2011

Student evaluation of teaching: Individual differences and bias effects

Verena Sylvia Bonitz
Iowa State University

Student evaluation of teaching: Individual differences and bias effects

by

Verena Sylvia Bonitz

A dissertation submitted to the graduate faculty

In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Psychology

Program of Study Committee:

Lisa Larson, Major Professor

Patrick Armstrong

Douglas Bonett

David Vogel

Donna Kienzler

Iowa State University

Ames, Iowa

2011

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The goal of this experimental study was to evaluate the influence of course type, instructor and student gender, and student individual differences (domain-specific vocational interests and confidence, personality, and gender role attitudes) on student evaluation of teaching (SET) scores. A sample of 610 college students (372 female) rated hypothetical instructors described in a vignette on eight common dimensions of teaching effectiveness. Mean SET ratings were not significantly different across instructor gender and course type. A series of multiple regressions revealed, however, that student individual differences explained a significant proportion of the variance in SET ratings. The most salient traits that were significantly related to SET ratings were agreeableness, conscientiousness, conventional and investigative confidence, and gender role attitudes. In addition, female students gave significantly higher mean ratings than male students independent of course type or instructor gender. This effect was eliminated when statistically controlling for students' individual differences. Overall, the findings of this study suggest that student individual differences can bias SET scores, which poses a threat to the validity of the ratings.

CHAPTER 1: INTRODUCTION

Student evaluation of teaching (SET) has become the most prevalent measure of teaching effectiveness across universities in the United States (e.g., Clayson, 2009; Pounder, 2008). The SET data are primarily used as diagnostic feedback tool for instructors (formative function), and as performance measures for personnel decisions such as hiring, tenure, and promotion (summative function) (Marsh & Dunkin, 1992). The use of SET data for these purposes is controversial, and both scholars and instructors alike have questioned the reliability and validity of the SET process (e.g., Aleamoni, 1987; Clayson, 2009; Pounder, 2008; Sproule, 2000). One frequent issue of contention concerns the susceptibility of SET scores to bias and manipulation (Crumbley & Reichelt, 2009; Pounder, 2007; W. M. Williams & Ceci, 1997). Therefore, the goal of the present experimental study was to conduct a systematic evaluation of a set of variables (course type, instructor and student gender, and student individual differences) for their biasing effect on SET ratings.

Bias in Student Evaluations of Teaching

A large proportion of SET research has been devoted to the issue of bias. Bias refers to the systematic introduction of irrelevant variables that may distort or disguise the relationship between variables (Heppner, Wampold, & Kivlighan, 2008; Messick, 1989). Bias therefore constitutes a threat to the validity of the measure, which can lead to unlawful discrimination and unfair evaluation practices. Research on SET bias has concentrated on a wide range of variables within the educational context, as well as the demographic characteristics of instructors and students. A multitude of variables have been shown to correlate with SET ratings, some of them yielding large and practically meaningful effects. However, it is not clear in most cases whether these variables are indeed instrumental in

student learning (which would justify an effect on SET scores), or whether they constitute bias. For example, consider the positive correlation between students' subject interest and SET ratings. This relation has been explained in two different ways (Marsh & Roche, 1997): The first explanation assumes that the instructor is an effective teacher who is able to elicit students' interest in the course material. Student interest in turn serves as a motivating force to engage in the material, to complete assignments, to attend classes and office hours, and to study for tests. All of these behaviors then lead to better learning outcomes. Thus, students reward the effective instructor with higher SET ratings, which reflects the amount they have learned in the course. In this case, student interest is not a biasing variable, but a valid indicator of the instructor's teaching ability.

The second explanation is based on the assumption that student interest is not related to the teaching effectiveness of the instructor, and therefore a bias variable. For example, it is possible that students who come into the course with an interest in the subject might transfer their liking of the content to the instructor (Halo effect). Students then express their liking of the content and the instructor through higher SET ratings irrespective of the actual teaching effectiveness of the instructor. In this case, instructors who teach popular subjects would have an unfair advantage over instructors who teach courses that most students are not inherently interested in.

