This is to certify that the
thesis entitled

SOURCE OF EXPERTISE IN SCORING KEY DEVELOPMENT AS A DETERMINANT OF THE NATURE OF THE CONSTRUCTS MEASURED

presented by

ABIGAIL K. QUINN

has been accepted towards fulfillment
of the requirements for the

M.A. degree in Psychology

Neal Schmitt
Major Professor's Signature

5/15/2009
Date

SOURCE OF EXPERTISE IN SCORING KEY DEVELOPMENT AS A DETERMINANT OF THE NATURE OF THE CONSTRUCTS MEASURED

By

Abigail K. Quinn

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Psychology

2009

ABSTRACT

SOURCE OF EXPERTISE IN SCORING KEY DEVELOPMENT AS A DETERMINANT OF THE NATURE OF THE CONSTRUCTS MEASURED

By

Abigail K. Quinn

The purpose of this study was to examine whether the selection of subject matter experts (SMEs) who provide scoring judgments during the development of a situational judgment test (SJT) has the potential to affect which constructs the SJT ultimately measures. It was hypothesized that different groups of SMEs would have different implicit theories of performance because of shared traits and experiences. Three groups of SMEs, graduate students, resident advisors, and undergraduate students provided scoring judgments for an SJT designed to measure college student performance. The scoring judgments made by the three groups were fairly similar. Correlations of scores based on the scoring keys developed by each group with a variety of performance criterion measures demonstrated that the scoring keys predicted performance equally well. Results suggest that SME choice did not impact the constructs measured by the SJT. Results of verbal protocol analyses conducted with a sub-group of each SME group indicate that the group members did provide different reasoning for their responses, but the differences did not affect construct measurement. Manipulation checks, however, demonstrated that the three groups chosen as SMEs did not differ in predicted ways. For this reason, the theory cannot be fully discounted. Limitations of the methodology are discussed thoroughly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Source of Expertise in Scoring Key Development as a Determinant of the Nature of the Constructs Measured

Introduction

My purpose in writing this paper is to challenge the practice of selecting subject matter experts based on convenience to help with test development and scoring of tests based on convenience. If our tests are designed to distinguish between individuals who will truly perform desirably versus individuals who will not, we need to ensure that the subject matter experts we select to provide information are actually experts in the occupation or discipline of interest. If subject matter experts (SMEs) are selected fairly arbitrarily (for example, graduate students high in cognitive ability but without appropriate expertise), then do the measures they help us develop truly tap the constructs we are targeting? In other words, if a graduate student without expertise in an area develops a scoring key for a selection instrument designed for a specific occupation, then won't the individuals who score highly on that instrument be more similar to the graduate student than to the experienced worker in that occupation?

One area in which SMEs play a critical role is in the development of scoring rubrics for exercises in which several alternative courses of action are available. The concern about the level of expertise necessary to play this important role has surfaced most recently in the literature on situational judgment tests (SJTs).

I suggest that different SMEs utilize different criteria for scoring SJTs. I will present research suggesting that SMEs drawn from the same groups (either broadly or specifically defined; for example, all employees within one organization or the members of a specific team within an organization) should reference similar criteria because, over

time, these groups tend to become fairly homogenous with respect to past experiences and traits, both of which contribute to decision making in a test-taking situation. Similarly, they should reference different criteria than SMEs in different groups. If different groups of SMEs are referencing different criteria when they develop scoring keys for SJTs, then we need to be precise about whom we select as our group of SMEs to ensure that they are referencing the criteria we purport to measure.

I will begin by reviewing literature that demonstrates that SJTs are valid and practical predictors of performance and suggest that, although researchers have yet to discover what variables affect the constructs measured by an SJT, trying to do so is a worthy goal because of how widely SJTs are used in practice and because understanding the constructs measured will allow users to better tailor the SJT items to the knowledge, skills, and abilities required on the job. In order to better understand, predict, and control applicant responses, we really need to know what the tests measure. For example, knowledge of what a test measures can help us in understanding why subgroup differences occur and can also help us to make informed decisions about how to put measures together in a single selection system without creating redundancy.

I will describe the process by which an SJT is typically developed and review how different aspects of the development process may influence ultimate construct measurement. I will focus specifically on the development of a scoring key. I argue that in order to fully understand the constructs measured by an SJT, we have to understand the nature of the scoring for that SJT. I will present evidence to suggest that individuals drawn from different groups reference different criteria when responding to test items (which, in the case of SJTs, is how we most commonly develop scoring keys) and that

individuals within groups are more likely to reference similar criteria than individuals between groups.

Review of Literature on SJT Validity and Subgroup Differences

A situational judgment test (SJT) is a low-fidelity simulation in which a test-taker is presented with a work-related situation accompanied by a list of responses to the situation and is asked to select either the best and worst options or those he or she would be most and least likely to perform (Motowidlo, Dunnette, & Carter, 1990). SJTs are widely used in employee selection contexts for three main reasons. The first reason is that they have been shown to be highly predictively valid; the second reason is that SJTs exhibit smaller subgroup differences than other common selection procedures, and the third reason is that applicants tend to react favorably to the measures and to perceive them as face valid.

In 2001, McDaniel. Morgeson, Finnegan, Campion, and Braverman conducted a meta-analytic review of the criterion-related validity of SJTs and estimated the population validity of such measures to be .34. This indicates that approximately 12 percent of the variance in subsequent job performance can be explained by a test-taker's score on the SJT. They also suggested that this value may be downwardly biased because the estimates of SJT validity were not corrected for range restriction. They estimated that the criterion-related validity of most SJTs should fall within the range of .21 to .41 and that the specific validity coefficient for an individual test may be moderated by a number of factors such as the content included in the test or whether or not the test was developed based on a job analysis. Perhaps another moderator of SJT validity, which will be

explored further in this paper, might be the suitability of the rubric developed for scoring the test.

SJTs have typically been shown to have lower levels of adverse impact than cognitive ability or verbal ability measures. Pulakos and Schmitt (1996) compared the adverse impact levels of a verbal ability measure with those of an SJT. Whereas the verbal ability test led to a standardized mean difference (d) between Whites and African Americans of 1.03, the SJT had a d of .41, a reduction in d of .62. The findings of Motowidlo and Tippins (1993) of a White-African American d of .32 and of Motowidlo et al. (1990) of White-African American d's of .14 and .29 in two different samples support the findings of Pulakos and Schmitt and indicate that the adverse impact levels of SJTs are typically lower than those of traditional cognitive ability tests (White – African American d = 1.00; Sackett, Schmitt, Ellingson, & Kabin, 2001). Whetzel, McDaniel, and Nguyen (2008) conducted a meta-analysis of race differences on SJT performance. They found a White – African American d of .38, a White – Hispanic d of .24, and a White – Asian d of .29. They also found evidence that mean race differences in SJT scores are largely due to differences in cognitive ability. The more an SJT overlapped with a measure of cognitive ability, the higher the racial differences found for that SJT. Although the reasons for why SJTs demonstrate less adverse impact than more traditional tests are not yet defined, Sackett et al. (2001) described evidence that supplementing cognitive predictors with noncognitive predictors relevant to job performance (yet not correlated with the cognitive predictors) can lead to reductions in adverse impact. It is probable that SJTs are measuring a variety of constructs in addition to cognitive ability that are related to performance.

Test takers and test developers tend to like SJTs because they at least appear to be more highly related to the jobs of interest than many more typical selection assessments (Bauer & Truxillo, 2006). Because SJTs include descriptions of situations that individuals might encounter on the job, they appear more face valid than more typical or abstract assessments, such as cognitive ability measures, which may seem less related to experiences on the job to a lay person.

Despite the validity and practicality of SJTs, researchers have been unable to determine exactly what SJTs measure. A number of researchers have attempted to understand the construct validity of SJTs by correlating SJT scores with other established measures. The results of such studies suggest that different SJTs are not consistently measuring the same constructs, but that individual SJTs can be designed to measure different constructs.

Perhaps the largest and most conclusive body of work in this area is the literature that relates SJT scores to cognitive ability. McDaniel et al. (2001) conducted a meta-analysis and found a mean corrected correlation of .46 with general cognitive ability measures, although there was significant variability around the mean (with a credibility interval of .17 to .75). Their results indicated that SJTs vary considerably in regard to their relationship with cognitive ability.

Because SJTs tend to include job-specific situations, it might be expected that job knowledge would be highly correlated with SJT scores. Similar to cognitive ability, however, meta-analytic results do not support consistency across SJTs. McDaniel & Nguyen (2001) reported an average correlation of .07, excluding a large study in which the correlation was negative (Clevenger & Haaland, 2000). Once again, the credibility

interval was large (-.14 to .29), indicating that SJTs vary considerably in the strength of their relationship with job knowledge.

Perhaps not surprisingly, SJTs have also been found to vary considerably in their relationship to personality measures. In their meta-analyses, McDaniel and Nguyen (2001) and McDaniel et al. (2007) examined the relationship between SJT scores and the Big Five personality traits. For almost all of the traits (excluding Openness, for which there were fewer estimates of effect size), the credibility intervals were large, once again indicating that SJTs vary greatly in their relationships to various constructs (Chan & Schmitt, 2006). The widely varying correlations with measures of relatively well-established constructs suggests that SJTs may be developed, either purposefully or inadvertently, to measure a variety of constructs.

If SJTs can be developed to measure a variety of constructs, the question remains as to what aspects of SJT development affect construct measurement. In the following section, I explore the little empirical evidence and speculation that exists about how SJT development may affect construct measurement.

Development of SJTs

There are four basic steps in the development of an SJT that may affect construct measurement. The first step is the generation of a set of item stems, or situations, that serve as the base of each question. Next, the response options for each item stem are developed. Test developers must make a decision about what type of instructions to provide to test takers and then, finally, a scoring key is developed. Although my focus in this paper is on how scoring key development affects construct measurement, I will first discuss how the three other parts of development may also do so.

The development of an SJT begins with the development of item stems, the descriptions of situations which are the most basic part of each question. The situations used in the final form of an SJT are typically derived from critical incidents collected from subject matter experts. The critical incidents are typically culled and refined by the test developer, who seeks to avoid too much redundancy and to cover as much of the desired topic area as possible. After choosing a final set of critical incidents, the test developer transforms the incidents into situation-based questions of similar length that use consistent language.

McDaniel and Nguyen (2001) described the following four characteristics along which most item stems could be distinguished: fidelity, length, complexity, and comprehensibility. I will use these categorizations to examine how each characteristic may affect the construct measured in the final SJT.

The fidelity of the item stem refers to how similar the presentation of the situation in the item stem is to how the situation would occur in real life. Chan and Schmitt (1997) compared performance on two versions of the same SJT with varying fidelities. The item stems of the first SJT were presented as videotaped vignettes whereas the item stems of the second SJT were presented as written descriptions of the vignettes. The video-taped vignettes would be considered to be of higher fidelity than the written descriptions. Although Chan and Schmitt were able to demonstrate that the content (and thus the intended constructs measured) in the two versions of the SJT was identical, they found that performance on the written version was correlated with scores on a reading comprehension test (an additional and unintended construct) whereas performance on the

videotaped version was not. Chan and Schmitt's findings suggest that the presentation style of item stems may affect which constructs an SJT measures.

McDaniel and Nguyen's (2001) second, third, and fourth categories (length, complexity, and comprehensibility) are probably strongly related, so I will address them together. Item stems can be written very simply or can involve complex situations (necessitating the need, for example, to deal with multiple actors with conflicting interests). In many cases, more complex item stems will be longer because it takes more words to describe more intricate situations. The complexity and length of item stems may also be related to the level of comprehensibility, the ease with which the meaning and implications of the situation are discerned. Sacco et al. (2000) found evidence in two studies (Sacco, Scheu, Ryan, & Schmitt, 2000; Sacco, Schmidt, & Rogg, 2000) that performance on item stems with greater detail and complexity was related to reading-level indices. In other words, the implication is that SJTs consisting of item stems with a higher level of detail may be measuring more cognitively-loaded constructs, such as reading level, regardless of whether or not such constructs are relevant to performance in the situations described. McDaniel et al. (2001) reported seemingly contradictory findings in their meta-analytic study. They found that SJTs with less detailed questions were more highly related to general cognitive ability ($r = .56$) than those with more detailed questions ($r = .47$). Although the empirical findings are contradictory, all three studies indicate that the complexity and comprehensibility (and possibly the length) of item stems may affect what constructs are measured.

Finally, it seems necessary to point out that item stem content may affect the constructs measured by an SJT. Although there is little empirical evidence to suggest

that intentionally writing item stems to measure specific constructs is the primary predictor of the constructs ultimately measured, the underlying theory or set of competencies used to develop a set of item stems is often presumed to affect construct validity as intended.

Once the item stems for an SJT have been developed, response options are developed for each stem. Response options are typically developed by a pool of subject matter experts who provide examples of typical or desired responses to the situation described in each item stem. Response options may also be written by the test developer. In either method of response generation, SMEs typically review the developed responses to weed out unrealistic and redundant responses.

McDaniel and Nguyen's (2001) four characteristics of item stems (fidelity, length, complexity, and comprehensibility) that may affect construct measurement are probably applicable to item responses, as well. Most SJTs, even those that present item stems by video, provide written item stems. It is plausible, however, to present videotaped item responses along with item stems, which suggests that the evidence of the effects of fidelity on construct measurement are applicable to a discussion of item responses as well as item stems. It also seems plausible to extend the findings and predictions that item stem complexity, length, and comprehensibility may affect construct measurement (possibly by affecting the cognitive resources required to respond). Sacco et al. (2000), however did not find evidence to support the idea that the stem-level reading effects would extend to the level of the response option.

In several studies, researchers have attempted to develop response options that reflect varying levels of a personality trait with the goal of developing an SJT that

measures that trait. For example, Motowidlo, Hooper, and Jackson (2006) developed an SJT with five item stems designed to tap extraversion, five item stems to tap agreeableness, and five item stems to tap conscientiousness. For each item stem, they then developed response options designed along a continuum of the trait of interest. So, for example, for an item designed to measure agreeableness, the item responses described behaviors that ranged from disagreeable to agreeable behaviors. The responses were also coded by graduate research assistants and the coding was used to determine if a response indicated a high or low level of the trait. Test-takers' responses were correlated with their scores on personality measures of agreeableness, extraversion, and conscientiousness. For agreeableness and extraversion, the SJT was correlated with these two personality scores, but correlations with conscientiousness were low and nonsignificant. Several other studies have found similar, conflicting results (Beauregard, 2000; Motowidlo, Diesch, & Jackson, 2003; Ployhart & Ryan, 2000; Porr & Ployhart, 2004; Trippe & Foti, 2003), but evidence suggests that it may be possible to intentionally design item response options to tap specific constructs.

A number of researchers have suggested that the transparency of response options is yet another possible aspect that may affect construct measurement, although this idea has yet to be tested empirically (Hooper, Cullen, & Sackett, 2006; Weekley, Ployhart, & Holtz, 2006). If certain response options are more socially desirable and test takers endorse those items, then the SJT would have the potential to become a test of social desirability rather than a valid predictive measure. Potential solutions to this problem include designing response options to be of equal social desirability or matching each

socially desirable response option with one that is equal on social desirability but not predictive of performance.

In addition to developing the item stems and response options, test developers have to decide what type of response instructions to provide to test takers. The two most common instruction types are behavioral tendency instructions and knowledge instructions (McDaniel, Hartman, Whetzel, & Grubb, 2007). Behavioral tendency instructions ask test takers to select the responses that best describe how they would behave in the given situation. In contrast, knowledge instructions ask test takers to respond with what they think the most (and least) effective response would be. There is some evidence to suggest that the type of instructions selected by test developers may affect what constructs an SJT measures. In their meta-analysis, McDaniel et al. (2007) found evidence that behavioral tendency instructions are more highly correlated with personality measures whereas knowledge instructions are more highly correlated with measures of cognitive ability, regardless of SJT content. Ployhart and Ehrhart (2003) made a similar distinction between two types of instructions ("would do" versus "should do") and found that scores on an SJT identical in content, but with different response instructions, were more similar across conditions using the same instruction type. In other words, instructions asking what an individual "would do" measured something different than instructions asking what an individual "should do" in a given situation, despite the item stems and response options being identical. McDaniel and Nguyen (2001) suggested (although they did not test empirically) that SJTs with knowledge instructions may be less resistant to faking (or measuring social desirability) than behavioral tendency instructions. They suggested that SJTs with knowledge instructions

measure the same type of knowledge from all participants whereas SJTs with behavioral tendency instructions measure behavioral tendency from non-fakers and knowledge about effectiveness from fakers. The described research provides evidence and theory that suggests that even the seemingly inconsequential choice about how to instruct applicants to complete an SJT can have implications for what constructs are measured by the test.

Although there has been little empirical work investigating how the development of item stems, response options, and response instructions may affect construct measurement, the work that does exist suggests that even slight or unintentional alterations in any of these three parts of an SJT may affect which constructs are measured. For example, in the example of the fidelity of item stems, Chan and Schmitt (1997) showed that identical situational content presented in written versus video format adds the measurement of an additional construct. For this reason, developers of SJTs must be cognizant of all of the choices they make about how to develop each aspect of an SJT and how that may influence construct measurement.

The fourth part of SJT development which I believe may influence construct measurement is the development of a scoring rubric. Although I believe that this process, which is often heavily influenced by SMEs, may impact construct measurement, there is little empirical work on this topic. The development of a scoring key and how it may affect construct measurement is also probably the aspect of SJT development that has the most relevance for the development of other measures. This aspect of SJT development and how it may affect construct measurement will be my focus for the remainder of this paper.

Scoring Key Development

Empirical versus Rational Keying

The two most common methods for developing scoring keys for an SJT are empirical and rational keying. In the empirical method, response options are selected and weighted based on their ability to differentiate membership in higher and lower performing criterion groups. In the rational method, response options are weighted based on the opinion of SMEs that the options tell us something important about the constructs targeted in the instrument.

In order to create a scoring key using the empirical method, one must have a criterion measure of interest on which there is significant variability in order to create at least two dichotomous groups (high and low performers), although empirical keys can be developed against continuous criterion measures as well. The main benefit of this method is, of course, that it maximizes the prediction of a specific, external criterion (England, 1971; Mitchell & Klimoski, 1982). In this case the SJT should be measuring the same construct(s) as the chosen criterion, although, if there is ambiguity about the nature of the constructs underlying the criterion, this will translate to the SJT as well. There are the following three main methodological problems with empirical keying: reliance on an external criterion, a potential lack of generalizability, and the potential decay of validity over time.

