Pulakos E. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C. L. Cooper, & I. T. Robertson (Eds.), *International review of industrial and organizational psychology*, (pp. 115-137). New York, NY: John Wiley & Sons

Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior*, *17*(6), 639-655.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 21*(2), 277-287.

Shaw, E. J. (2015). An SAT® Validity Primer. *College Board*.

Simpson, C. (2016). *Effects of standardized tests on students' well-being.* Harvard Graduate school of Education. Retrieved from https://projects.iq.harvard.edu/files/eap/files/c.simpsoneffectsoftestingonwellbeing516.pdf

Smit-Voskuijl, O. F. (2005). Job Analysis: Current and future perspectives. In Evers, A., Anderson, N., & Smit-Voskuijl, O. (Eds.), *Blackwell Handbook of Personnel Selection*. Malden, MA: Blackwell.

Sternberg, R. J., Wagner, R. K., & Okagaki, L. (1993). Practical intelligence: The nature and role of tacit knowledge in work and at school. In J. M. Puckett & H. W. Reese (Eds.) *Mechanisms of everyday cognition* (pp. 205-227). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

St-Sauveur, C., Girouard, S., & Goyette, V. (2014). Use of Situational Judgment Tests in Personnel Selection: Are the different methods for scoring the response options equivalent? *International Journal of Selection and Assessment*, *22*(3), 225-239.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Tourangeau, R. (1984). Cognitive sciences and survey methods. *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: The National Academies Press.

Trapmann, S., Hell, B., Hirn, J. O. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Journal of Psychology*, *215*(2), 132-151.

Truxillo, D. M., Donahue, L. M., & Kuang, D. (2003). Job sample tests, performance testing, and competency testing. In S. N. Haynes, E. M. Heiby, & M. Hersen (Eds.) *Comprehensive handbook of psychological assessment* (pp. 345-370).

Van der Linden, D., Te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315-327.

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

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*(3), 679-700.

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

Weekley, J.A., Ployhart, R.E., & Holtz, B.C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J.A Weekly, & R.E. Ployhart, (Eds.), *Situational judgment tests: Theory measurement and application* (pp. 157 – 182). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educational Assessment*, *20*(1), 23-45.

Whelpley, C. E. (2014). How to Score Situational Judgment Tests: A Theoretical Approach and Empirical Test. (Doctoral Dissertation). Retrieved from VCU Scholars Compass.

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.

Ziegler, M., Schmidt-Atzert, L., Buhner, M., & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: Questionnaire, semi-projective, and objective. *Psychology Science*, *49*(4), 29.