Based on the available SET research, however, it is not possible to distinguish between several alternative explanations for the relations between such variables and SET ratings due to methodological limitations. The first limitation concerns the preponderance of non-experimental field research. Many studies involve the post-hoc analysis of existing SET data gathered in actual classrooms. Despite the high degree of external validity, the

inferences that can be drawn from such research are limited due to the lack of manipulation of independent variables, random assignment, and control of possible confounds.

The second issue that limits the interpretability of the available data on SET bias is the paucity of studies that have focused on the mechanisms that explain the observed correlations between various background variables and SET ratings. The inconsistency in findings across studies, both with regard to the direction and the size of the effects, suggests the presence of moderator and mediator variables that alter the relation between target variable and SET ratings. Most research has remained at the descriptive level, examining one variable at a time (a notable exception is the interaction between instructor and student gender). However, only the simultaneous evaluation of multiple variables allows for identification of moderating or mediating effects.

The Present Study

The purpose of the present study was to address the methodological shortcomings of prior SET research by using an experimental design to examine a set of variables simultaneously for their effect on SET ratings. The study was implemented as follows: Students were asked to rate a hypothetical instructor on eight different dimensions of teaching effectiveness. The use of hypothetical vignettes allowed controlling for variables that might otherwise confound the results. Further, by removing the instructor-student interaction and learning process, it became possible to separate the effect of response style/bias from potentially meaningful effects a variable has on learning. In other words, it was possible to differentiate between alternative explanations for the relation between the target variable and SET ratings. Since students did not actually interact with the instructor, it

could be assumed that any systematic differences in their responses to the vignettes were entirely due to response style and bias.

Two factors were manipulated across the instructor vignettes, namely type of course taught by the instructor (counseling vs. research methods) and instructor gender (male vs. female). Students were randomly assigned to one of the four experimental conditions; these were male instructor/counseling, female instructor/counseling, male instructor/research methods, and female instructor/research methods. A variety of student background variables were tested for their effects on the vignette ratings; these were student gender and gender role attitudes, broad and domain-specific vocational interests and confidence, and the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism). The following section summarizes any prior research on the effect of these variables on SET ratings, and gives the rationale for their inclusion in the present study.

Course Type

SET ratings have been shown to vary by academic discipline. On average, instructors in the arts and humanities tend to receive the highest ratings, followed by the social/biological sciences and business; instructors in computer science, engineering, and the physical sciences tend to obtain the lowest ratings (e.g., Basow & Montgomery, 2005; Cashin, 1990; Ory, 2001).

Based on these findings, two specialty areas within psychology (counseling psychology and research methods/statistics) were manipulated across the instructor vignettes used in the present study. The rationale for this choice was as follows: Psychology students tend to be most interested in content courses dealing with human relations, developmental psychology, and clinical/counseling psychology. Research methods and statistics courses on

the other hand are often viewed by students as boring, difficult, and irrelevant to their future career plans (Conners, McCown, & Roskos-Ewoldsen, 1998; Early, 2007; Manning, Zachar, Ray, & LoBello, 2006; Vittengl et al., 2004). In addition, students tend to delay or avoid enrolling in methods and statistics courses (Lauer, Rajecki, & Minke, 2006).

Gender and Gender Role Attitudes

There is some preliminary empirical evidence to support a link between instructor gender and SET ratings. However, the extent and the direction of this effect might depend on the specific pairing of instructor and student gender, and complex interpersonal dynamics based on gender role stereotypes. Therefore, instructor gender, student gender, and student gender role attitudes were all included as variables in the present study.