The effectiveness of the empirical key depends entirely on the adequacy of the criterion measure used to represent the construct(s) of interest (Thayer, 1977). By using an external criterion to create a scoring key, the prediction of that external criterion is maximized, but that creates uncertainty about why the measure is effective. By using an external criterion measure to score an SJT, the SJT takes on any problems or

inadequacies in the criterion measure. While the empirical keying method ensures that there is a relationship with some outcome variable of interest, it also has the potential to add another level of error in the measurement of the underlying constructs. Not only are we left uncertain about why the SJT accurately predicts scores on the criterion measure, but also we have to assume that the criterion measure adequately tapped the underlying constructs of interest in the first place.

In addition, the generalizability of the scoring key is dependent on both the reference group and the sample who takes the SJT and the criterion measure. First of all, the sample who takes the measures must adequately represent the group of people to which the scores are supposed to generalize (the reference group) in order for validity estimates to generalize. Second, scoring weights developed from a specific sample will, to a certain extent, capitalize on sample-specific factors, which will result in high validity estimates (Hogan, 1994). Thus, it is essential to conduct cross-validation studies on multiple samples representing the reference group in order to avoid the effects of idiosyncratic factors in the data.

A third methodological issue with empirical keying is the tendency for the validity of the measurement to diminish over time (Reilly & Chao, 1982; Thayer, 1977). Three possible reasons for this documented decay in validity of empirically-developed scoring keys over time are changes in the nature of the external criterion over time, shifts in the nature of the reference group, and lack of security of the scoring rubric (or an increased familiarity with desirable responses).

When developing a scoring key using the rational method, subject matter experts (SMEs) are typically asked to select the best and worst responses and/or to rate the

effectiveness of each response option. The scoring key is then developed by weighting response options based on the best and worst ratings of the SMEs and based on item effectiveness ("best" items should also be rated as relatively effective and "worst" items should be rated as relatively ineffective; Motowidlo et al., 1990). The main benefit to this method is that it is theoretically based. It is assumed that subject matter experts are using their expertise to critically analyze and judge how the response options should be weighted using their "theory" of what constitutes effective performance. Unfortunately, however, in practice, we generally do not know the rationale that the "experts" use to determine their responses. In other words, their "theories" are implicit, rather than explicit. Asking SMEs to respond as though they were test takers assumes that they are capable of accurately introspecting and reporting how they would behave or how they think others should behave. SMEs are generally selected because of their expertise in the field (or their demonstration in the past that they can behave desirably, which has led to their success in their field or organization), but they may or may not be aware of the reasons why they have been successful or what behaviors have best served them. When we use SMEs to rationally score SJTs, we assume they are using their underlying "theory" of successful job performance to develop a scoring key. The "theories" are never explicitly stated and may vary considerably across SMEs or groups of SMEs.

From where do the Criteria SMEs Use Come?

In this section, I argue that SMEs, because of their background and expertise, may differ in their theories of performance. Moreover, I argue that there are subgroups of SMEs whose implicit theories are similar to other members' of the subgroup and dissimilar to the theories of members of other subgroups. First, however, I consider what

processes might create similar views of performance or implicit performance constructs that SMEs may use in making their judgments about scoring items.

Theories about the cognitive processes of test-takers (or SMEs) suggest that individuals responding to specific items reference criteria that come from their past experiences as well as individual differences (Motowidlo, Hooper, & Jackson, 2006; Ployhart, 2006). It is probably not accurate to assume that all individuals whom we might consider experts in a field consider successful performance or successful resolution of a situation in a given SJT item in the same way. If they are not viewing successful performance the way that we think they are, then our test is not measuring the constructs that we think (or go so far as to claim) it measures.

Ployhart (2006) described a model of determinants of predictor response processes. The basic concept was that there are a variety of ways in which latent constructs affect test takers' responses to an individual test item. I believe that we can analyze a SME's judgment about the best and worst response to an SJT item in the same way. A test taker (or SME) engages in the following four related and sequential phases of cognitive processing in determining his/her response: 1) comprehension of the item (in this case the situation), 2) retrieval of relevant information from long-term memory, 3) the forming of a judgment using that information, and 4) the choice of a response option based on that judgment. All four of the phases are informed by latent constructs or criteria engaged by the test-taker (or SME). When we ask SMEs to help us with scoring, we tend to assume that the constructs they reference at each step in the process of making their responses will be fairly consistent across SMEs, but we have no evidence to support this. For example, in rating the same response options for an item stem, one subject

matter expert may retrieve information and make a judgment based heavily on cognitive ability whereas another SME may do so using agreeableness as the main construct of interest. In reality, of course, each SME is probably referencing multiple constructs at each step of the process of responding to a single item, only some of which may really be relevant to the task. It is by averaging responses across multiple SMEs that we may be able to average across idiosyncrasies to find the responses that are overall selected based on the true constructs of interest. Of course, this assumes that groups of SMEs do not share idiosyncrasies, a point I will challenge later. This brings me to the other theory I am utilizing in my explanation.

Motowidlo, Hooper, and Jackson (2006) have suggested the relevance of implicit trait policy (ITP) to test-taker endorsement of SJT response options. Similarly to Ployhart's theory, I am applying this theory to subject matter experts determining response effectiveness. ITP theory suggests that there are stable differences between individuals in their beliefs about the importance of various personality traits for determining behavioral effectiveness. In other words, to frame this in terms of Ployhart's predictor response processes model, individuals' beliefs that certain traits are important to use in certain situations will affect the criteria they reference after reading the item and how they apply those criteria to their judgment of which response option to choose. For example, an individual whose ITPs weigh agreeableness highly will judge the effectiveness of the response options based on how well they represent agreeable behavior whereas an individual whose ITPs weigh conscientiousness highly will use that as his/her criterion of interest in determining response option effectiveness. It is necessary to emphasize again that, when we ask SMEs to create a scoring rubric, we are

assuming that the SMEs we choose have the most relevant ITPs for the situations in our measure. These ITPs come from their pre-existing individual differences and are shaped by personal experiences.

Theoretically, it appears likely that different individuals will rate the effectiveness of different response options differently based on their own criteria of interest. This leaves open the question of from where these criteria come or how individuals determine which criteria to reference. The theories about the cognitive processes test-takers use in making decisions about how to respond to test items suggest that the criteria that SMEs use in their judgments come from individual differences and from past experiences. In the past, it has been assumed that all individuals designated as SMEs use the same criterion of interest when evaluating response effectiveness. I am not arguing that each SME will employ entirely different criteria in judging response options, but I am arguing that similarities or differences in background and experience will produce subgroups of SMEs who will provide different judgments about scoring keys. In the next section, I consider what about individuals chosen to serve as SMEs makes them similar or different in terms of what might influence their judgments about scoring the response options.

Influences that Produce Judgment Similarities/Differences

There are a number of theories or hypotheses in a variety of disciplines within psychology and sociology that suggest that individuals within groups will have more similar traits and experiences than individuals between groups. Some of these theories include attraction-selection-attrition, socialization or fit, organizational demography, evolutionary theories about tradeoffs between personality traits, and niche-picking. These theories suggest that individuals with similar individual differences and

experiences tend to aggregate together or to be drawn to similar work groups, organizations, and occupations and that membership in these groups tends to lead group members to become more similar in a variety of ways including their ideas and judgments over time. I will briefly describe each of these ideas and provide exemplary research studies that support these propositions.

Schneider's theory of Attraction-Selection-Attrition (ASA; 1987) suggests that organizations are likely to attract and select individuals who have similar personality traits and work values and that, over time, individuals who are not a good fit with the organization will leave (Cable & Judge, 1996, 1997; Judge & Cable, 1997). In the recruitment process, individuals who are similar to existing employees are both more likely to be recruited by the organization and to apply for jobs. Individuals who observe a misalignment between their characteristics and those of organization members may avoid applying to the organization in the first place, may remove themselves from the application process, or may turn down a job offer. If an applicant whose characteristics are not aligned with those of existing employees is offered a job and accepts it, he or she is more likely to leave the organization. Over time, this process leads to the homogenization of or a restriction in the range within an organization in terms of personality traits and personal values. This theory has been extended further through empirical work to apply to the work group level and also to apply to past experiences in addition to traits and values. If ASA theory is correct, we should see a greater similarity in the judgments or approaches to problem situations among a set of experienced SMEs than among those who are relatively recently confronted by the organization.

To support the attraction portion of ASA theory, Judge and Cable (1997) found evidence that the types of organizations to which job seekers are attracted are related to their personality traits. Specifically, they found evidence that job seekers who score high on neuroticism are less attracted to innovative and decisive organizational cultures, that job seekers who score high on extraversion are attracted to aggressive and team-oriented organizational cultures and less attracted to supportive cultures, that job seekers who score high on openness to experience are more attracted to innovative organizational cultures and less attracted to detail-and team-oriented cultures, that job seekers high on agreeableness are attracted to supportive and team-oriented cultures and less attracted to aggressive, outcome-oriented and decisive cultures, and finally, that job seekers high on conscientiousness are attracted to detail-oriented, outcome-oriented, and rewards-oriented cultures and less attracted to innovative cultures.

To support the selection portion of ASA theory, Cable and Judge (1997) examined the hiring decisions of interviewers. They found that interviewers who perceived an applicant's values to be highly congruent with those of the company predicted that those employees would fit well with the organization. They were also highly likely to recommend that those employees be hired.

To support the attrition portion of ASA theory, Jackson et al. (1991) found that heterogeneity of top-management teams predicted turnover within the groups. The attributes on which they measured heterogeneity included age and experience outside of the industry (both of which were significant predictors of turnover), as well as tenure, education level, college alma mater, possession of a business management degree, and military experience (none of which were significant predictors).

According to this theory, there is reason to believe that SMEs drawn from the same groups will possess similar traits and experiences because, over time, new members who are attracted to the group tend to have a lot in common with current members, members who stay with the group are those who are most similar to begin with, and because members of groups become more similar to one another over time.

Socialization research leads to similar premises. There is evidence to suggest that, during the process in which new employees are socialized to an organization, their values and those of the organization become even more closely aligned. Cable and Parsons (2001) proposed that when newcomers to an organization learn during the socialization process that their values differ from those of the organization, they experience dissonance because their values do not match those normative for success. They suggested that two alternative ways of dealing with this dissonance are for individuals to alter their values or to leave the organization. This process would ensure that, over time, those employees that stay with the organization will have values aligned with those of the organization. Cable and Parsons found that, even when they controlled for initial (or pre-employment) congruence of personal and organizational values, new employees' values tended to become aligned with their organization when employers' used specific socialization tactics which led to positive social interaction and support with existing employees.

Cable and Judge (1996) conducted a study in which they examined job seekers' perceptions during a recruiting cycle at a university. At Time 1, in the spring, job seekers rated the attractiveness of a job's attributes, their perceived fit with the company and job, and their perceptions of the company's values. At Time 2, the job seekers provided

measures of their individual differences, mainly in their values, and the importance of fit in their job search and choice. At Time 3, after working for approximately 5 months, the participants completed a survey about their perceived fit with the organization and job and their job attitudes. Cable and Judge found evidence that perceived values congruence between applicants and organizations was predictive of their fit perceptions, that fit perceptions predicted job choice intentions, that perceived values congruence with the organization at which the participants ultimately accepted a job positively affected their perceptions of fit as employees, and that fit perceptions predicted positive outcomes, such as reduced turnover.

Another related theory, organizational demography, suggests that a measure of the aggregate of the demographic information about organization members influences behavior independently of individual-level attributes. The idea behind this theory is that demographic variables can serve as a proxy for measures that are less objective, such as attitudes, because demographic variables are directly measurable. Therefore, there is an assumption that individuals from the same demographic groups have similar attitudes, behave similarly, and make similar judgments. This may be a big assumption, but there is some evidence to suggest that the methodology of organizational demography can be useful, which indicates that it may be accurate at least some of the time (Lawrence, 1997; Pfeffer, 1983). According to this theory, there is reason to expect that SMEs drawn from the same environments will, in general, possess similar demographic descriptors (for example, organizational tenure, marital status, or gender) and that, those who do possess similar descriptors will produce similar judgments. This is because those demographic variables serve as indicators or predictors of similar attitudes, which influence how they

make similar decisions. This theory lacks a process explanation for how demographic variables influence attitudes which influence outcomes (in this case, ratings of response options for SJT item stems); it simply predicts (with some success) that they will.

Wagner, Pfeffer and O'Reilly (1984) found empirical support for the theory of organizational demography. They examined the demography of top-management teams, predicting that members most similar in terms of date of entry into the organization would be less likely to leave the organization than those members who differed along that variable. Overall, they found support for this hypothesis; the larger the distribution of dates-of-entry within a group, the higher the proportion of the group that left. The theory of organizational demography predicts that groups consisting of members who are more demographically similar will be more successful in the long term. One mechanism that predicts this success is that homogeneous groups are less likely to have high rates of turnover.

There is some theory in evolutionary psychology that suggests that the expression of certain personality traits has benefits for certain groups, so individuals with those traits tend to be attracted to those groups where they can be successful (Nettle, 2006). One example that has been discussed is that, among university students, academic success is strongly positively correlated with neuroticism among those students resilient enough to cope with its negative effects (McKenzie, 1989; McKenzie, Taghavi-Knosary, & Tindell, 2000). This theory developed from a theory of tradeoffs, which is an explanation of heritable variation between humans. In other words, it is an attempt to explain why humans have different traits if a specific profile of traits would be most adaptive for the human experience. The theory suggests that there is not a specific profile of traits that

would be most adaptive to the human experience, but rather that individuals who are successful in life (or are considered "fit" for survival) are attracted to environments where the traits they possess are adaptive.

In most cases, SMEs are chosen to provide expertise on devising a scoring key because they have been successful performers in their environment (i.e., they are high ranking employees within an organization or high performing college students). Thus, according to evolutionary theory, it is likely that SMEs from the same environment will possess similar traits because they will have been attracted to their environment specifically because of the adaptive utility of their particular traits within that environment.

In their study, McKenzie, Taghavi-Khonsary, and Tindell (2000) examined a group of university students who, over a period of three years, were enrolled in a higher education course of study in London. The program offered an opportunity to leave with a Certificate of Higher Education after one year of successful study or to leave with a Diploma of Higher Education after two years of successful study. Successful attainment of the Diploma allowed the students to continue with one additional year of study to earn one final award. McKenzie et al. found that, for students with high scores on a measure of coping, neuroticism was highly correlated with academic achievement; this correlation increased for each higher level of educational attainment in the program (i.e. from certificate to diploma to degree). Students who possessed high levels of neuroticism, but also scored highly on a measure of coping ability, were able to channel their capabilities towards higher educational attainment. This finding supports the idea that individuals who are successful in specific domains may possess similar profiles of personality traits.

SMEs chosen for their expertise in a specific area (i.e. an area in which they have been successful) thus, seem likely to have a similar profile of traits.

Caspi, Roberts, and Shiner (2005) described niche-building processes, whereby individuals create, seek out, and/or end up in environments that are highly correlated with their personality traits. Once individuals are in those environments, there may be processes within the environments which then promote the persistence of trait-related behaviors and inhibit or discourage opportunities for changing those behaviors. These environments can range from occupations to workplaces to social situations and more. According to the niche-building theory, SMEs with similar traits will create, seek out, and end up in environments that foster those traits. Thus, individual environments (for example, organizations) will tend, over time, to have and to keep employees with similar personality traits. For example, an individual who is highly extraverted is likely to seek out a workplace where he or she feels comfortable expressing that trait and also where that trait is encouraged. The individuals in that workplace, where extraversion is encouraged, are, in turn, more likely to recruit and hire an extraverted person for positions in their work group. Thus, certain workplaces will be more likely to contain extraverts than introverts and vice versa.

Magnus, Diener, Fujita, and Pavot (1993) conducted a study in which they sought to determine whether personality traits have the capacity to influence objectively positive and negative life events (for example, getting married was coded as a positive event whereas the death of a close family member was coded as a negative event). In 1986, they collected measures of the Big 5 personality traits from a group of undergraduates. Four years later, the former undergraduates responded to a mailed checklist about

objective life events they had experienced. Those events were coded as positive or negative by an independent sample. Magnus et al. found, as they predicted, that individuals high in extraversion reported more positive events, whereas individuals high in neuroticism reported more negative events. Because the events checklist consisted of objective and fairly salient events, it is unlikely that the difference in reported events was due to a reporting bias (i.e. highly neurotic participants were probably not simply remembering more negative events than positive events). The findings of this study support the niche-picking process described by Caspi, Roberts, and Shiner (2005) whereby the personality traits that individuals have influence the situations they find themselves in. Because people with similar personality traits are likely to find themselves in similar situations, it seems likely that because SMEs from the same group have been attracted to a single situation, they probably possess similar personality traits.

These theories drawn from disparate fields within psychology and sociology all suggest that members of a group are likely to have more similar individual differences and past experiences than non-members. The more specific the group, the more likely an individual is to be similar to the other members. For example, an employee of an organization is likely to be more similar to the members of his or her work group than to the members of another work group within the organization and is likely to be more similar to members of the organization than to members of a different organization. These similarities in individual differences in traits and past experiences are likely to affect the judgments that an SME makes when evaluating situations and response options to create a scoring rubric for an SJT. Based on these theoretical notions, I develop the hypotheses presented in the next section of the paper.

Hypotheses

To investigate whether group membership does affect judgments made by SMEs in creating scoring rubrics for an SJT, I plan to use an SJT designed to measure college student performance and three groups of SMEs: graduate students, undergraduate resident advisors (mentors), and undergraduate students drawn from the Psychology subject pool (largely first and second year students). I have three hypotheses about the nature of the differences between these three groups.

H1: There will be significant differences between the three groups of SMEs in the effectiveness ratings of the response options in the SJT, such that graduate students will rate options related to academics highest, mentors will rate options related to leadership and service highest, and undergraduates will rate options related to social life highest.

Several theories and bodies of research (i.e., ASA, socialization, etc.) suggest that individuals with similar backgrounds or experiences, traits, and values will process information and make decisions in similar ways. My premise is that graduate students highly value academic pursuits, partially because they have been successful in their own academic pursuits and because they have been rewarded for their efforts in this domain. I also posit that their family background was one in which academic pursuits were valued and rewarded and that their parents themselves may have served as high-achieving academic role models. In addition, evidence (McKenzie, Taghavi-Khonsary, & Tindell, 2000) suggests that high achieving graduate students may be highly neurotic in addition to being highly conscientious.

I posit that mentors will also value academic performance, but will value a broader array of aspects of student life in making their judgments. They have been drawn to serve the university, which indicates that they value leadership and service. Because of their training and socialization as RAs, they will value interpersonal competence. I also predict that they will possess personality attributes that are relevant to both academic success and sociability, including high levels of conscientiousness, extraversion, and agreeableness.

Finally, I suggest that undergraduate students in the Psychology subject pool tend to be students relatively new to the university. As they adjust to life away from home, largely surrounded by peers for the first time, their judgments of how to succeed in college will be based mainly on interpersonal and social competence. I also suggest that, similarly to the mentors, the undergraduate students will, as a group, score highly on measures of extraversion and agreeableness. Please see the attached table (Appendix A) for a visual presentation of the predicted differences among the three SME groups.

H2: There will be significant differences in the correlates of scores based on scoring keys developed from different SME groups.

Given my arguments above about the nature of the three different SME groups, I predict that (H2a) scores based on the scoring key developed by graduate students will correlate more highly with a measure of students' academic success and with a measure of neuroticism than will the scoring keys developed by the other two SME groups; that (H2b) scores based on the scoring key developed by mentors will correlate more highly with a measure of leadership and service-related performance than will the scoring keys developed by the other two SME groups; and that (H2c) scores based on the scoring key

developed by undergraduates will correlate more highly with measures of students' social competence and sociability than will the scoring keys developed by the other two SME groups.

H3: There will be significant differences between groups of SMEs in references made to specific dimensions based on a verbal protocol analysis of remarks made while they are making judgments about option favorability, such that graduate students will tend to refer to the importance of academic success, mentors will refer to the importance of leadership and services, and undergraduates will refer to the importance of social success.

I plan to ask a small number of SMEs in each of these three groups to talk through their thought process out loud while making their scoring judgments. My specific predictions are that (H3a) graduate students will make the highest number of academically-related spoken references while making judgments in the SJT task, (H3b) mentors will reference a broad array of criteria (academic, leadership and service-oriented, and social) for making judgments during the verbal protocol task, and (H3c) undergraduates will make the highest number of socially-related spoken references during the verbal protocol task.

Method

Participants and Procedures

Phase 1: Coding of SJT Response Options

In this phase, 13 graduate students in Industrial/Organizational Psychology read each of the SJT items and response options and rated whether they believed each response option was a measure of academically-oriented behavior, interpersonally-oriented behavior, or a behavior influenced by a broad array of both academic and nonacademic domains. The participants were also asked to provide an estimate of how confident they were in their rating of the category to which each response option belongs. The purpose of this phase was to demonstrate that the SJT instrument would allow for sufficient differentiation between the groups along the predicted dimensions. Please reference Appendix B for the instructions for this task.

The ratings of the response options along with ratings of confidence in the choice can be found in Table 1.

Table 1

Coding of Response Options in Phase 1

Item	Option	Dimension	% Endorsing	Confidence (M and SD)
1	a	Leadership/Service	76.9	*M*=4.54, *SD*=.66
	c	Leadership/Service	92.3	*M*=4.46, *SD*=.52
	d	Leadership/Service	76.9	*M*=3.85, *SD*=1.07
2	c	Academic	76.9	*M*=3.08, *SD*=.86
	d	Social	84.6	*M*=4.31, *SD*=.75
3	b	Leadership/Service	76.9	*M*=4.15, *SD*=.55
	d	Academic	76.9	*M*=3.92, *SD*=.86
	f	Social	100.0	*M*=3.85, *SD*=.80
4	c	Academic	84.6	*M*=4.38, *SD*=.51
	d	Academic	76.9	*M*=3.85, *SD*=.80
7	b	Academic	92.3	*M*=4.23, *SD*=.93
	c	Academic	84.6	*M*=4.46, *SD*=.88
8	a	Leadership/Service	92.3	*M*=4.00, *SD*=.91
9	a	Social	92.3	*M*=4.23, *SD*=.83

Table 1 (cont'd)

10	c	Academic	84.6	*M*=4.23, *SD*=.73
	f	Social	84.6	*M*=4.31, *SD*=.75
11	c	Academic	92.3	*M*=4.62, *SD*=.51
	e	Academic	76.9	*M*=3.62, *SD*=.96
12	b	Academic	92.3	*M*=3.92, *SD*=.76
13	b	Leadership/Service	92.3	*M*=4.00, *SD*=1.00
	c	Leadership/Service	92.3	*M*=3.69, *SD*=1.18
	d	Social	84.6	*M*=4.46, *SD*=.52
	e	Leadership/Service	92.3	*M*=3.77, *SD*=.93
14	b	Academic	92.3	*M*=4.00, *SD*=1.00
	f	Leadership/Service	76.9	*M*=4.08, *SD*=.49
15	c	Leadership/Service	76.9	*M*=4.38, *SD*=.65
	d	Social	92.3	*M*=4.31, *SD*=.63
	e	Academic	100.0	*M*=4.46, *SD*=.52
16	a	Social	92.3	*M*=3.92, *SD*=1.19
	b	Social	84.6	*M*=4.38, *SD*=.65
	d	Social	84.6	*M*=3.15, *SD*=.99
17	d	Social	100.0	*M*=4.85, *SD*=.38
18	e	Academic	75.0	*M*=3.25, *SD*=1.14
19	b	Social	100.0	*M*=4.38, *SD*=.77
20	a	Academic	84.6	*M*=1.31, *SD*=.85
21	a	Academic	84.6	*M*=4.15, *SD*=.90
	b	Academic	76.9	*M*=3.54, *SD*=1.05
22	a	Academic	84.6	*M*=4.62, *SD*=.65
	b	Academic	76.9	*M*=3.69, *SD*=.85
	d	Social	92.3	*M*=4.54, *SD*=.78
23	a	Academic	76.9	*M*=3.77, *SD*=1.17
	b	Academic	92.3	*M*=4.54, *SD*=.66
	c	Academic	76.9	*M*=3.77, *SD*=.93
25	a	Social	76.9	*M*=4.31, *SD*=.75
	b	Leadership/Service	84.6	*M*=3.62, *SD*=.96
26	a	Academic	76.9	*M*=3.54, *SD*=.78
	c	Academic	100.0	*M*=4.92, *SD*=.28
27	b	Social	100.0	*M*=4.38, *SD*=.65
	c	Social	84.6	*M*=3.92, *SD*=1.04
	d	Social	76.9	*M*=4.00, *SD*=1.00
29	a	Leadership/Service	92.3	*M*=4.38, *SD*=.51
	d	Social	92.3	*M*=4.23, *SD*=.83
30	a	Academic	76.9	*M*=4.00, *SD*=1.22
	b	Social	76.9	*M*=4.15, *SD*=1.07
	c	Academic	92.3	*M*=4.23, *SD*=1.17
31	b	Academic	92.3	*M*=4.08, *SD*=1.19
	e	Academic	92.3	*M*=4.28, *SD*=1.12
	g	Social	84.6	*M*=4.31, *SD*=.63
33	a	Academic	84.6	*M*=4.23, *SD*=.93

Table 1 (cont'd)

	b	Social	92.3	*M*=4.00, *SD*=.91
	c	Academic	100.0	*M*=3.92, *SD*=1.19
	d	Academic	84.6	*M*=3.54, *SD*=.97
34	a	Academic	84.6	*M*=3.92, *SD*=1.04
	b	Academic	76.9	*M*=4.31, *SD*=.63
	c	Academic	84.6	*M*=3.54, *SD*=.97
	d	Academic	76.9	*M*=4.00, *SD*=1.08
	e	Social	84.6	*M*=4.08, *SD*=.76
35	c	Leadership/Service	76.9	*M*=4.31, *SD*=.75
	d	Leadership/Service	76.9	*M*=4.00, *SD*=.82

Out of 189 response options in the SJT instrument (see Appendix C), at least 70% of the Phase 1 participants agreed on a single dimension for 69 (37%) of the response options. Thirty-four (49.3%) of the 69 response options were classified as related to the academic dimension, 14 (20.3%) were classified as related to the leadership/service dimension, and 21 (30.4%) were classified as related to the social dimension. Phase 1 was considered a successful demonstration that the SJT contained sufficient content to discriminate between the three SME groups along the predicted dimensions.

Phase 2: Scoring Key Development

The participants of Phase 2 were 28 graduate students (different from those recruited for Phase 1), 28 mentors (resident advisors), and 33 undergraduate students. For each item, participants were asked to select the response option that they would be most and least likely to select if faced with that situation. They were also asked to rate the effectiveness of each response option for that item. The participants were also asked to provide information about themselves to be used as "manipulation checks" to ensure that the groups of SMEs really did differ along the variables predicted (measures of academic performance, leadership and service-related behavior, interpersonal

competence, personality, values, family support of education, and parental education). Please reference Appendices D, E, F and G for the instructions and content of these tasks.

The responses of 2,753 applicants to the university (collected in 2004 for another, ongoing research effort; Schmitt et al., 2007) were used to assess the correlates of the SJT scored based on the judgments of the different SME groups. Of the 2,696 students who reported their sex, 35.6% were male and 62.3% were female. Of the 2,524 students who reported their ethnicity, 55.2% were White, 23.2% were African American, 7.5% were Asian, 5.7% were Hispanic, and 8.3% identified as multi-racial or other.

Phase 3: Verbal Protocol Analysis

In this phase, five graduate students, five resident advisors, and five undergraduate students were trained in the verbal protocol technique. They were given the following instructions, "As you answer the following six questions, please try to say all of your thoughts out loud as you go. Please describe your thoughts, feelings and choices about what you are doing and reading. It might be your reactions, reasoning, or even something you are reminded of. Please don't censor your thoughts. Even if a thought does not seem relevant to the task, it is of interest to me. Remember, everything you say to me today will be kept confidential." After the training, the participants were given two practice SJT items (See Appendix H). They completed those two items using the think-aloud protocol and the researcher answered any questions they had and encouraged them to feel comfortable. After each participant felt comfortable with the verbal protocol technique, he or she was digitally recorded while completing six of the SJT items using this technique.

The six SJT items were selected for the verbal protocol based on the coding by the Phase 1 participants. The items (2, 3, 14, 15, 22, and 29) were chosen because the Phase 1 participants had categorized the response options for each of the items as belonging to multiple dimensions. For example, Phase 1 participants identified 15c as leadership/service, 15d as social, and 15e as academic. Therefore, the item was expected to elicit different patterns of response from each of the three SME groups.

After completing those six items, the participants completed the final 30 items using the same technique as the other participants participating in the scoring key development. The participants were also asked to provide information about themselves to be used as "manipulation checks" to ensure that the groups of SMEs really did differ along the variables predicted (measures of academic performance, leadership and service-related behavior, interpersonal competence, and personality). Please reference Appendices D, E, F and G for the instructions and content of these tasks. The researcher created a typed transcript of each recording.

Measures

SJT

The SJT used in this study was developed to reflect 12 dimensions of college performance. The development process is described fully by Oswald et al. (2004). To summarize, the item stems were taken from existing measures and adapted by the researchers to reflect the 12 dimensions of college student performance. Additional item stems and response options were developed by undergraduates. The final measure consists of 36 items (see Appendix C). The items reflect a range of academic, interpersonal, and intrapersonal situations.

Indicators to be Linked with Scores Based on SMEs' Scoring Keys

GPA. Yearly and cumulative four-year grade point average information were collected from the registrar's office at each of the universities for the years 2004-2008. In addition, the universities provided the four-year cumulative GPA for each of the students. Because the different schools from which data were collected varied in selectivity, college GPA was corrected. First, GPA for each university was standardized. The standardized grades were then regressed on the scores for college admissions tests along with a set of dummy variables representing each college and university. The coefficients for the dummy variables indicated the differences in grades that would be expected for students with comparable admissions scores at the different universities. Finally, GPA was adjusted for each participant by their school's regression coefficient so that students at universities with higher average admissions scores received a relatively higher adjusted college GPA, and students at universities with lower average scores received a relatively lower adjusted college GPA. This correction was made for all of the GPA variables.

BARS and BOS. Seven items from the behaviorally-anchored rating scale and seven sub-scales from the behavioral observation scale were used as indicators of the dimensions which the SME groups were predicted to value differently. Those items and sub-scales were designed to measure knowledge and mastery of general principles, continuous learning, leadership, interpersonal skills, social responsibility, adaptability, and ethics and integrity (see Appendices D and E). The BARS measures were collected from students who participated in follow-up data collections at the end of their first,

second, third, and eighth semesters. The BOS measures were collected from students who participated in the follow-up data collection at the end of their third semester.

Personality measures. The subscales of the 50-item International Personality Item Pool (Goldberg, 1999) designed to measure conscientiousness, neuroticism, extraversion, and agreeableness were administered to the applicant pool (see Appendix F). The personality measures were collected at the first data collection in early fall of 2004 when participants had just started at their universities.

Measures of Differentiation between SME groups ("Manipulation Check")

In order to measure whether the SME groups really differed from one another along the predicted dimensions (see Appendix A), the SMEs were asked to provide self-ratings on several measures. First, they were asked to fill out the BARS and BOS of the seven dimensions described above (see Appendix D) in reference to their own behavior. Next, they completed a rating scale of the importance of the 12 dimensions of college student performance (see Appendix G). This scale was to be used in conjunction with self-ratings to determine whether the personal characteristics and values of members of each SME group differed in the predicted ways. Next, they were asked to complete the Conscientiousness, Neuroticism, Extraversion, and Agreeableness dimensions of the IPIP (Goldberg, 1999). Finally, they were asked to provide information about their parental education and an estimate of their college GPA. The predicted differences between SME groups can be seen in Appendix A.

Results

Manipulation Check

In order to examine whether the members of the SME groups differed along the predicted dimensions, I conducted mean difference tests on each of the indicators to determine whether the members of the SME groups differed on each measure as predicted (see Appendix A for predictions and Table 2 for results of analyses). ANOVA tests that indicated significant differences between groups were followed by post hoc Tukey tests to determine the nature of the group differences.

Table 2

Manipulation Checks

Measure			Support for Hypotheses?
College GPA	Predicted:	Graduates would be highest	Yes
	Found:	Graduates (M = 3.73, SD = .23) and mentors (M = 3.54, SD = .23) had significantly higher GPA than undergrads (M = 3.12, SD = .54). Graduates and mentors did not differ significantly ($F(2,88) = 19.91, p < .001$).	
		Self-rated Behavior (BOS and BARS)	
Knowledge	Predicted:	Graduates would be higher than undergrads	Partial
	Found:	Graduate students (M = 2.34, SD = .66) and mentors (M = 2.20, SD = .49) both had significantly higher scores on the BOS than did undergraduates (M = 1.55, SD = .34), but were not significantly different from one another ($F(2,88) = 21.35, p < .001$). The groups did not differ on BARS ratings ($F(2,84) = 2.24$, *ns*).	
Continuous Learning	Predicted:	Graduates would be higher than undergrads	Yes
	Found:	Graduate students (M = 3.69, SD = .76) had significantly higher scores on the BOS than did both mentors (M = 3.23, SD = .72) and undergraduates (M = 3.04, SD = .65), who did not differ from one another ($F(2,88) = 6.75, p < .05$).	

Table 2 (cont'd)

		Graduates (M = 4.11, SD = .74) were higher than undergraduates (M = 3.40, SD = .84) on BARS. Mentors (M = 3.61, SD = .79) did not differ from either group on BARS ($F(2,87) = 6.12$, $p < .01$).	
Leadership Skills	Predicted:	Mentors would be highest	Partial
	Found:	Graduate students (M = 2.33, SD = .82) and mentors (M = 2.76, SD = .77) both had significantly higher BOS scores than did undergraduates (M = 1.88, SD = .58), although they did not differ from each other ($F(2,88) = 11.26$, $p < .001$). Mentors (M = 4.39, SD = .63) were higher on BARS than both grads (M = 3.39, SD = .74) and undergrads (M = 3.63, SD = 1.10), which don't differ from one another ($F(2,87) = 10.45$, $p <.001$).	
Interpersonal Skills	Predicted:	Mentors and undergraduates would be higher than graduates	No
	Found:	No differences (BOS: $F(2,88) = 1.32$, *ns*; BARS: $F(2,87) = 1.50$, *ns*).	
Social Responsibility	Predicted:	Mentors would be highest	Partial
	Found:	Mentors (M = 2.42, SD = .52) were higher than undergraduates (M = 1.95, SD = .58) on BOS, but graduates (M = 2.28, SD = .82) did not differ from either group ($F(2,88) = 4.30$, $p < .05$). The groups did not differ on BARS ($F(2,87)$ = 1.27, *ns*).	
Adaptability	Predicted:	Mentors would be highest	No
	Found:	No differences (BOS: $F(2.88) = 1.03$, *ns*; BARS: $F(2,87) = .18$, *ns*).	
Ethics	Predicted:	Mentors would be highest	No
	Found:	No differences (BOS: $F(2,88) = .16$, *ns*; BARS: $F(2,87) = .81$, *ns*).	
		Importance Ratings	
Knowledge	Predicted:	Graduates would be highest	No
	Found:	No differences ($F(2,88) = 1.53$, *ns*).	
Continuous Learning	Predicted:	Graduates would be highest	No
	Found:	No differences ($F(2,88) = .64$, *ns*).	
Artistic	Predicted:	No prediction (exploratory)	n/a
	Found:	No differences ($F(2,88) = 1.94$, *ns*).	
Multicultural	Predicted:	No prediction (exploratory)	n/a

Table 2 (cont'd)

	Found:	No differences ($F(2,88) = 1.17$, *ns*).	
Leadership	Predicted:	Mentors would be highest	Partial
	Found:	Mentors ($M = 6.04$, $SD = .84$) and Undergrads ($M = 5.94$, $SD = .83$) thought it more important than graduates ($M = 5.32$, $SD = .86$), but did not differ from each other ($F(2,88) = 6.04$, $p < .05$).	
Interpersonal	Predicted:	Undergraduates would be highest	No
	Found:	No differences ($F(2,88) = .64$, *ns*)	
Social Responsibility	Predicted:	Mentors would be highest	Partial
	Found:	Mentors ($M = 6.00$, $SD = 1.02$) and undergrads ($M = 5.88$, $SD = .93$) thought it more important than graduates ($M = 4.82$, $SD = 1.66$), but did not differ from each other ($F(2,88) = 7.96$, $p = .001$).	
Health	Predicted:	No prediction (exploratory)	n/a
	Found:	No differences ($F(2,88) = .35$, *ns)*	
Career	Predicted:	No prediction (exploratory)	n/a
	Found:	No differences ($F(2,88) = .18$, *ns*)	
Adaptability	Predicted:	Mentors would be highest	n/a
	Found:	No differences ($F(2,88) = .58$, *ns*)	
Perseverance	Predicted:	No prediction (exploratory)	n/a
	Found:	No differences ($F(2,88) = .11$, *ns*)	
Ethics	Predicted:	Mentors would be highest	Partial
	Found:	Mentors ($M = 6.68$, $SD = .82$) and undergrads ($M = 6.55$, $SD = .67$) thought it more important than Grads ($M = 5.71$, $SD = 1.08$), but did not differ from each other ($F(2,88) = 10.45$, $p < .001$).	
		Personality	
Agreeableness	Predicted:	Mentors and undergrads would be higher than graduates	No
	Found:	No differences ($F(2,88) = .57$, *ns*)	
Conscientiousness	Predicted:	Graduates and mentors would be higher than undergrads	No
	Found:	No differences ($F(2,88) = .35$, *ns*)	
Emotional Stability	Predicted:	Mentors and undergrads would be higher than graduates	No
	Found:	No differences ($F(2,88) = .01$, *ns*)	
Extraversion	Predicted:	Undergrads would be highest	No
	Found:	No differences ($F(2,88) = .55$, *ns*)	

Table 2 (cont'd)

		Experience	
Family Support of Education	Predicted:	Graduates would be highest	No
	Found:	No differences ($F(2,88) = 1.72$, *ns*)	
Mother's Education	Predicted:	Graduates would be highest	No
	Found:	No differences ($F(2,88) = .01$, *ns*)	
Father's Education	Predicted:	Graduates would be highest	No
	Found:	No differences ($F(2,88) = 1.02$, *ns*)	

GPA

As predicted, graduate students reported the highest college GPA ($M = 3.73$, $SD = .23$). The GPA reported by graduate students was significantly greater than that of undergraduates ($M = 3.12$, $SD = .54$), but the GPA of mentors ($M = 3.54$, $SD = .23$) did not differ from either group ($F(2,88) = 19.91$, $p < .001$).