Instructor gender. The instructor variable that has probably received the most attention in the empirical literature is instructor gender. Research on the relation between instructor gender and SET scores has yielded mixed and complex findings: Some studies have found higher global SET scores for male instructors compared to female instructors (e.g., Sidanius & Crane, 1989; B. P. Smith, 2009), some have found the reverse pattern of results (e.g., Basow & Montgomery, 2005; Whitworth, Price, & Randall, 2002), while others have not found any systematic gender differences (e.g., Feldman, 1992, 1993; G. Smith & Anderson, 2005). However, more fine-grained analyses have revealed an interesting pattern. When female teachers received higher ratings than men, it was usually on dimensions that captured the interpersonal relations between instructor and students: Generally, women were praised for being caring, empathetic, approachable, and for fostering a good relational climate in the class room (e.g., Bachen, McLoughlin, & Garcia, 1999; Basow & Montgomery, 2005; Basow, Phelan, & Capotosto, 2006; Bennett, 1982; Centra & Gaubatz,

2000; Kierstead, D'Agostino, & Dill, 1988). Men, on the other hand received higher ratings on dimensions such as course planning, competence, knowledge, and organization skills (e.g., Basow et al., 2006; B. P. Smith, 2009). In addition, men have been rated higher than women in physical science disciplines (e.g., Basow & Silberg, 1987; Potvin, Hazari, Tai, & Sadler, 2009).

Student gender. The picture becomes even more complicated when one considers the effects of student gender, gender role stereotypes, and academic discipline. There seems to be a complex interaction between student gender and the gender of the instructor. Although some studies have shown that female students compared to male students tend to indiscriminately give higher ratings in general (e.g., Bachen et al., 1999; Badri, Abdulla, Kamali, & Dodeen, 2006; Darby, 2006a; Santhanam & Hicks, 2002), other research indicates a same sex preference for instructors. In several studies, female students gave higher ratings to female instructors while male students preferred male instructors (e.g., Das & Das, 2001; Lueck & et al., 1993; Ory, 2001). However, others found that the same sex preference was limited to female students, while men did not indicate any instructor gender preference (e.g., Bachen et al., 1999; Centra & Gaubatz, 2000).

Gender role attitudes. There is also some evidence that gender role dynamics between students and instructors can affect SET ratings. Research has shown that female instructors who do not conform to traditional feminine gender roles (i.e. being nurturing, deferring, nice, and relational) tend to be perceived negatively by both male and female students (Bachen et al., 1999; Basow & Montgomery, 2005; Basow et al., 2006; Bennett, 1982; Martin, 1984). The same might also hold true for male instructors who do not behave in traditionally masculine ways (Swaffield, 1996). Although these effects occur to some

extent across all four possible gender pairings, the influence of gender role stereotypes seems to be most pronounced for the male student / female instructor pairing, with the result that women teachers are held to a higher standard by their male students (Basow et al., 2006; Martin, 1984; Pounder, 2007).

Lastly, the extent of gender and gender role interactions might also be dependent on the academic discipline. For example, Basow and Montgomery (2005) have found that female instructors in the humanities and social sciences were rated higher then male instructors on interpersonal SET dimensions, but the effect was reversed for instructors in the physical sciences (male instructors were rated higher than female instructors on interpersonal characteristics). Overall, however, no consistent effects have been shown across studies with regard to academic discipline as a moderator of the gender-SET relation (Bachen et al., 1999; Basow & Montgomery, 2005; Centra & Gaubatz, 2000).

Student Individual Differences

Little research has been conducted on how students' background (e.g., in terms of their interests or personality) might influence their rating behavior. The only student characteristics that have been investigated for their influence on SET scores are the students' interest in the course subject, and student personality and social style.

Domain-specific vocational interests. Previous research has found that students' higher interest in the course content was associated with higher SET scores (e.g., Granzin & Painter, 1973; Greimel-Fuhrmann & Geyer, 2003; Marsh & Roche, 1997). However, it is not possible based on the research design of these studies to decide whether student interest in the course was preexisting or instilled by the instructor. The present study attempts to fill this gap in the literature by evaluating how students' preexisting interest in a variety of domains

affect their SET ratings. Both broad and narrow domains were included; these were the six Holland (1997) interest types, and three specific domains that matched the chosen course types, namely interest in statistics, social science, and social service.