Knowledge and Continuous Learning Behaviors

In support of predictions, graduate students ($M = 3.69$, $SD = .76$) rated themselves significantly higher than mentors ($M = 3.23$, $SD = .72$) and undergraduates ($M = 3.04$, $SD = .65$) on continuous learning on the BOS ($F(2,88) = 6.75$, $p < .05$). On the BARS, graduate students ($M = 4.11$, $SD = .74$) rated themselves significantly higher than did undergraduates ($M = 3.40$, $SD = .84$), but mentors ($M = 3.61$, $SD = .79$) did not differ from either group ($F(2,87) = 6.12$, $p < .01$). In partial support of predictions, graduate students ($M = 2.34$, $SD = .66$) and mentors ($M = 2.20$, $SD = .49$) rated themselves significantly higher than undergraduates ($M = 1.55$, $SD = .34$) on the knowledge scale of the BOS ($F(2,88) = 21.35$, $p < .001$), but not the BARS (on which the groups did not differ; $F(2,84) = 2.24$, *ns*).

Leadership, Social Responsibility, Adaptability, and Ethics Behaviors

In partial support of predictions, mentors (M = 4.39, SD = .63) rated themselves as higher than both graduate students (M = 3.39, SD = .74) and undergraduates (M = 3.63, SD = 1.10) on the leadership scale of the BARS ($F(2,87) = 10.45$, $p < .001$). On the BOS, however, both mentors (M = 2.76, SD = .77) and graduate students (M = 2.33, SD = .82) rated themselves as greater than undergraduates (M = 1.88, SD = .58) on leadership ($F(2,88) = 11.26$, $p < .001$). Again, in partial support of predictions, mentors (M = 2.42, SD = .52) rated themselves as higher than undergraduates (M = 1.95, SD = .58) on social responsibility on the BOS ($F(2,88) = 4.30$, $p < .05$), but neither group differed from graduate students (M = 2.28, SD = .82). The groups did not differ on social responsibility measured by the BARS ($F(2,87) = 1.27$, *ns*). Contrary to predictions, the three groups did not differ on ratings of adaptability (BOS: $F(2.88) = 1.03$, *ns*; BARS: $F(2,87) = .18$, *ns*) or ethics (BOS: $F(2,88) = .16$, *ns*; BARS: $F(2,87) = .81$, *ns*).

Interpersonal Behaviors

Contrary to predictions that undergraduates would rate themselves the strongest on interpersonal behaviors, the three groups did not differ on ratings of interpersonal behaviors (BOS: $F(2,88) = 1.32$, *ns*; BARS: $F(2,87) = 1.50$, *ns*).

Ratings of Importance for College Students

In partial support of predictions, Mentors (M = 6.04, SD = .84) and Undergrads (M = 5.94, SD = .83) thought leadership more important than graduates (M = 5.32, SD = .86), but did not differ from each other ($F(2,88) = 6.04$, $p < .05$). Mentors (M = 6.00, SD = 1.02) and undergrads (M = 5.88, SD = .93) thought social responsibility more important than graduates (M = 4.82, SD = 1.66), but did not differ from each other ($F(2,88) = 7.96$, $p = .001$). Mentors (M = 6.68, SD = .82) and undergrads (M = 6.55, SD = .67) thought

ethics more important than Grads ($M = 5.71$, $SD = 1.08$), but did not differ from each other ($F(2,88) = 10.45$, $p < .001$). The groups did not differ on the other dimensions (Knowledge: $F(2,88) = 1.53$, *ns*; Continuous Learning: $F(2,88) = .64$, *ns*; Artistic: $F(2,88) = 1.94$, *ns*; Multicultural: $F(2,88) = 1.17$, *ns*; Interpersonal: $F(2,88) = .64$, *ns*, Health: $F(2,88) = .35$, *ns*; Career: $F(2,88) = .18$, *ns*; Adaptability: $F(2,88) = .58$, *ns*; Perseverance: $F(2,88) = .11$, *ns*).

Personality Traits

Contrary to predictions, the groups did not differ on agreeableness ($F(2,88) = .57$, *ns*), conscientiousness ($F(2,88) = .35$, *ns*), emotional stability ($F(2,88) = .01$, *ns*), or extraversion ($F(2,88) = .55$, *ns*).

Demographics

Contrary to expectations, the groups did not differ on family support of education($F(2,88) = 1.72$, *ns*), or mother's ($F(2,88) = .01$, *ns*) or father's level of education ($F(2,88) = 1.02$, *ns*).

Summary of Findings for Manipulation Check

Overall, there was very little evidence that the three SME groups differed on the dimensions along which they were predicted to differ. There was some evidence to suggest that graduate student SMEs were more academically inclined than the other two groups (they had higher grades in college and reported more continuous learning behaviors). Mentors did not demonstrate more leadership and service related behaviors and undergraduates did not report being skilled in interpersonal interactions. The groups had similar values about the importance of the various dimensions for college student performance, similar personality traits, experienced similar family support for their

education, and reported that their parents had similar levels of education. Despite the finding that the groups did not differ in the expected ways, I continued with the planned analyses for two reasons; I wanted to explore whether my hypothesis that different groups would provide different judgments would still be supported, even if the differences in judgments were not of the expected type, and I also viewed this research as a developmental experience.

Hypothesis 1

Hypothesis 1 was that there would be significant differences in the effectiveness judgments of the different groups of SMEs (graduate students, mentors, and undergraduates) for response options classified into each of the three hypothesized dimensions (academic, leadership/service, and social). In order to examine this hypothesis, I compared the average ratings of effectiveness for each of the 69 response options that were identified by the Phase 1 participants as classifiable into the three dimensions across groups. I expected that the options for which I found significant differences in ratings of effectiveness by each group to reflect the different hypothesized lay theories of each SME group. Specifically, I expected that, out of the three SME groups, graduate SMEs would rate response options classified as academic as most effective, that mentor SMEs would rate leadership/service options as most effective, and undergraduate SMEs would rate social response options as most effective.

In order to make comparisons across groups, I used the following two different analytic techniques: analysis of variance and standardized mean difference comparisons. First, I conducted analysis of variance comparisons across the ratings of effectiveness for each of the 69 response options. The results of this analysis can be found in Table 3. As

can be seen in the table, the ANOVAs indicated that the three groups of SMEs differed on the effectiveness ratings for only seven of the 69 response options. Of those seven response options, post hoc Tukey tests indicated that Hypothesis 1 was partially supported for only three of those response options. For option 16a, which Phase 1 participants categorized into the social dimension, post hoc analyses indicated that undergraduates ($M = 4.39$, $SD = .72$) rated the option as significantly more effective than mentors ($M = 3.69$, $SD = .88$), but that neither group differed in their ratings of effectiveness from the graduate students ($M = 4.14$, $SD = .71$). For option 16b, which Phase 1 participants also categorized into the social dimension, post hoc analyses indicated that, once again, undergraduates ($M = 3.55$, $SD = 1.18$) rated the option as significantly more effective than mentors ($M = 2.77$, $SD = 1.14$), but that neither group differed in their ratings of effectiveness from the graduate students ($M = 3.23$, $SD = 1.05$). For options 16a and 16b, Hypothesis 1 was partially supported because undergraduate SMEs rated those response options classified as social as more effective than did mentors, although graduate students did not differ from either group. For option 23a, which Phase 1 participants classified into the academic dimension, post hoc analyses indicated that graduate students ($M = 4.21$, $SD = .74$) rated the option as significantly more effective than did mentors ($M = 3.58$, $SD = .81$), but that the ratings of undergraduates ($M = 3.97$, $SD = .87$) did not differ from either group. For option 23a, Hypothesis 1 was partially supported because graduate SMEs rated the option, which was classified as academic, as more effective than did mentors, although undergraduates did not differ in their ratings from either group.

The results from the ANOVA indicated very little support for Hypothesis 1. Very little evidence was found to suggest that the SME groups differed in their ratings of effectiveness for the response options. Of the 69 response options that were expected to elicit differences in ratings, only seven items elicited significantly different ratings across groups and only three of those options were in the expected direction.

In addition to the analysis of variance, I also compared the standardized mean differences (d) across each pair of SME groups.

Table 3

Comparison of Effectiveness Ratings of the SJT Response Options across SME Groups

Item	Option	Classified Dimension	F-test	Tukey test results	d for G-M	d for G-U	d for M-U
1	a	Leadership/Service	F(2,84) = 1.31		-0.29	-0.40	-0.12
	c	Leadership/Service	F(2,84) = 6.97*	U>M	0.38	-0.57	-1.09
	d	Leadership/Service	F(2,84) = .82		0.07	-0.24	-0.32
2	c	Academic	F(2,84) = .52		0.17	0.27	0.10
	d	Social	F(2,84) = .45		0.20	-0.06	-0.24
3	b	Leadership/Service	F(2,84) = 3.00		-0.66	-0.43	0.20
	d	Academic	F(2,84) = .14		-0.08	-0.13	-0.06
	f	Social	F(2,83) = .83		0.28	0.31	0.01
4	c	Academic	F(2,84) = .02		0.04	0.05	0.02
	d	Academic	F(2,84) = 1.09		-0.34	0.00	0.35
7	b	Academic	F(2,83) = .28		0.04	-0.14	-0.21
	c	Academic	F(2,83) = 1.06		0.11	0.37	0.25
8	a	Leadership/Service	F(2,84) = .13		0.04	0.12	0.10
9	a	Social	F(2,84) = .60		-0.29	-0.12	0.19
10	c	Academic	F(2,84) = 1.15		-0.12	-0.39	-0.26
	f	Social	F(2,83) = .50		0.27	0.06	-0.20
11	c	Academic	F(2,82) = .22		0.14	-0.05	-0.17
	e	Academic	F(2,84) = .08		-0.06	-0.09	-0.05
12	b	Academic	F(2,84) = 1.16		-0.35	-0.34	-0.04
13	b	Leadership/Service	F(2,84) = .18		0.16	0.09	-0.07
	c	Leadership/Service	F(2,84) = .16		0.12	0.14	0.04
	d	Social	F(2,83) = .35		0.22	0.00	-0.20
	e	Leadership/Service	F(2,83) = .14		0.13	0.01	-0.13
14	b	Academic	F(2,82) = 1.48		0.20	-0.25	-0.47
	f	Leadership/Service	F(2,84) = 3.47*	G>U	0.55	0.64	0.16
15	c	Leadership/Service	F(2,84) = 1.53		-0.09	-0.44	-0.34
	d	Social	F(2,84) = 1.57		0.28	0.47	0.17

Table 3 (cont'd)

	e	Academic	F(2,84) = 1.80		0.13	-0.33	-0.51
16	a	Social	F(2,84) = 5.88*	U>M	0.57	-0.34	-0.88
	b	Social	F(2,84) = 3.45*	U>M	0.47	-0.24	-0.67
	d	Social	F(2,84) = .49		-0.19	-0.25	-0.05
17	d	Social	F(2,83) = 1.36		-0.47	-0.07	0.35
18	e	Academic	F(2,84) = 1.69		0.47	0.12	-0.37
19	b	Social	F(2,84) = 2.80		0.55	0.51	-0.10
20	a	Academic	F(2,84) = .56		0.01	-0.23	-0.25
21	a	Academic	F(2,84) = .65		0.04	-0.22	-0.32
	b	Academic	F(2,84) = .29		-0.07	-0.19	-0.13
22	a	Academic	F(2,83) = .55		-0.01	-0.23	-0.29
	b	Academic	F(2,83) = 1.58		0.45	0.16	-0.34
	d	Social	F(2,84) = .24		-0.02	0.15	0.18
23	a	Academic	F(2,84) = 4.21*	G>M	0.83	0.30	-0.46
	b	Academic	F(2,84) = 1.83		-0.03	-0.44	-0.47
	c	Academic	F(2,84) = .38		-0.21	-0.20	0.00
25	a	Social	F(2,84) = .47		0.20	-0.04	-0.25
	b	Leadership/Service	F(2,84) = 3.17*	U=M=G	-0.04	-0.58	-0.51
26	a	Academic	F(2,84) = 1.09		-0.12	-0.36	-0.28
	c	Academic	F(2,84) = 2.49		-0.52	-0.48	-0.02
27	b	Social	F(2,83) = .14		0.00	-0.11	-0.12
	c	Social	F(2,83) = .11		0.11	-0.01	-0.13
	d	Social	F(2,83) = .94		-0.37	-0.27	0.07
29	a	Leadership/Service	F(2,84) = .33		-0.20	-0.18	0.02
	d	Social	F(2,83) = 1.33		0.13	-0.29	-0.43
30	a	Academic	F(2,84) = 5.10*	U>M	0.22	-0.58	-0.94
	b	Social	F(2,84) = .09		-0.03	-0.09	-0.08
	c	Academic	F(2,84) = 2.67		0.36	-0.25	-0.63
31	b	Academic	F(2,83) = .96		0.39	0.10	-0.27
	e	Academic	F(2,83) = 1.31		0.24	-0.21	-0.42
	g	Social	F(2,84) = 1.37		0.40	0.38	0.07
33	a	Academic	F(2,84) = 1.03		0.16	-0.21	-0.41
	b	Social	F(2,84) = .65		0.29	0.03	-0.26
	c	Academic	F(2,84) = .20		0.18	0.04	-0.13
	d	Academic	F(2,84) = .88		0.21	-0.16	-0.33
34	a	Academic	F(2,84) = .23		-0.07	-0.16	-0.11
	b	Academic	F(2,84) = .07		-0.08	-0.09	-0.02
	c	Academic	F(2,84) = .09		0.09	0.00	-0.10
	d	Academic	F(2,84) = .80		0.20	-0.15	-0.32
	e	Social	F(2,84) = .96		0.36	0.28	-0.07
35	c	Leadership/Service	F(2,84) = 1.47		-0.05	-0.39	-0.40
	d	Leadership/Service	F(2,84) = 2.95		-0.58	-0.51	0.01

Note. The p-value for F-tests marked with an asterisk was less than .05. U = undergraduate group, M = mentor group, G = graduate student group.

In the last three columns of Table 3, the d-values are reported for each pair (Graduate-Mentor, Graduate-Undergraduate, Mentor-Undergraduate). As may be observed in the table, a fair number of the differences would be considered meaningful according to Cohen's standards (1977; d-values of 0.2-0.5 are considered a small difference, 0.5-0.8 a moderate difference, and 0.8 and above a large difference). In order to evaluate whether the three SME groups differed meaningfully in their ratings of effectiveness using d-values, I examined all of the meaningful (according to Cohen's criteria, 0.2 and above) d-values for each dimension relative to the total number of response options that were classified as belonging to each dimension. I created profiles to indicate which types of findings would strongly and weakly support or refute Hypothesis 1. For example, 34 of the response options were classified as belonging to the academic dimension. Of those 34 response options, 3 indicated that graduate students rated the options as more effective than undergraduates and 11 indicated that graduate students rated the options as more effective than mentors, both of which would be considered support for the hypothesis (see Table 4). In contrast, for 15 options, undergraduates rated them as more effective than graduate students and, for 3 items, mentors rated them as more effective than graduate students. Overall, the pattern of responses indicated that although graduate students rated academic options as more effective than did mentors, undergraduates rated academic options as even more effective than did graduate students. This pattern is contradictory to Hypothesis 1, which predicted that graduate students would rate academic options as more effective than either of the other SME groups.

Table 4

Comparison of Differences across Groups for Ratings of Effectiveness of Options in the Academic Dimension (k = 34)

	Support		Against	
	G>U	G>M	U>G	M>G
k	3	11	15	3
%	8.82	32.35	44.12	8.82

Overall, the pattern of response for options classified into the leadership/service dimension (k = 14; see Table 5) indicated that although mentors thought leadership/service items were more effective than graduates did, undergraduates rated leadership/service items more effective than did mentors.

Table 5

Comparison of Differences across Groups for Ratings of Effectiveness of Options in the Leadership/Service Dimension (k=14)

	Support		Against	
	M>G	M>U	G>M	U>M
k	4	1	2	5
%	28.57	7.14	14.29	35.71

Once again, this pattern is counter to the predictions of Hypothesis 1, that mentors would rate the leadership/service options as most effective relative to the other two SME groups. The pattern of response for options classified into the social dimension (k = 21; see Table

6) indicated that undergraduates rated social options as more effective than mentors and that undergraduates and graduates rated social options as approximately equally effective. This pattern partially supports the prediction of Hypothesis 1 that undergraduates would report the highest effectiveness ratings for the social options.

Table 6

Comparison of Differences across Groups for Ratings of Effectiveness of Options in the Social Dimension (k=21)

	Support		Against	
	U>G	U>M	G>U	M>U
k	5	8	5	1
%	23.81	38.10	23.81	4.76

Taken together, the ANOVA analyses and standardized mean difference comparisons for the 69 options for which the three SME groups were hypothesized to differ in ratings of effectiveness across the predicted dimensions indicate very little support for Hypothesis 1.