Vocational interests are commonly defined as "patterns of likes, dislikes, and indifferences regarding career-relevant activities and occupations" (Lent, Brown, & Hackett, 1994, p.88). Among the many models that have been developed for classifying people based on their interests, Holland's (1959, 1997a) typology is probably the most commonly used system. Holland specified six distinct types (known as the RIASEC types realistic, investigative, artistic, social, enterprising, and conventional) that are structurally arranged in a hexagon (see Figure 1). The six types can be described as follows: Individuals with high realistic interests value practical tasks, they enjoy working with their hands, they like working with tools and are mechanically inclined. This type tends to enjoy the outdoors and athletic activities. Examples of majors with a high realistic component are mechanical engineering and agricultural sciences. Investigative types enjoy working with theories and abstract ideas. They like to conduct research and other intellectual and scientific pursuits. Examples of investigative majors are chemistry and biology. Artistic personalities value creativity, originality, aesthetics, and they are often non-conforming and independent. Typical majors in the artistic category are music and graphic design. Social types like to directly work with people in cooperative environments where the emphasis is on helping, instructing, and supporting others. Majors that are high on the social dimension are education and social work. Enterprising types enjoy leading and persuading others, selling, and managing. They highly prize status and often work in corporate environments. Representative enterprising majors are business and law. The conventional type is

characterized by high interest in activities that require attention to detail, organization skills, and efficiency. Typical conventional majors are accounting and computer science.

As illustrated in Figure 1, Prediger (1982) modified the RIASEC taxonomy by specifying two underlying bipolar dimensions, namely *people vs. things* (PT), and *data vs. ideas* (DI), that provide a more parsimonious explanation of the structure of the Holland model (see also Rounds & Tracey, 1993). The people/things dimension denotes the extent to which an occupation involves impersonal tasks as opposed to interpersonal relations with others. The data/ideas dimension represents the degree to which the occupational tasks are intrapsychic (thinking, using knowledge and insight) as opposed to having an external and data-related focus.

Domain-specific self-efficacy. No studies could be found in which students' domain specific self-efficacy beliefs have been tested for their effect on SET ratings. However, it is possible that students who have confidence in their domain-specific abilities might feel more positive about the instructor of a course in this area. Therefore, students' confidence in their ability to perform activities related to the six RIASEC types was assessed in parallel to their interest in these domains.

Personality. Only two studies could be located that address the relation between students' personality/social style and their SET ratings. To study the impact of students' personality style on SET ratings, Munz and Munz (1997) correlated both students' mood trait (positive and negative affectivity) and mood state at the time of SET administration with their SET ratings. The authors found no link between trait variables and SET ratings, but there was a modest positive correlation between mood state at the time of evaluation and SET scores. Schlee (2005) studied the influence of social style on instructor preferences in a

sample of different groups of business majors, namely the people-oriented majors, marketing and management, in comparison to quantitative areas (economics, accounting, and finance). The analyses indicated that students in different majors had different social styles: The proportion of students with an expressive social style (characterized by descriptors such as excitable, enthusiastic, stimulating, dramatic, and friendly) was significantly higher in people-oriented majors than in quantitative majors. Conversely, students in quantitative areas were more likely to belong to the "driver" social style (strong-willed, independent, tough, dominating, harsh, and efficient) or to the analytical social style (orderly, industrious, persistent, serious, exacting, and critical). With regard to instructor preference ratings, there was an interaction between the social styles of students and instructors: Students had a tendency to rate instructors higher when they matched their own social style. For example, students with expressive styles appreciated instructors who were responsive and caring, but these characteristics were perceived negatively (as weak and acquiescing) by students with an analytical social style.

The results from these two studies regarding the role of personality in students' rating behavior are inconclusive. Therefore, in order to gain a better understanding of how students' personality traits influence their SET ratings, a measure of the Big Five personality traits (neuroticism, agreeableness, openness, conscientiousness, and extraversion) was included in this study.