Hypothesis 2

Hypothesis 2 was that there would be significant differences in the correlates of scores based on the scoring keys developed from different SME groups with external criteria measures. In order to examine this hypothesis, correlations between each of the three scoring keys (graduate, mentor, and undergraduate) and the external variables were computed. Tests of the significance of the differences in the correlations in the predicted directions (see Appendix A) would be considered support (or lack of support) for Hypothesis 2. In order to test for differences among the correlations of the scores

developed from the keys for each group with each criterion, the procedure for testing the heterogeneity of a set of correlated correlations described by Meng, Rosenthal, and Rubin (1992) was used. The results for each of the analyses can be found in Table 7. Although the scoring keys for each of the three groups were correlated with most of the criteria measures, there were no differences in the relationships between the three scoring keys and any of the criteria. Thus, there was no support for Hypothesis 2.

Table 7

Comparisons of Correlations for Hypothesis 2

Criterion Measure		USJT	MSJT	GSJT	chi-sq.	df
First Year GPA	r	.32	.31	.31	.17	2
	N	1519	1519	1519		
Second Year GPA	r	.32	.31	.31	.09	2
	N	1383	1383	1383		
Third Year GPA	r	.30	.29	.29	.14	2
	N	1318	1318	1318		
Fourth Year GPA	r	.25	.24	.25	.09	2
	N	1231	1231	1231		
Cumulative 4-Year GPA	r	.31	.29	.29	.58	2
	N	1867	1867	1867		
1st semester BARS Knowledge	r	.20	.22	.23	.66	2
	N	1140	1140	1140		
1st semester BARS Continuous Learning	r	.05	.08	.08	.48	2
	N	1140	1140	1140		
1st semester BARS Leadership	r	.04	.09	.08	1.62	2
	N	1137	1137	1137		
1st semester BARS Interpersonal	r	.05	.09	.08	.91	2
	N	1140	1140	1140		

Table 7 (cont'd)

1st semester BARS Social Responsibility	r	.07	.10	.08	.68	2
	N	1139	1139	1139		
1st semester BARS Adaptability	r	.00	.01	.01	.03	2
	N	1139	1139	1139		
1st semester BARS Ethics	r	.21	.24	.23	.43	2
	N	1141	1141	1141		
1st year BARS Knowledge	r	.22	.20	.23	.35	2
	N	980	980	980		
1st year BARS Continuous Learning	r	.03	.07	.06	.82	2
	N	982	982	982		
1st year BARS Leadership	r	.12	.17	.16	2.38	2
	N	982	982	982		
1st year BARS Interpersonal	r	.13	.17	.16	1.41	2
	N	977	977	977		
1st year BARS Social Responsibility	r	.14	.19	.17	1.77	2
	N	978	978	978		
1st year BARS Adaptability	r	0.06	.09	.09	.82	2
	N	983	983	983		
1st year BARs Ethics	r	.20	.21	.23	.49	2
	N	984	984	984		
3rd semester BARS Knowledge	r	.26	.25	.27	.26	2
	N	872	872	872		
3rd semester BARS Continuous Learning	r	.05	.07	0.06	.25	2
	N	873	873	873		
3rd semester BARS Leadership	r	.14	.19	.20	2.28	2
	N	871	871	871		
3rd semester BARS Interpersonal	r	.11	.15	.15	.97	2
	N	870	870	870		
3rd semester BARS Social Responsibility	r	.13	.18	.17	1.16	2
	N	870	870	870		

Table 7 (cont'd)

3rd semester BARS Adaptability	r	.03	0.06	0.06	.52	2
	N	868	868	868		
3rd semester BARs Ethics	r	.24	.25	.25	.07	2
	N	873	873	873		
4th year BARS Knowledge	r	.20	.19	.20	.01	2
	N	594	594	594		
4th year BARS Continuous Learning	r	.05	.05	.06	.02	2
	N	592	592	592		
4th year BARS Leadership	r	.12	.16	.14	.59	2
	N	592	592	592		
4th year Interpersonal	r	.15	.18	.17	.31	2
	N	590	590	590		
4th year BARS Responsibility	r	.09	.13	.11	.63	2
	N	594	594	594		
4th year BARS Adaptability	r	.01	.02	.01	.03	2
	N	593	593	593		
4th year BARS Ethics	r	.20	.21	.21	.05	2
	N	591	591	591		
BOS Knowledge	r	.06	.08	.07	.32	2
	N	872	872	872		
BOS Continuous Learning	r	.07	.09	.09	.21	2
	N	865	865	865		
BOS Leadership	r	0.02	.07	.08	2.11	2
	N	872	872	872		
BOS Interpersonal	r	.15	.15	.16	.04	2
	N	865	865	865		
BOS Social Responsibility	r	.09	.14	.13	1.25	2
	N	872	872	872		
BOS Adaptability	r	.11	.10	.10	.15	2
	N	848	848	848		
BOS Ethics	r	.17	.22	.20	1.38	2

Table 7 (cont'd)

	N	871	871	871		
Conscientiousness	r	.21	.23	.23	1.19	2
	N	2686	2686	2686		
Emotional Stability	r	.03	.06	.05	1.24	2
	N	2685	2685	2685		
Extraversion	r	.02	.04	.03	.62	2
	N	2685	2685	2685		
Agreeableness	r	.27	.31	.29	4.39	2
	N	2686	2686	2686		

Note. Correlations above .05 are significant at $p < .05$. The significant Chi-square value with 2 degrees of freedom is 5.99.

Hypothesis 3

Hypothesis 3 was that there would be significant differences between groups of SMEs in references made to criteria during a verbal protocol analysis. Because the results from the manipulation check and Hypothesis 1 indicated very little support for the prediction that the groups differed along the predicted dimensions, I developed 26 new coding categories from the content of the verbal protocols (see Appendix I). The new coding categories referred to reasons for or against choosing a response option. Three graduate students in Industrial/Organizational Psychology who were not part of the Phase 1 data collection and who were unaware of the hypotheses and theoretical background of the study coded the verbal protocol analyses based on the new coding scheme.

To systematize the coding task, I first extracted all phrases from the transcriptions of the verbal protocols that referred to reasons for selecting or not selecting a specific response option. There were 360 distinct phrases or sentences extracted from the verbal

protocols. The coders then coded the short phrases or sentences rather than content analyzing entire verbal protocol transcriptions. The coders were blind to group membership of the "speaker" and were also unaware that there were different groups of speakers. They were presented with the item stem and response option corresponding to each of the phrases extracted from the transcripts.

After the coding was complete, I determined that all three coders had agreed on codes for 149 (41%) of the 360 phrases. Two out of the three coders agreed on codes for 294 (82%) of the 360 phrases. For the following analyses, I assigned codes to the 294 phrases based on the agreement of two out of three of the coders.

In order to compare whether the three SME groups differed in terms of the reasons they verbally assigned to why they did or did not select a response option, I created a frequency count for how many SMEs from each group responded according to each code for each item. Because each group of SMEs provided different numbers of code-able phrases, I then divided the frequency for each code by the total number of phrases elicited from each group to create a proportion. All of the proportions are presented in Table 8. The SJT items, responses, and most common codes are presented in Figure 1.

Table 8

Results from the Verbal Protocol Analysis

	Item 1			Item 2			Item 3		
Code	Grads	Mentors	Undergrads	Grads	Mentors	Undergrads	Grads	Mentors	Undergrads
1	0.04	0.06	0.00	0.08	0.00	0.00	0.00	0.00	0.00
2	0.04	0.06	0.00	0.04	0.00	0.00	0.04	0.00	0.00
3	0.14	0.12	0.11	0.08	0.05	0.13	0.04	0.07	0.09
4	0.11	0.18	0.33	0.25	0.38	0.25	0.17	0.30	0.36
5	0.00	0.12	0.11	0.04	0.05	0.00	0.00	0.00	0.00

Table 8 (cont'd)

6	0.04	0.00	0.11	0.00	0.00	0.25	0.00	0.04	0.00
7	0.11	0.06	0.11	0.00	0.00	0.00	0.08	0.04	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.09
9	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
10	0.00	0.00	0.00	0.08	0.10	0.00	0.08	0.11	0.09
11	0.00	0.00	0.00	0.08	0.00	0.00	0.17	0.11	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
13	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.04	0.05	0.13	0.00	0.00	0.00
15	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.09
16	0.04	0.12	0.00	0.08	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00
18	0.11	0.18	0.11	0.04	0.10	0.13	0.08	0.15	0.00
19	0.00	0.06	0.00	0.00	0.05	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.05	0.00	0.04	0.00	0.09
21	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00
22	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.08	0.05	0.00	0.08	0.07	0.09
24	0.00	0.00	0.00	0.04	0.10	0.00	0.00	0.04	0.00
25	0.04	0.00	0.00	0.00	0.05	0.00	0.04	0.00	0.09
26	0.00	0.00	0.00	0.04	0.00	0.00	0.04	0.00	0.00

	Item 4			Item 5			Item 6		
Code	Grads	Mentors	Undergrads	Grads	Mentors	Undergrads	Grads	Mentors	Undergrads
1	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.15	0.00	0.00	0.11	0.07	0.00	0.07	0.07	0.00
4	0.08	0.17	0.40	0.47	0.67	0.20	0.50	0.47	0.42
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.16	0.07	0.10	0.00	0.00	0.00
7	0.23	0.11	0.20	0.05	0.00	0.20	0.07	0.00	0.08
8	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.27	0.17
10	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17
12	0.08	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.08	0.06	0.07	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.16	0.20	0.30	0.00	0.00	0.00
16	0.15	0.17	0.07	0.05	0.00	0.10	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17

Table 8 (cont'd)

18	0.00	0.06	0.00	0.00	0.00	0.10	0.00	0.13	0.00
19	0.00	0.06	0.13	0.00	0.00	0.00	0.00	0.07	0.00
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.08	0.11	0.00	0.00	0.00	0.00	0.21	0.00	0.00
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	0.00	0.11	0.07	0.00	0.00	0.00	0.00	0.00	0.00
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	All Items		
Code	Grads	Mentors	Undergrads
1	0.04	0.01	0.00
2	0.03	0.01	0.00
3	0.11	0.06	0.05
4	0.26	0.34	0.34
5	0.01	0.03	0.02
6	0.04	0.02	0.06
7	0.09	0.04	0.11
8	0.01	0.01	0.02
9	0.03	0.04	0.03
10	0.04	0.05	0.02
11	0.05	0.03	0.03
12	0.03	0.02	0.00
13	0.03	0.01	0.02
14	0.01	0.01	0.03
15	0.03	0.03	0.08
16	0.05	0.04	0.03
17	0.00	0.01	0.03
18	0.05	0.11	0.05
19	0.00	0.04	0.03
20	0.01	0.01	0.02
21	0.00	0.00	0.02
22	0.00	0.01	0.00
23	0.06	0.04	0.02
24	0.01	0.03	0.00
25	0.03	0.03	0.03
26	0.02	0.00	0.00

Figure 1

SJT Items and Response Options from the Verbal Protocol and the Most Common Codes per Item

Item 1

You find that you are eating more fattening and greasy food than normal, and that you have not been getting sufficient exercise. You have gained 15 pounds but find it difficult to change your eating and exercising habits. How would you deal with this situation?

a. Start slowly by cutting out two snacks a day.
b. Don't worry about it. You only live once, so eat what you want.
c. Get help from someone with experience in this area, such as a health professional or nutritionist.
d. Get some friends together and exercise together. There is power in numbers.
e. Try to establish a regular exercise routine, and focus on eating healthy foods.

Most Common Codes for Responses to Item 1 (Across Options)

(4) It would not solve the problem.
    33% of undergraduate responses
    18% of mentor responses
    11% of graduate student responses
(18) I think this would work well because of what I know about human nature (for example, setting clear goals helps motivate people).
    18% of mentor responses
    11% of graduate student responses
(3) Something else would need to be added for this option to work.
    14% of graduate student responses
(7) I just don't like to do this or I am not willing to do this.
    11% of graduate student responses

Item 2

You are part of a three-person group working on a class project with a quickly approaching deadline. One member of the team is not pulling his/her weight and avoids assignments, complains about the amount of work that has to be done, and says the project doesn't really matter anyway. While you are all classmates, you seem to be the group leader. What would you do?

a. Divide the workload evenly among members of the group, making sure everyone knows they are responsible for their share. If the group member still does not pull his/her own

Figure 1 (cont'd)

weight, bring it up with the instructor.

b. Speak with the group member and offer him/her encouragement to complete their portion of the project. If the group member still does not contribute, bring it up with the instructor.

c. Try to get the team member motivated to do a share of the work. If that doesn't help the situation, just put more effort into the project yourself in order to complete it.

d. Just do the group member's portion of the assignment in addition to your own, and tell the instructor about the situation.

e. See if the person could be removed from your group.

f. Consult with the group member who is not a problem about the most appropriate course of action, and then act on whatever you both decide.

Most Common Codes for Responses to Item 2 (Across Options)

(4) It would not solve the problem.
    38% of mentor responses
    25% of graduate student responses
    25% of undergraduate responses
(6) It would take too much time or require too much effort.
    25% of undergraduate responses

Item 3

You and five other students must have a report ready within 48 hours. The last time the six of you worked together, you became the leader. You know that one of the group members did no work whatsoever on the last occasion, yet she is in your group again. This time it is necessary that all members pull their own weight. What would you do?

a. Let her know that you are aware that she did not do any work last time, and that this time it is necessary that she fully contribute.

b. Do your entire end of the work and ensure that the instructor is aware that you did your share, regardless of what the other members do.

c. Explain to the group that the professor will be made aware of who contributed what to the project, and ensure that this happens.

d. Stress the importance that everyone fully contributes his or her share to the project.

e. Work as closely with her as possible (e.g. assign both of you a related task) so as to offer encouragement and to ensure that her work gets done.

f. Assign her a specific task with a specific timeframe. If she does not do the work, ask to have her re-assigned, and have the group pick up her work.

Figure 1 (cont'd)

Most Common Codes for Responses to Item 3 (Across Options)

(4) It would not solve the problem.
    36% mentor responses
    30% undergraduate responses
    17% graduate student responses
(11) It is too confrontational or would offend other people.
    17% of graduate student responses

Item 4

You are working together with other classmates on a project. Your group keeps running into a variety of problems that threaten to cause the project to be late. The other group members want to just plan to submit it late. Another option would be to devote much more time than planned to the project and possibly get it in on time. What would you do?

a. Try to get it done, but plan to submit it late.
b. Ask the instructor for help or for an extension. If that doesn't work, just try your best and do what you can or turn it in late.
c. Motivate the group to devote more time and work together to get it done.
d. Have the group decide what to do.
e. Work hard to finish it because there are consequences for being late, and meeting deadlines is important to you.
f. Tell the instructor your situation, and ask for advice.

Most Common Codes for Responses to Item 4 (Across Options)

(4) It would not solve the problem.
    40% of undergraduate responses
    17% of mentor responses
(16) This is something I've done in the past and it worked for me.
    17% of mentor responses
(7) I just don't like to do this or I am not willing to do this.
    23% of graduate responses

Item 5

An event in the news makes you wonder about the history behind it. What would you do?

a. Do some research, looking up all the facts for yourself.

Figure 1 (cont'd)

b. Do a quick Internet search to see if you could find any information.
c. Think about it briefly, then move on.
d. Ask others what they know about the topic.
e. Resolve to read the newspaper more often.

Most Common Codes for Responses to Item 5 (Across Options)

(4) It would not solve the problem.
    67% of mentor responses
    47% of graduate student responses
(15) It would work quickly or with little effort.
    30% of undergraduate responses

Item 6

One of your friends' roommates frequently parties until late at night, often returning to the room after drinking, engaging in loud and obnoxious behavior. Your friend finds that he/she cannot study or sleep well because of this, but also feels reluctant or afraid to talk with the resident assistant about this. What action would you take?

a. Approach the resident assistant on behalf of your friend.
b. Talk to the roommate yourself, and explain that his/her behavior bothers your friend.
c. Tell your friend to talk with the roommate and let him/her know that the behavior is not acceptable.
d. Offer to let your friend stay with you when necessary.
e. Suggest to your friend that he/she talk it out with the roommate, and offer to be available as a neutral third party when the two have the conversation.

Most Common Codes for Responses to Item 6 (Across Options)

(4) It would not solve the problem.
    50% of graduate student responses
    47% of mentor responses
    42% of undergraduate responses

In the last three columns of Table 8, the proportions are compared across all of the items.

Across all of the items, the most common reasoning SMEs provided was code 4, "This response would not be effective because it would not solve the problem."

An analysis of the most common reasons members of each group provided by item is somewhat more informative. For item 1, which described an unwanted weight gain and asked what the participant would do in that situation, 33% of the undergraduate responses were coded 4. Responses by mentors were most often coded 4 (18%) and 18 (18%), "I think this would work well because of what I know about human nature (for example, setting clear goals helps motivate people)." Responses by graduate students were most often coded 3 (14%), "Something else would need to be added for this option to work," closely followed by 11% of the responses coded as 4, 7, and 18 each. Code 7 was "I just do not like this or I am not willing to do this." Although there was significant overlap in the reasons that each group gave for their responses (most commonly 4 or 18), differences emerged between groups not only in reasoning but in diversity of reasoning. Whereas responses by graduate students were diverse, one third of the responses by undergraduates were based on the same reasoning.

For item 2, the largest proportion of responses for graduate students (25%) and mentors (38%) were coded 4. The responses by undergraduates were clustered at 4 (25%) and 6 (25%), "It would take too much time or require too much effort." For item 3, the largest proportion of responses for mentors (30%) and undergraduates (36%) was coded 4. The responses by graduate students were clustered at 4 (17%) and 11 (17%), "It is too confrontational or would offend other people." For item 4, the responses for undergraduates were most frequently coded 4 (40%). The responses for mentors were most frequently coded 4 (17%) and 16 (17%), "This is something I've done in the past and it worked for me." The responses for graduate students were most frequently coded 7 (23%). For item 5, the responses for graduate students (47%) and mentors (67%) were

most frequently coded 4. The responses for undergraduate students were most frequently coded 15 (30%), "It would work quickly or with little effort." For item 6, the responses for all three groups, graduate students (50%), mentors (47%), and undergraduates (42%) were most frequently coded 4.

Overall, there was a lot of diversity in responses from each group. The most common reasoning across all groups for all items was code 4, but analyses of patterns at the item level revealed more information. Undergraduates tended to prefer response options that would work quickly or without a lot of effort (together the codes accounted for 22% of their responses to item 1, 25% of their responses to item 2, and 40% of their responses to item 5) and to avoid options that would take too much time or effort. Mentors referred to what they knew about human nature or what had worked for them in the past (together, the codes accounted for 30% of their responses to item 1 and 23% of responses to item 4). There was the least amount of agreement among the reasoning by graduate students between and within items, suggesting that there was a lot of diversity of reasoning within that group.

Discussion

The purpose of this study was to explore whether the selection of the SMEs who provide scoring judgments during the development of an SJT has the potential to affect which constructs the SJT ultimately measures. In order to examine this question, three groups of SMEs (graduate students, resident advisors or mentors, and undergraduate students drawn from the Psychology subject pool) provided scoring judgments for an SJT intended to measure college student performance. The first hypothesis was that members of the three groups of SMEs would rate the effectiveness of the response options according to different implicit theories of desirable performance. Specifically, it was predicted that graduate student SMEs would view options related to academic performance as most effective, that mentors would view options related to leadership and service as most effective, and that undergraduates drawn from the subject pool would rate options related to social performance as most effective. Data regarding differences on dimensions relevant to these hypothetical differences between groups were collected from each of the three groups.

The theories presented in the introduction that support the idea that distinct groups of individuals develop similar implicit theories through shared experiences and traits all assume that groups differ from one another. In order to ensure that the groups of SMEs in this study did differ from one another and in expected ways, data were collected from the SMEs as a manipulation check. The manipulation check indicated that, although the groups differed in some of the expected ways, they were actually quite similar. For GPA, knowledge, and continuous learning self-rated behaviors, for which it was predicted that graduate students would be highest, there was evidence that undergraduates were indeed

lower than graduate students, but mentors were similar to both groups. On self-rated leadership behaviors and social responsibility, for which mentors were expected to be the highest, there was evidence that undergraduates were lower than mentors, but graduate students were similar to both groups. On self-rated adaptability and ethics behaviors, for which mentors were expected to be the highest, the groups did not differ at all. On self-rated interpersonal behaviors, for which undergraduates were expected to be the highest, the groups did not differ. The three groups had very similar ideas about what behaviors are important in college. They also scored very similarly on the personality traits of agreeableness, conscientiousness, emotional stability, and extraversion. In general, the three groups of SMEs were not very different. This finding can be seen as a contradiction of the theories indicating that groups become similar over time through a wide variety of processes or (more likely) the three groups chosen for this study did not meet the true definition of groups who share experiences and distinct traits.

Despite the finding that the three SME groups did not differ on all variables as expected, I proceeded with the planned analyses. The data did not support the first hypothesis. Although the three groups of SMEs did not agree about the effectiveness of a few of the response options, they did agree about the effectiveness of most of the response options. In addition, even for those response options about which they did not agree, the differences were not always in the expected direction. The second hypothesis was that scores on the SJT based on the scoring keys developed by the three different groups of SMEs would differentially predict criterion measures of college student performance. The second hypothesis was also not supported by the data. Although the scoring keys developed from the responses of the three different groups of SMEs did

predict a broad array of performance criteria, there were no differences in predictive capability across groups. This finding indicates that the three scoring keys were measuring the same set of constructs.

In order to further explore the reasoning behind the judgments of the SMEs in the scoring task, a sub-group of each SME group participated in a verbal protocol analysis task in which they reasoned aloud as they made their judgments for six of the SJT items. The third hypothesis was that members of the three SME groups would articulate different reasons for making their scoring judgments. Because of the lack of support for the original hypotheses, the original predictions that graduate students would refer to academic reasons, mentors would refer to leadership and service reasons, and undergraduates would refer to social reasons for making their judgments were ignored and a new set of codes were developed based on the content of the transcripts from the verbal protocols. Based on this new set of codes, there is some evidence that the groups did differ in the reasoning that they used for making their scoring judgments. The most common reasoning that each group gave for their decisions was that they disregarded an option because they believed that it simply would not work. There was relatively little agreement among graduate students for why they chose and rejected response options, suggesting that graduate students did not share an implicit theory of performance or at least that they did not articulate a uniform theory. In contrast, members of both the mentor and undergraduate groups responded in a relatively consistent manner using similar reasoning. Mentors tended to refer to what they knew about human nature and to what had worked for them in the past. Undergraduates tended to prefer options that they

believed would take little time or effort and reject options that they perceived would take a lot of time or effort.

The theoretical argument in the introduction was that different groups have different implicit theories of performance that come from their shared experiences and traits. Despite some evidence that the members of the three groups used different reasoning to make scoring judgments, there were no practical effects of the differences in reasoning for this study. Scores based on scoring keys developed by each group of SMEs were equivalent in terms of predicting a wide array of criteria, including GPA, a variety of behaviors, and personality traits. As will be discussed below, there are a number of limitations to the methodology of this study that may have impacted the results, but it must be emphasized that the main finding in this study is that the choice of SMEs for scoring key development did not impact construct measurement. In fact, although both graduate students and mentors could be viewed as "experts" in college student performance, both groups having demonstrated high performance as undergraduates, the undergraduates who contributed scoring judgments were not selected based on their success as students and, thus, would not typically be considered "experts" on performance. Their judgments, however, were as effective in predicting performance criteria as were the judgments of the graduate students and mentors.

Limitations and Future Research

There were a number of limitations to this study that may have affected the findings. First, the expert groups may not have been distinct enough to have found effects. Second, the dimensions along which the groups were expected to differ may not have been representative of true differences between the groups. Third, college student

performance may be a construct about which many people have good judgment. Fourth, the verbal protocol procedure may not have been conducted equally effectively across all three groups, and fifth, experts may not be very good at explaining how they make their decisions.

First, the choice of groups for SMEs may have limited the findings. The findings of the manipulation check indicated that the groups were fairly similar. The prediction was that groups have members with similar experiences and traits, so it may be that the groups were too similar in terms of experience and traits to have had different implicit theories of college student performance. It is also possible that the groups were not similar enough within groups to have led to distinct differences between groups. The graduate student SMEs were drawn from several departments and, thus, may not have had a shared identity or shared experiences as graduate students. It may be that specific departments within the graduate school attract certain types of students, but that graduate students in general do not share a profile of traits and characteristics. Mentors would seem likely to share a profile of experiences and traits because they are all attracted to and belong to a single program within the university, but the program is very large and there may be a large amount of diversity among mentors. The undergraduates shared only the experience of being students at the same large university and taking at least one class in the Psychology Department.

Future research should examine more distinct SME groups. For example, in this study, it may have been possible to examine a different set of groups that would still have been experts in college student performance. For example, the groups could have come from different universities. It is possible that individuals at different universities may

have different norms or lay theories of performance. I also could have included a faculty expert group. It may be that faculty members, who have experience teaching, advising, and mentoring undergraduates, have different theories about effective student performance than do students and former students. It should be emphasized that the scoring keys developed by all three expert groups effectively predicted performance along a variety of dimensions. The question here is whether experts from another university or setting might develop a scoring key that also effectively predicted student performance, but with differences in which constructs were best predicted.

Second, the dimensions along which the groups were expected to differ were not representative of real differences between the groups. In order to find meaningful differences between the groups on ratings of effectiveness if they truly existed, it would have been necessary to have developed dimensions that represented real differences between the groups. It is important to note that the findings that the three scoring keys did not predict the performance measures differently would not have been different even if the differences between the groups had been better determined. If a similar study is conducted to examine differences between groups of experts in the future, the choice of dimensions along which the experts are expected to differ should be based on pilot-testing, focus groups, or data of some kind. This would allow differences to become more apparent if differences in judgments of different groups really can have meaningful effects. However, use of different dimensions would make direct comparisons of the groups impossible.

The measure chosen for the study was an SJT intended to measure college student performance. It may be that the choice of the measure limited the possibility of finding

differences between groups because there is a high level of knowledge about appropriate behavior for college students among adults in general. The target of the measure is high school students who have no experience at the college level. The participants who took the SJT in 2004 were new college students in their first few days or weeks at school. All of the subject matter experts had been attending college for at least one year by the time they provided scoring judgments. It may be that the necessary level of knowledge for an expert is developed during the first year of college, so that students at the end of their first year are capable experts whereas students who have not yet entered college differ in their levels of practical judgment. If there is a common theory of college student performance among most people who have completed at least their first year of college, then there is a very large pool of experts who could legitimately provide judgments for scoring this measure. However, there may be many types of constructs for which there is a not a high level of agreement among the implicit theories of different groups.

Future research might also consider selecting a measure of a construct about which there are fewer "experts." If it is true that college student performance is a topic about which many people have a similar theory of performance, there may be many other types of measures for different lay theories of performance among different expert groups that might have a more profound effect on construct measurement. If this is the case, the findings of this study could be misleading.

Although the verbal protocol procedure did yield some differences between groups, there were some flaws in the procedure. First, the verbal protocol procedure was completed with the undergraduate student participants from the subject pool before the mentors and graduate students. The undergraduate students also provided the fewest

number of use-able phrases of the three groups. It is possible that I was less proficient at conducting the protocol procedure when working with the first five participants. I did follow the same protocol and provide the same instructions, but I may have unintentionally created a more comfortable atmosphere later on that led to more (or more useful) elicitation from mentors and graduate students. It is also possible that mentors and graduate students felt more comfortable with a graduate student researcher or that undergraduates tend to be less verbally fluent than mentors and graduate students. Mentors are selected for their positions in part for their interpersonal verbal skills and graduate students are selected for admission to graduate school in part for their high verbal scores on the Graduate Record Exam (GRE). Also, undergraduates earned class credit whereas members of the other groups earned money for their participation. It is possible that money was a stronger motivator than class credit. In addition, I had no training before leading participants through the verbal protocol procedure and the exercise may have been more effective at getting at participants' real thoughts if I were more skilled.

The other limitation with the verbal protocol is the technique itself. There is some evidence to suggest that experts may not be capable of explaining the real reasons why they make certain judgments. Glaser, Lesgold, and Gott (1991) suggested that when an individual reaches the level of automaticity (or the level of expertise) in a subject, he or she may no longer be able to verbalize intended behaviors or processes that are necessary for successful task performance. In addition to the problems with the procedures for the verbal protocol, it may be that subject matter experts may have trouble articulating the

true reasons for their decisions verbally. They may be unaware of the true rationale for their judgments and, thus, unable to explain them to others.

It is possible that, at least for some SJTs, the constructs measured are so well defined by the item stems and response options that the scoring system has little effect. In this case, the choice of SMEs for the development of critical incidents that become items and for the development of response options would have a much greater effect on construct measurement than the choice of SMEs for scoring. Although there has been research about the impact of the fidelity, length, complexity, and comprehensibility of item stems on construct measurement (Chan & Schmitt, 1997; McDaniel et al,, 2001; Sacco et al., 2000), there has been little research about how the content of item stems may impact construct measurement. There is also some evidence that the content of response options can be developed with the purpose of measuring specific constructs (for example, Motowidlo, Hooper, & Jackson, 2006). Future research further exploring how the development of item stems and response options impacts the constructs an SJT measures may be especially useful. However, we should not completely disregard the hypothesis that aspects of SJT development other than item development may impact construct measurement. As discussed earlier, there is also evidence that other aspects of SJT development, such as instruction sets, may impact construct measurement, even with identical stems and responses (McDaniel et al., 2007; McDaniel & Nguyen, 2001; Ployhart & Ehrhart, 2003).

Practical Implications

Because the results were contrary to the predictions in this study, the implications of the findings are a little bit confusing to decipher. On theone hand, the implications

seem to be that the choice of experts for SJT scoring judgment may not be as important as one might think. It may be possible for a variety of individuals to serve as subject matter experts and provide scoring judgments for an SJT without impacting construct measurement. On the other hand, it is hard to imagine that careful expert choice is not important. It is possible that all of the SMEs in this study were appropriate experts for the measurement of college student performance or that they were at least "equally good" experts at the measurement of college student performance. It may be that there are topics about which individuals and groups tend to share implicit theories of performance and that there are other types of topics about which groups differ in their theories of performance. The methodology of this study may have limited the findings because college student performance may fall in the former category. Practically, any incumbents with a certain level of experience (in this case, one year in college) may be able to serve as good experts for scoring judgments. It is still important, however, to determine what that necessary and sufficient level of experience is prior to selecting SMEs.

Appendix A

Expected Relationships between Scores Based on the Scoring Keys Developed by each SME Group and External Measures and Manipulation Check

External Measures	Graduate Students	Resident Advisors	Undergraduates
Neuroticism	+		
Conscientiousness	+	+	
GPA (Academic Performance)	+	+	
Family Support of Education (Manipulation check only)	+	+	
Parental Education (Manipulation check only)	+	+	
Leadership and Service-Related Behaviors (Leadership, Social Responsibility, Adaptability, Ethics and Integrity)		+	
Interpersonal Behaviors		+	+
Extraversion		+	+
Agreeableness		+	+

Appendix B

Coding Exercise for I/O Graduate Students

Please read each of the following situations and the responses to each situation carefully. After you read each response option, please indicate whether you think it would be selected by someone who strongly values academics, someone who strongly values leadership and service, someone who strongly values their social life, or whether you cannot determine who would select that item based on the above categories.

Please try to select one of the categories for each response option and only rate it as undetermined if you absolutely cannot decide. If you think that a response option would be selected by people from more than one of the above categories, you may rate it as such. However, as much as possible, please select only one category for each response option.

After you have selected a category for each response option, please rate your confidence in your rating of that response option on the following scale: 5 = fully confident, 4 = somewhat confident, 3 = neither confident nor lacking confidence, 2 = somewhat lacking in confidence, 1 = completely lacking confidence.

Thank you!

Example item:

You only have so much time in a day. What do you spend the most time on?

	Academic	Leadership and Service	Social Life	Undetermined	Confidence
a. Doing your homework.	X				5
b. Serving food at the local soup kitchen.		X			4
c. Hanging out with your friends.			X		5
d. Playing Nintendo by yourself.				X	3

Appendix C

Situational Judgment Test

Instructions: We asked a large group of college students to describe situations that they have faced in college; they then explained how they dealt with those situations. Each question that follows reflects one of those situations, along with a list of alternative ways they said they would respond to the situation. Please read each situation and then read all of the alternatives presented. Then, indicate which way you think you would MOST LIKELY respond. It might not be exactly what you would do in the situation, but it should be the alternative that comes closest to what you think you would actually do. Next, decide which alternative you would be LEAST LIKELY to take in the situation, and record your answer. After you have selected the responses you would be most and least likely to make, please rate the effectiveness of each of the response choices for the situation on a scale from 1 (very ineffective) to 5 (very effective).

1. After a local disaster, the Red Cross asked for volunteer blood donors. Because of a medical condition, you cannot donate blood. How would you react in this situation?

a. Encourage others to donate blood.
b. Donate money to the Red Cross instead.
c. Volunteer your time to generate money for the Red Cross.
d. Volunteer to give out cookies and help at the blood drives.
e. Ask the Red Cross if you could help them in any other way.

2. You find that you are eating more fattening and greasy food than normal, and that you have not been getting sufficient exercise. You have gained 15 pounds but find it difficult to change your eating and exercising habits. How would you deal with this situation?

a. Start slowly by cutting out two snacks a day.
b. Don't worry about it. You only live once, so eat what you want.
c. Get help from someone with experience in this area, such as a health professional or nutritionist.
d. Get some friends together and exercise together. There is power in numbers.
e. Try to establish a regular exercise routine, and focus on eating healthy foods.

3. You are part of a three-person group working on a class project with a quickly approaching deadline. One member of the team is not pulling his/her weight and avoids assignments, complains about the amount of work that has to be done, and says the project doesn't really matter anyway. While you are all classmates, you seem to be the group leader. What would you do?

a. Divide the workload evenly among members of the group, making sure everyone knows they are responsible for their share. If the group member still does not pull his/her own weight, bring it up with the instructor.

b. Speak with the group member and offer him/her encouragement to complete their portion of the project. If the group member still does not contribute, bring it up with the instructor.
c. Try to get the team member motivated to do a share of the work. If that doesn't help the situation, just put more effort into the project yourself in order to complete it.
d. Just do the group member's portion of the assignment in addition to your own, and tell the instructor about the situation.
e. See if the person could be removed from your group.
f. Consult with the group member who is not a problem about the most appropriate course of action, and then act on whatever you both decide.

4. You have very much wanted to be a teacher, but you failed the entrance exam into the College of Education. This exam is not given again for a year. What would you do?

a. Change majors to something similar that does not require an entrance exam.
b. Take a year off to earn some money, and then retake the exam.
c. Take additional relevant classes, and seek advice on how to best prepare for the examination the next year.
d. Take other requirements or courses of interest to you for a year, and then retake the examination next year.

5. A fellow student allows you to listen to threatening phone calls that have been placed on his/her answering machine by another student. The student does not want you to tell anyone but thinks the caller may be capable of causing physical harm. What would you do?

a. Try to talk the friend into calling the police and warn him/her to not walk around alone.
b. Talk to the resident assistant about it.
c. Contact the police yourself if you think there is any real threat of physical harm.
d. Find out who is making the calls; if it is another student, confront him/her – singly or jointly.
e. Unless the friend knows something that he/she is not saying, there is no reason *not* to call the police – so call them if your friend won't.
f. Have the friend change his/her phone number, and have it unlisted.

6. You have been standing in line for the restroom for some time after a campus event, and someone cuts into the line ahead of you. What would you do?

a. Politely inform the person that there is a line and hopefully he/she will move to the back.
b. Say aloud to someone near you how rude it is that people cut in line.
c. Give the person dirty looks, and try to squeeze him/her out of line.
d. Scold the person for not respecting others.
e. Be annoyed but not do anything. It's just one more person.

f. Calmly cut in line back in front of them.

7. You are interested in finance, but do not have further finance courses for at least another semester. What would you do?

a. Wait until the next semester, and take another class then.
b. Try to register for an alternative finance course as an elective.
c. Use the semester to do some independent study so that you are well prepared for the next course.
d. Get involved in on-campus finance clubs or investment games.
e. See if you could be a teacher's assistant for a finance class.

8. As a leader of a student organization, you asked a committee member to track the use of important and costly supplies. In response, he/she developed forms requiring the organization's committee members to indicate when and how they used various supplies. This committee member then complains that no committee members are completing these forms. How would you handle this situation?

a. Explain the importance of tracking to the committee, and request that everyone comply with the request.
b. Ask everyone to respect the committee member's hard work and effort by cooperating.
c. Limit access to the supplies until people start filling out the forms, or have penalties for not complying.
d. Designate another committee member to be in charge of tracking and enforcing the information requests.
e. Ask the committee if there is a misunderstanding about the forms and for suggestions on improving them.

9. Your roommate, usually a tidy person, has recently experienced some personal difficulties, thus becoming quite distracted and leaving much of the household responsibilities to you. You have discussed your concerns, and empathetically requested that he/she resume sharing in the responsibilities as soon as possible. A month passes, and you are still doing too much of the housework. What would you do?

a. Find out more about his/her problem and try to deal with that first.
b. Stop doing all of the household responsibilities to show him/her what it's like.
c. Talk with him/her again, and explain that you are suffering as a result of his/her behavior.
d. Tell him/her that if he/she doesn't help, you will move out.
e. Do your share of the work, and put anything of your roommate's that affects you in his/her area of the room.

10. After you arrive on campus, you begin to socialize with a group of students who drink regularly, even though all are underage. By the end of the term, you realize that you are

drinking several drinks at least three nights a week, but you don't know how to withdraw from the group in which this is normal routine behavior. What action would you take?

a. Ask a close friend to help watch out for your best interests, and pursue other activities with other people.
b. As long as you keep your grades up, it is not a problem.
c. Explain to the group that you are concerned about falling behind if you continue the behavior, and concentrate more on your studies instead.
d. Join alternative groups such as campus clubs and sports, or maybe even take an evening or early morning job.
e. Just socialize with the group less frequently.
f. Continue socializing with the group, but don't always drink when they do.

11. You have been having trouble with a class in which everyone else seems to be doing well. Your homework comes back with unsatisfactory grades week after week, and your test scores have been marginally passing. How would you proceed?

a. Find a study group to work with you.
b. Talk to the professor and to friends in the class, and read more.
c. Get tutoring, and study more frequently for this class.
d. Seek help from someone in the class who is doing well.
e. Talk to the professor or TA to find out what you are doing wrong, compare notes with others and seek out tutoring.
f. Stay calm and continue to do the best you can.

12. There is a seminar being held on campus that would expand your understanding of a class topic, but the seminar time conflicts with the class schedule. What would you do?

a. Skip the class, and go to the seminar because it is related to the class.
b. Go to class because it might cover what the seminar would cover.
c. Go to class and talk to someone that went to the seminar.
d. Get advice from the professor and then decide what to do.

13. You are the student coordinator for the gym, and it's 4:30 P.M. You have just been informed that there is no heat in the gym. As it is the middle of winter and very cold, you know this will be a problem. There is a student dance being held in the gym that night at 7:00 P.M., and there are no alternative facilities in which to hold the number of people expected at this event. What would you do?

a. Let everyone know that the dance is postponed or called off.
b. Call maintenance, and see if they can fix it.
c. Look for small heaters to fill the room.
d. Call people and check the consensus opinion about what to do.
e. Find a group of rooms as an alternative location.
f. Inform the students to dress warmly for the dance.

14. You and five other students must have a report ready within 48 hours. The last time the six of you worked together, you became the leader. You know that one of the group members did no work whatsoever on the last occasion, yet she is in your group again. This time it is necessary that all members pull their own weight. What would you do?

a. Let her know that you are aware that she did not do any work last time, and that this time it is necessary that she fully contribute.
b. Do your entire end of the work and ensure that the instructor is aware that you did your share, regardless of what the other members do.
c. Explain to the group that the professor will be made aware of who contributed what to the project, and ensure that this happens.
d. Stress the importance that everyone fully contributes his or her share to the project.
e. Work as closely with her as possible (e.g. assign both of you a related task) so as to offer encouragement and to ensure that her work gets done.
f. Assign her a specific task with a specific timeframe. If she does not do the work, ask to have her re-assigned, and have the group pick up her work.

15. You are working together with other classmates on a project. Your group keeps running into a variety of problems that threaten to cause the project to be late. The other group members want to just plan to submit it late. Another option would be to devote much more time than planned to the project and possibly get it in on time. What would you do?

a. Try to get it done, but plan to submit it late.
b. Ask the instructor for help or for an extension. If that doesn't work, just try your best and do what you can or turn it in late.
c. Motivate the group to devote more time and work together to get it done.
d. Have the group decide what to do.
e. Work hard to finish it because there are consequences for being late, and meeting deadlines is important to you.
f. Tell the instructor your situation, and ask for advice.

16. You grew up in a small farming community and moved into a dorm area in which all students were from an urban background. They seem to have different concerns and interests, and they often just stare blankly when you talk about your background and experiences. How would you react?

a. Ask them questions about their experiences in the hopes that they will develop some interest in your background.
b. Find other places to make friends with people who also come from farming communities.
c. Try to talk to just one person, on his/her own, about what life was like for you growing up.
d. Ask others about their experiences and ask if they have any questions of you.
e. Voice your feelings about the staring, and limit the talking about your background.

17. You have set ideas about what music is pleasing to the ear, and a friend is pushing you to join his/her at a concert that she thinks you would enjoy. The band will be playing a type of music that you prefer to avoid. What would you do?

a. You would not go, but you would decline as politely as possible.
b. If the ticket is free, you would go; otherwise you would not attend.
c. You would go to the concert with an open mind, hoping that you might appreciate it.
d. You would go because of your friend.

18. You know that a group of students in your class cheats on exams by putting formulas into scientific calculators or into cell phones. The professor has clearly warned against such activity, but you are not sure what she would do if she knew what these students were doing. What action would you take?

a. Try doing the same thing until people start getting caught.
b. Study the way you know best, don't cheat, but don't turn in the other students either.
c. You would do nothing; it's none of your business.
d. You would mention it to the professor so she can deal with the problems in the class.
e. Don't tell the professor, but make sure it is clear you are not involved in case they get caught.
f. Send the professor an anonymous message about what is going on.

19. You and your friends know that an attractive mutual friend has been dating another person for nearly a year. However, one of your friends tries his/her best to get a date with this individual. How would you react?

a. If the acquaintance is in a happy relationship, tell your friend to wait and rethink it. If not, it is ok to get a date.
b. Support my friend.
c. Tell your friend to wait until the person is single or to just forget about the person.
d. Tell your friend that it is inappropriate to interfere in the relationship.
e. Just be annoyed at your friend. Do not get involved.

20. When you first started school, you planned to major in an area in which you are no longer interested, and now your grades are not as good as you would like. You know that you do not want to major in this subject. What would you do?

a. Explore other options, and try to change your major to something you like.
b. Take classes or ask friends about other majors.
c. Change majors if isn't a huge setback. Otherwise, make the best of it.
d. Ask your advisor if the major has more interesting classes that you haven't taken yet.

21. Because of family problems, you realize that your parents can no longer support you financially at the same level as they have, and you do not have enough money to continue in school. What plans would you make?

a. Apply for student financial aid or get a part-time job.
b. Ask other family members for money to finish school.
c. Drop out of school and save money for going back.
d. Take fewer classes because of the lower level of finances.

22. An event in the news makes you wonder about the history behind it. What would you do?

a. Do some research, looking up all the facts for yourself.
b. Do a quick Internet search to see if you could find any information.
c. Think about it briefly, then move on.
d. Ask others what they know about the topic.
e. Resolve to read the newspaper more often.

23. You are finding a particular class dull and boring, and you are having difficulty staying awake. What would you do?

a. Do what you can to stay awake, such as drinking caffeine or sitting toward the front of the class.
b. Read the class material beforehand to make the lecture more interesting.
c. During the lecture, do some studying that is required for the course.
d. Make sure you are getting enough sleep every school night.
e. Skip the class if it is that dull and boring to you.

24. In the summer and fall, you walked to class and participated in various outdoor sports. When cold weather came, you took the bus and no longer participated in sports. You find that you are gaining weight. What action would you take?

a. Participate in indoor sports and start working out indoors.
b. Try not to eat as much, or eat different kinds of food.
c. Walk to classes more, go to the gym and watch what you eat.
d. Work out in your room.
e. Talk to an expert in diets and see if you can find someone who will encourage you to start working out again.
f. Not relevant due to physical disability.

25. One of your friends tells a joke that makes fun of people of a particular ethnic background. What would you do?

a. Laugh if it is funny and no one from that group is present.
b. Leave the room.

c. Nothing. Probably laugh if it is funny; it is just a joke.
d. Point out the offensiveness of the remark to the friend, and indicate your lack of tolerance for similar remarks.
e. Laugh if it is funny, but warn him/her to be careful in the future about where the joke is told.
f. Do not laugh; show displeasure by ignoring the joke.

26. Your grade for a particular class is based on three exams, with no class attendance requirement. All of the homework requirements for the class are posted on the professor's web site. What would you do?

a. Attend class for as long as you feel that it is helping your grades.
b. Do all the homework, but only go to some of the lectures. It's the exams that count.
c. Go to all the classes anyway. The professor may say something important.
d. Skip classes, but if you did poorly on the first exam, start going to classes.
e. There is no need to go to classes. Just get the homework done, and pass the exams.

27. There is a concert coming up that you think will be fantastic, but no one you know is interested in going with you. What would you do?

a. Go by yourself and find someone else at the concert that went alone.
b. Try to find someone else to go with you, but if you cannot then you would not go.
c. Ask your best friend to go even if you knew that he/she wasn't as excited as you were.
d. Get two tickets and offer a free ticket to anyone you know that might want to go.

28. You share a dorm room with three other students. One half-hour before you are expecting a guest, you get home to find the place completely trashed. There is no sign of any of your roommates. What would you do?

a. Clean up the mess as much as possible before the guest arrives. Then speak with your roommates immediately upon their return, so your guest knows how concerned you were about the mess.
b. Leave the mess and explain the situation to your guest.
c. Leave the mess and take the guest somewhere else.
d. Clean up the mess as much as possible before the guest arrives. Then, without the guest around, ask the roommates why the place was trashed so badly and what can be done in the future to avoid this situation.

29. One of your friends' roommates frequently parties until late at night, often returning to the room after drinking, engaging in loud and obnoxious behavior. Your friend finds that he/she cannot study or sleep well because of this, but also feels reluctant or afraid to talk with the resident assistant about this. What action would you take?

a. Approach the resident assistant on behalf of your friend.

b. Talk to the roommate yourself, and explain that his/her behavior bothers your friend.
c. Tell your friend to talk with the roommate and let him/her know that the behavior is not acceptable.
d. Offer to let your friend stay with you when necessary.
e. Suggest to your friend that he/she talk it out with the roommate, and offer to be available as a neutral third party when the two have the conversation.

30. You are searching for a major that interests you and think you might be interested in psychology. You do not know much about preparation to be a psychologist or what kinds of opportunities exist for careers in this area. What action would you take?

a. Talk to an advisor in psychology to see what career options are available.
b. Talk with a friend who is a psychology major to see what it is about.
c. Take an introductory psychology course to see what areas in psychology there are.
d. Look up job listings for psychologists on the Internet.

31. You are interested in several different classes/disciplines, but don't know anything about future educational or career opportunities in these areas. What steps would you take to get informed?

a. Go to an advisor or knowledgeable professional who might tell you more and be able to answer your questions.
b. Research topics using available resources like relevant books and Internet web sites.
c. Attempt to obtain some hands-on experience, like internships.
d. Use the school's resources such as career services and career counselors.
e. Take some introductory classes in the area of interest to see if you want to pursue that area further.
f. Think about your interests and try to figure out which of them fit with the different disciplines.
g. Ask friends and family for advice and information. If possible, ask a friend who is familiar with the area.

32. In a class of 50 students, you discover that a group of your friends have worked out a scheme to share answers on an exam. The professor has vision problems and will likely never notice. You are not doing very well in the course. What would you do in these circumstances?

a. Avoid being around these friends.
b. It is not exactly honest but under the circumstances, the scheme is OK. You would join them.
c. Do your own work, and do not tell the professor about the scheme because it is not your problem.
d. Cheat and get a good grade.
e. Tell the professor about the scheme.

f. Study for the exam, but join the scheme as a backup strategy for the test.

33. You see a painting that intrigues you. You know nothing about it other than the artist's name. What would you do?

a. Look up the artist on the Internet to see if you can find some of his/her other work.
b. Ask others if they know anything about the artist.
c. Do some research to find out what you want to know.
d. Look for help at the library, asking for books about this artist.
e. Enjoy the painting, but leave it at that.

34. Your professor has just given you a project that will obviously require the whole semester to complete. She gave you all the details you need to get started, but you are not sure how the project should proceed from there. She does not appear to intend to give you any more information in class. What would you do?

a. Work out the project to the best of your ability, and approach the professor if you get stuck.
b. Generate some ideas, and then go to office hours to see how the professor responds to them.
c. Ask the professor about the project after class.
d. Visit the professor or a teaching assistant during office hours to discuss the project.
e. Talk to other students to get an idea of what they are doing.
f. Try to get an idea of whether or not other students seem confused. If so, bring the issue up with the professor during class.

35. You are part of a committee to reduce cross-cultural tension in your dorm. A group of students in your dorm complain to you that people always convey holiday greetings to them that are not associated with their religion or culture. They request that their differences be respected. How would you address this problem?

a. Ask the group politely to ignore the greetings, realizing that the people had good intentions.
b. Tell the well-wishers respectfully to please refrain from making specific holiday greetings.
c. Have a meeting at which people can discuss their differences and hopefully work out an understanding.
d. As part of the committee, make all cultural holidays visible so that people can be aware of diversity.
e. Tell them to respond with a meaningful greeting of their own.

36. A friend on your floor is always organizing "social" activities – including trips to local bars. Aside from the fact that this person is underage and failing some classes, you realize that the individual is drinking half a dozen or more drinks at least three or four times a week. No one else seems to know or to be concerned about the person. What would you do?

a. Talk to him/her about easing up on the alcohol, explaining that it will not help with classes, which should be the main reason for being in college.
b. Use humor to broach the topic and offer alternatives to this usual "social" activity.
c. Bring up the situation with the floor's resident assistant.
d. Try to get him/her involved in other activities.
e. Talk to the person to determine subtly if there are other issues that need to be addressed, and refer him/her to help if appropriate.
f. Talk to other people on the floor, and discuss ways to address the situation.
g. Ask once about this behavior and see where the discussion leads, then leave him/her to pursue his/her own course of action.

Appendix D

Behaviorally Anchored Rating Scales

The following questions ask you to rate your skills in 12 different areas during the PAST SIX MONTHS. Read the definition of each are, then use the behavioral examples provided to help you rate yourself most accurately.

Knowledge and mastery of general principles is defined as: Gaining knowledge and mastering facts, ideas and theories and how they interrelate, and the relevant contexts in which knowledge is developed and applied.

1. Very low (for example: rarely studying for tests, slacking off on assignments)
2.
3. Average (for example: sometimes studying for tests, putting some effort into assignments)
4.
5. Very high (for example: studying hard for tests, putting a great deal of effort into assignments)

Continuous learning is defined as: Being intellectually curious and interested in continuous learning. Actively seeking new ideas and new skills, both in core areas of study as well as in peripheral or novel areas.

1. Very low (for example: only learning the minimum amount required for class, rarely searching out information on topics that interest you on the internet or at the library)
2.
3. Average (for example: sometimes learning a little more than what is required for courses on class topics that interest you, occasionally searching out interesting topics on the internet or at the library)
4.
5. Very high (for example: frequently learning extra information beyond what is covered in classes on topics that interest you, often searching out interesting topics on the internet or at the library)

Leadership is defined as: Demonstrating skills in a group, such as motivating others, coordinating groups and tasks, serving as a representative for the group, or otherwise performing a managing role in a group.

1. Very low (for example: avoids being in charge of group projects, always waiting for others to assign work to you on group tasks)
2.

3. Average (for example: sometimes coordinating group tasks or activities when asked, speaking up in groups when you have an idea about the direction the group should go)
4.
5. Very high (for example: often taking charge in group activities, motivating others in groups, representing the groups that you're involved in to others)

Interpersonal skills is defined as: Communicating and dealing well with others, whether in informal social situations or more formal school-related situations. Being aware of the social dynamics of a situation and responding appropriately.

1. Very low (for example: picking fights with other people, keeping thoughts/or feelings bottled up, letting emotions explode, saying inappropriate things)
2.
3. Average (for example: usually expressing thoughts and feelings effectively, thinking about what situation you're in and what type of behavior is appropriate, usually remaining calm when interacting with others)
4.
5. Very high (for example: almost always clearly and calmly expressing thoughts and feelings, listening carefully to others and responding appropriately)

Social responsibility is defined as: Being responsible to society and the community, and demonstrating good citizenship. Being actively involved in the events in one's surrounding community, which can be at the neighborhood, town/city, state, national, or college/university level. Activities may include volunteer work for the community, attending city council meetings, and voting.

1. Very low (for example: very rarely voting, rarely participating in community activities or volunteer work, littering)
2.
3. Average (for example: voting in major elections, occasionally participating in community activities, signing petitions)
4.
5. Very high (for example: voting in all major and local elections, actively participating in community activities, helping out neighbors and other community members)

Adaptability and life skills is defined as: Adapting to a changing environment (at school or home), dealing well with gradual or sudden and expected or unexpected changes. Being effective in planning one's everyday activities and dealing with novel problems and challenges in life.

1. Very low (for example: frequently getting upset when unexpected events force you to change your plans, rarely leaving extra time in your schedule in case things don't go according to plan)
2.

3. Average (for example: sometimes getting upset when unexpected events force you to change your plans, sometimes leaving a little bit of extra time in your schedule in case things don't go according to plan)
4.
5. Very high (for example: rarely getting upset when unexpected events force you to change your plans, almost always leaving enough time to get everything done even if things don't go according to plan)

Ethics and integrity is defined as: Having a well-developed set of values, and behaving in ways consistent with those values. In everyday life, this probably means being honest, not cheating (on exams or in committed relationships), and having respect for others.

1. Very low (for example: cheating on exams, frequently telling lies, worrying very little about being an ethical person)
2.
3. Average (for example: usually acting honestly, but sometimes telling lies, might consider cheating on an exam under certain circumstances)
4.
5. Very high (for example: almost always behaving honestly, never cheating on an exam, never unfaithful a significant other)

Appendix E

Behavioral Observation Scales

The following items ask about various experiences you may have had during college. You will be asked to estimate the number of times you had each experience. For some questions, it might be difficult to remember exactly how many times, so take your best guess. It might help to think of examples of times when you had each experience. (Please note that the items were not collected in the order as shown. The items were collected across scales based on the response scale appropriate for each item.)

Please rate how many times during college you…

(Knowledge and Mastery of General Principles)

1. Were on the dean's list
    A. 0 times
    B. 1-2 times
    C. 3-4 times
    D. 5-6 times
    E. 7-8 times
    F. 9-10 times
    G. more than 10 times

2. Were invited to be part of a research group
    A. 0 times
    B. 1-2 times
    C. 3-4 times
    D. 5-6 times
    E. 7-8 times
    F. 9-10 times
    G. more than 10 times

3. Were invited to join an honor society
    A. 0 times
    B. 1-2 times
    C. 3-4 times
    D. 5-6 times
    E. 7-8 times
    F. 9-10 times
    G. more than 10 times

4. Were recognized publicly by a professor for your class work
    A. 0 times
    B. 1-4 times
    C. 5-9 times
    D. 10-14 times

E. 15-19 times
F. 20-24 times
G. more than 24 times

5. Won or maintained a competitive academic scholarship
   A. 0 times
   B. 1-4 times
   C. 5-9 times
   D. 10-14 times
   E. 15-19 times
   F. 20-24 times
   G. more than 24 times

6. Were paid to tutor a classmate in a course
   A. 0 times
   B. 1-4 times
   C. 5-9 times
   D. 10-14 times
   E. 15-19 times
   F. 20-24 times
   G. more than 24 times

7. Won an award for an academic project
   A. 0 times
   B. 1-4 times
   C. 5-9 times
   D. 10-14 times
   E. 15-19 times
   F. 20-24 times
   G. more than 24 times

8. Participated in an academic competition
   A. 0 times
   B. 1-10 times
   C. 11-20 times
   D. 21-30 times
   E. 31-40 times
   F. 41-50 times
   G. more than 50 times

(Continuous Learning)
1. Enrolled in a class outside of school to learn more about a subject you were interested in
   A. 0 times
   B. 1-2 times
   C. 3-4 times

D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

2. Conducted an experiment not required by class
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

3. Attended a lecture or talk not required or rewarded by classes
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

4. Offered information to a teacher that went beyond the information in the course textbook
A. 0 times
B. 1-10 times
C. 11-20 times
D. 21-30 times
E. 31-40 times
F. 41-50 times
G. more than 50 times

5. Read ahead in a class textbook because you were interested in the subject
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

6. Read an educational or scientific magazine
A. Never
B. Less than once per year
C. At least once per year

D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

7. Devoted a regular practice time to develop a skill or a better understanding of something that interests you
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

8. Researched (e.g., getting a book or looking on the internet) and learned more information about a topic or question that you found interesting
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

9. Read a book or article related to something you found interesting
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

(Leadership)
1. Worked as a formal representative for your college or university
A. 0 times
B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

2. Were the team captain or leader for an official school or club sports team
A. 0 times

B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

3. Started a new club, organization, or other official group
A. 0 times
B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

4. Were appointed or elected officer in a club, professional society, or other organized interest group
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

5. Organized a community event (e.g., a walkathon, a neighborhood picnic, a voter registration drive)
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

6. Encouraged non-participating members of a group to be more active
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

7. Acted as the leader of a team for a class project

A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

8. Delegated tasks to a group of people
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

(Interpersonal skills)
1. Were told by a friend to stop saying something offensive or embarrassing
A. 0 times
B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

2. Helped other people resolve a dispute
A. 0 times
B. 1-10 times
C. 11-20 times
D. 21-30 times
E. 31-40 times
F. 41-50 times
G. more than 50 times

3. Did or said something that seriously offended someone
A. 0 times
B. 1-10 times
C. 11-20 times
D. 21-30 times
E. 31-40 times
F. 41-50 times
G. more than 50 times

4. Hosted a party or large social gathering

A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

5. Comforted a friend who was upset
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

6. Made "small talk" with someone you didn't know very well
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

7. Introduced yourself to others at a party or social gathering
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

(Social Responsibility)
1. Voted in a school election
A. 0 times
B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

2. Voted in a local or national election

A. 0 times
B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

3. Participated in a protest or demonstration
A. 0 times
B. 1-2 times
C. 3-4 times
D. 5-6 times
E. 7-8 times
F. 9-10 times
G. more than 10 times

4. Signed a petition
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

5. Were a member of a community outreach organization (e.g., Boy/Girl Scouts or Big Brother/Sister)
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

6. Participated as a member of an official political organization
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

7. Donated money or items to a charity organization

A. 0 times
B. 1-10 times
C. 11-20 times
D. 21-30 times
E. 31-40 times
F. 41-50 times
G. more than 50 times

8. Organized or participated in a community event
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

9. Were involved in volunteer work
A. Never
B. Less than once per year
C. At least once per year
D. At least once per semester
E. At least once per month
F. At least once per week
G. Almost every day

(Adaptability)
1. Missed deadlines for class projects, work, or other important obligations
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

2. Were late paying a bill
A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

3. Were late to a class, meeting, or other appointment

A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

4. Asked for an extension on an assignment because you didn't leave enough time to finish

A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

5. Had to back out of prior meetings or responsibilities because you had trouble managing all of your responsibilities

A. 0 times
B. 1-4 times
C. 5-9 times
D. 10-14 times
E. 15-19 times
F. 20-24 times
G. more than 24 times

6. Went to class unprepared when you could have completed work or readings

A. 0 times
B. 1-10 times
C. 11-20 times
D. 21-30 times
E. 31-40 times
F. 41-50 times
G. more than 50 times

7. Produced a poor product for class (e.g. paper, presentation) because you did not start working on it early enough

A. 0 times
B. 1-10 times
C. 11-20 times
D. 21-30 times
E. 31-40 times
F. 41-50 times
G. more than 50 times

(Ethics and Integrity)

1. Lied on a formal document (e.g., school form, work application)
    A. 0 times
    B. 1-2 times
    C. 3-4 times
    D. 5-6 times
    E. 7-8 times
    F. 9-10 times
    G. more than 10 times

2. Received a warning from a landlord or were evicted from an apartment
    A. 0 times
    B. 1-2 times
    C. 3-4 times
    D. 5-6 times
    E. 7-8 times
    F. 9-10 times
    G. more than 10 times

3. Were arrested for a misdemeanor or received a citation
    A. 0 times
    B. 1-4 times
    C. 5-9 times
    D. 10-14 times
    E. 15-19 times
    F. 20-24 times
    G. more than 24 times

4. Were investigated by the campus judicial advisory board
    A. 0 times
    B. 1-4 times
    C. 5-9 times
    D. 10-14 times
    E. 15-19 times
    F. 20-24 times
    G. more than 24 times

5. Were issued a parking or speeding ticket
    A. 0 times
    B. 1-10 times
    C. 11-20 times
    D. 21-30 times
    E. 31-40 times
    F. 41-50 times

G. more than 50 times

6. Used a fake ID
   A. Never
   B. Less than once per year
   C. At least once per year
   D. At least once per semester
   E. At least once per month
   F. At least once per week
   G. Almost every day

7. Stole something or borrowed something without permission
   A. Never
   B. Less than once per year
   C. At least once per year
   D. At least once per semester
   E. At least once per month
   F. At least once per week
   G. Almost every day

8. Cheated on an exam, test, or classwork
   A. Never
   B. Less than once per year
   C. At least once per year
   D. At least once per semester
   E. At least once per month
   F. At least once per week
   G. Almost every day

9. Lied to someone to cover up something you did
   A. Never
   B. Less than once per year
   C. At least once per year
   D. At least once per semester
   E. At least once per month
   F. At least once per week
   G. Almost every day

Appendix F

IPIP Sub-scales

On the following pages, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes *you*. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Please read each statement carefully, and then fill in the bubble that corresponds to the number on the scale.

Response Options

1: Very Inaccurate
2: Moderately Inaccurate
3: Neither Inaccurate nor Accurate
4: Moderately Accurate
5: Very Accurate

(Conscientiousness)
- Am always prepared.
- Pay attention to details.
- Get chores done right away.
- Like order.
- Follow a schedule.
- Am exacting in my work.
- Leave my belongings around.
- Make a mess of things.
- Often forget to put things back in their proper place.
- Shirk my duties.

(Neuroticism)
- Am relaxed most of the time.
- Seldom feel blue.
- Get stressed out easily.
- Worry about things.
- Am easily disturbed.
- Get upset easily.
- Change my mood a lot.
- Have frequent mood swings.
- Get irritated easily.
- Often feel blue.

(Extraversion)
- Am the life of the party.

Feel comfortable around people.
Start conversations.
Talk to a lot of different people at parties.
Don't mind being the center of attention.
Don't talk a lot.
Keep in the background.
Have little to say.
Don't like to draw attention to myself.
Am quiet around strangers.

(Agreeableness)
Am interested in people.
Sympathize with others' feelings.
Have a soft heart.
Take time out for others.
Feel others' emotions.
Make people feel at ease.
Am not really interested in others.
Insult people.
Am not interested in other people's problems.
Feel little concern for others.

Appendix G

Rating Task and Demographics for SMEs

Please rate how important you believe the following domains are to success as a college student:

a. Very unimportant
b. Unimportant
c. Somewhat unimportant
d. Neither unimportant nor important
e. Somewhat important
f. Important
g. Very important

2. Knowledge and mastery of general principles is defined as gaining knowledge and mastering facts, ideas and theories and how they interrelate, and the relevant contexts in which knowledge is developed and applied. For example, studying for tests and putting effort into assignments.

3. Continuous learning is defined as: Being intellectually curious and interested in continuous learning. Actively seeking new ideas and new skills, both in core areas of study as well as in peripheral or novel areas. For example, learning more than what is required for courses, searching out interesting topics on the internet or at the library.

4. Artistic and cultural appreciation is defined as: Appreciating art and culture, either at an expert level or simply at the level of one who is interested. For example, attending plays, musical performances, art galleries or other artistic events, trying to learn about art and culture.

5. Appreciation for diversity is defined as: Showing openness, tolerance, and interest in a diversity of individuals and groups (e.g., by culture, ethnicity, religion, or gender). Actively participating in, contributing to, and influencing a heterogeneous environment. For example, speaking in a "politically correct" way, actively trying to learn about people from other cultures or groups, going to events sponsored by different cultural groups.

6. Leadership is defined as: Demonstrating skills in a group, such as motivating others, coordinating groups and tasks, serving as a representative for the group, or otherwise performing a managing role in a group.

7. Interpersonal skills is defined as: Communicating and dealing well with others, whether in informal social situations or more formal school-related situations. Being aware of the social dynamics of a situation and responding appropriately.

8. Social responsibility is defined as: Being responsible to society and the community, and demonstrating good citizenship. Being actively involved in the events in one's surrounding community, which can be at the neighborhood, town/city, state, national, or college/university level. Activities may include volunteer work for the community, attending city council meetings, and voting.

9. Physical and Psychological Health is defined as: Possessing the physical and psychological health required to engage actively in a scholastic environment. This would include participating in healthy behaviors, such as eating properly, exercising regularly, and maintaining healthy personal and academic relations with others, as well as avoiding unhealthy behaviors, such as alcohol/drug abuse, unprotected sex, and ineffective or counterproductive coping behaviors.

10. Career orientation is defined as: Having a clear sense of career one aspires to enter into, which may happen before entry into college, or at any time while in college. Establishing, prioritizing, and following a set of general and specific career-related goals.

11. Adaptability and life skills is defined as: Adapting to a changing environment (at school or home), dealing well with gradual or sudden and expected or unexpected changes. Being effective in planning one's everyday activities and dealing with novel problems and challenges in life.

12. Perseverance is defined as: Committing oneself to goals and priorities set, regardless of the difficulties that stand in the way. Goals range from long-term goals (e.g., graduating from college) to short-term goals (e.g., showing up for class every day even when the class isn't interesting).

13. Ethics and integrity is defined as: Having a well-developed set of values, and behaving in ways consistent with those values. In everyday life, this probably means being honest, not cheating (on exams or in committed relationships), and having respect for others.

How supportive is your family of your pursuits in higher education?

a. very unsupportive
b. unsupportive
c. somewhat unsupportive
d. neither unsupportive nor supportive
e. somewhat supportive
f. supportive
g. very supportive

What is the highest level of education attained by your mother?

a. below high school
b. high school diploma
c. two-year Associate's degree
d. four-year Bachelor's degree
e. Graduate degree

What is the highest level of education attained by your father?

a. below high school
b. high school diploma
c. two-year Associate's degree
d. four-year Bachelor's degree
e. Graduate degree

Appendix H

Instructions for Verbal Protocol Task

Practice Items for Verbal Protocol Participants

As you answer the following questions, please try say all of your thoughts out loud as you go. Please describe your thoughts, feelings and choices about what you are doing and reading. It might be your reactions, reasoning, or even something you are reminded of. Please don't censor your thoughts. Even if a thought does not seem relevant to the task, it is of interest to me. Remember, everything you say to me today will be kept confidential.

To help you get used to saying all of your thoughts out loud, we're going to go through two practice items. Once you feel comfortable with the think-aloud technique, then I will ask you to complete six additional items.

1. You are shopping when you notice a man robbing the store. What would you do?

a) Leave the store as quickly as possible and call the police.
b) Try to apprehend the robber yourself.
c) Follow the man and call the police as soon as he appears settled somewhere.
d) Nothing, as you do not wish to get involved in the matter.

2. Your professor recently passed out exams that your class took last week. Everyone except you was given an extra 5 points to make up for some errors the professor made in writing the test. What would you do?

a) Assume it was a mistake and speak to your professor.
b) Confront your professor regarding why are being treated unfairly.
c) Assume that the professor added 5 points to your score but forgot to indicate the score change on the test you received back.
d) Complain to the head of the department.
e) Drop the class.

Appendix I

Coding Scheme:
Reasons Why Someone Would or Would Not Choose a Response Option

The response would not be effective because...

1. It is too difficult or I don't have the qualities necessary to pull it off.
2. I've done this in the past and it did not work for me.
3. Something else would need to be added for this option to work.
4. It would not solve the problem.
5. It is too extreme.

The response might be effective, but it would be undesirable because...

6. It would take too much time or require too much effort.
7. I just don't like to do this or I am not willing to do this.
8. It is immature.
9. I don't have the right to do this.
10. It would be unfair to me or to others.
11. It is too confrontational or would offend other people.
12. It could be harmful to me (for example, it could backfire or overburden me).
13. I prefer to take responsibility myself rather than relying on others.
14. I prefer to seek others' opinions rather than make decisions unilaterally.

The response would be effective because...

15. It would work quickly or with little effort.
16. This is something I've done in the past and it worked for me.
17. This would be the kind or thoughtful option.
18. I think this would work well because of what I know about human nature (for example, setting clear goals helps motivate people).
19. Seeking help from an authority can be helpful.
20. I like this because it is fair.
21. It is what I am "supposed" to do (for example, it follows the rules).

Other reasons:

22. I wouldn't think of this on my own.
23. This option does not provide enough information for me to know if it would work.
24. Group problems and issues should be dealt with inside a group.
25. This option sounds like something I would do because of my personality.
26. I don't think this would work, but I know I would do it anyway.

References

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J.A. Weekley, & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 233-249). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Beauregard, R. S. (2000). *Construct explication of a situational judgment test: Addressing multidimensionality through item development, content analysis, and scoring procedures.* Unpublished doctoral dissertation, Wright State University, Dayton OH.

Cable, D. M., & Judge, T. A. (1996). Person-organization fit, job choice decisions, and organizational entry. *Organizational Behavior and Human Decision Processes, 67*, 294-311.

Cable, D. M., & Judge, T. A. (1997). Interviewers' perceptions of person-organization fit and organizational selection decisions. *Journal of Applied Psychology, 82*, 546-561.

Cable, D. M., & Parsons, C. K. (2001). Socialization tactics and person-organization fit. *Personnel Psychology, 54*, 1-23.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56*, 453-484.

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. (2006). Situational judgment tests: Method or construct? In J.A. Weekley, & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135-155). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Clevenger, J. P., & Haaland, D. E. (2000, April). *The relationship between job knowledge and situational judgment test performance.* Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

England, G. W. (1971). *Development and use of weighted application blanks.* (Bulletin No. 55). Minneapolis: University of Minnesota, Industrial Relations Center.

George, J. M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology, 75*, 107-116.

Glaser, R., Lesgold, A., & Gott, S. (1991). Implications of cognitive psychology for measuring in job performance. In A. K. Wigdor, B. F. Green, Jr. (eds.), *Performance Assessment for the Workplace* (Vol. 2, pp. 1-26). Washington, DC: National Academy Press.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, the Netherlands: Tilburg University Press.

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

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational threats to the use of SJTs: Faking, coaching, and retesting issues. In J.A. Weekley, & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 205-232). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

Jackson, S. E., Brett, J. F., Sessa, V. I., Cooper, D. M., Julin, J. A., & Peyronnin, K. (1991). Some differences make a difference: Individual dissimilarity and group heterogeneity as correlates of recruitment, promotions, and turnover. *Journal of Applied Psychology, 76*, 675-689.

Judge, T. A., & Cable, D. M. (1997). Applicant personality, organizational culture, and organization attraction. *Personnel Psychology, 50*, 359-394.

Lawrence, B. S. (1997). The black box of organizational demography. *Organization Science, 8*, 1-22.

Magnus, K., Diener, E., Fujita, F., & Pavot, W. (1993). Extraversion and neuroticism as predictors as objective life events: A longitudinal analysis. *Journal of Personality and Social Psychology, 65*, 1046-1053.

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

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

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

McKenzie, J. (1989). Neuroticism and academic achievement: The Furneaux factor. *Personality and Individual Differences, 10,* 509–515.

McKenzie, J., Taghavi-Knosary, M., & Tindell, G. (2000). Neuroticism and academic achievement: The Furneaux factor as a measure of academic rigour. P*ersonality and Individual Differences, 29,* 3–11.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology, 67*, 411-418.

Motowidlo, S. J., Diesch, A. C., & Jackson, H. L. (2003, April). *Using the situational judgment format to measure personality characteristics*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

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

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J.A. Weekley, & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57-81). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *American Psychologist, 61*, 622-631.

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

Pfeffer, J. (1983). Organizational demography. In L.L. Cummings & B.M. Staw (Eds.) *Research in organizational behavior* (pp. 299-357). Greenwich, CT: JAI Press.

Ployhart, R. E. (2006). The predictor response process model. In J.A. Weekley, & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 83-105). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Ployhart, R. E., & Erhart, 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*, 1-16.

Ployhart, R. E., & Ryan, A. M. (2000, April). *Integrating personality tests with situational judgment tests for the prediction of customer service performance.* Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Porr, W. B., & Ployhart, R. E. (2004, April). *The validity of empirically and construct-oriented situational judgment tests.* Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

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

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Sacco, J. M., Scheu, C., Ryan, A. M., Schmitt, N. W. (2000, April). *Understanding race differences on situational judgment tests using readability statistics.* Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Sacco, J. M., Schmidt, D. B., & Rogg, K. L. (2000, April). *Using readability statistics and reading comprehension scores to predict situational judgment test performance, black-white differences, and validity.* Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302-318.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437-454.

Thayer, P. W. (1977). Somethings old, somethings new. *Personnel Psychology, 30*, 513-524.

Trippe, M. D., & Foti, R. J. (2003, April). *An evaluation of the construct validity of situational judgment tests*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

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. Weekley, & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157-182). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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