Research Questions and Hypotheses

The formulation of specific hypotheses for this study is challenging for three main reasons: First, many of the variables included in the present study have not yielded consistent effects in previous studies. Second, previous findings suggest that there might be a multitude

of moderator and mediator variables that affect the relation between the target variable and SET ratings in complex ways. Finally, many of the variables included in the present study have not been examined at all in the prior literature on SET ratings. With these caveats in mind, three main questions were examined as part of the present study: 1) Are there any systematic differences in students' SET ratings across course type and instructor gender, the two variables manipulated in the instructor vignette? Is the effect of instructor gender moderated by the gender of the students, i.e., do students show a same-sex preference for instructors as prior findings suggest? 2) Which of the student individual difference variables are systematically related to SET ratings? Can these variables be integrated conceptually in a meaningful way? 3) Do female students on average give higher SET ratings than male students as the prior literature suggests? If so, can this gender difference in SET ratings be explained by gender differences in students' interests and personality traits? These three questions are discussed below, including conceptual considerations and possible hypotheses

Course type, instructor gender, and student gender as moderator. The first question that this study sought to answer related to the influence of instructor gender and course taught and a possible moderating effect of student gender. Research has consistently shown that students have a preference for human service/relations courses over quantitative courses (Conners et al., 1998; Early, 2007; Manning et al., 2006; Vittengl et al., 2004). Therefore, it was expected that, irrespective of instructor or student gender, students would give significantly higher mean ratings to the counseling psychology vignettes compared to the research methods vignettes.

Previous research has not shown consistent global effects of instructor gender on SET ratings. Instead, there is evidence that the effect is moderated by the gender of the student

(e.g., Basow & Montgomery, 2005; Centra & Gaubatz, 2000; Das & Das, 2001; Lueck & et al., 1993). Based on these findings, it was predicted that male students on average would give significantly higher ratings to male instructors relative to female instructors, while the ratings given by female students would show the reverse pattern (significantly higher mean ratings for female instructors compared to male instructors). Therefore, it was hypothesized that there would be a significant interaction between student gender and instructor gender.

Student individual differences and their conceptual integration. One central focus of this study was to examine the effect of student individual differences on SET ratings. However, the goal was not only to empirically identify the specific traits that significantly predict SET scores, but also to evaluate whether these traits can be integrated into a conceptually meaningful pattern based on existing research. Over the past two decades, researchers have attempted to create theoretical frameworks that integrate a variety of individual difference variables (e.g., ability, interests, and personality) in a comprehensive manner. For example, Ackerman (Ackerman, 1996; Ackerman & Heggestad, 1997) proposed the PPIK theory (where PPIK stands for the four main components of the theory: Intelligence-as-Process, Personality, Interests, and Intelligence-as-Knowledge), which describes how cognitive and non-cognitive factors (such as personality and interests) conjointly influence the development of adult intellect. Ackerman proposed four trait complexes that conceptually integrate specific interests, personality traits, and types of ability. The first trait complex, the social complex, is posited to include social and enterprising interests, and extraversion; no specific abilities are associated with the social complex. The second trait complex, the clerical/conventional complex, is thought to include conventional interests, conscientiousness, and perceptual speed abilities. The third complex,

science/math, integrates realistic and investigative interests, and visual perception and math reasoning abilities; no Big Five personality traits are thought to relate to this complex. The last complex, the intellectual/cultural complex is thought to include artistic and investigative interests, openness to experience, and abilities related to crystallized intelligence and ideational fluency.

The proposed links across individual difference domains have also been investigated empirically (Gasser, Larson, & Borgen, 2004; Larson & Borgen, 2002; Larson, Rottinghaus, & Borgen, 2002; Staggs, Larson, & Borgen, 2003, 2007). For example, Larson and colleagues (2002) conducted a meta-analysis of 24 studies in which correlations between the six RIASEC interest types and the Big Five personality traits were reported. They found that five of the 30 possible bivariate correlations were substantial, with mean correlation coefficients of $r = .48$ (artistic interests with openness), $r = .41$ (enterprising interests with extraversion), $r = .31$ (social interests with extraversion), $r = .28$ (investigative interests with openness), and $r = .19$ (social interests with agreeableness).

In order to give a more meaningful view of how individual differences might impact SET ratings, the results of the present study were interpreted conceptually based on the available empirical evidence for systematic correlations between the respective individual difference domains.

Student gender. There is some evidence in the literature that female students indiscriminately tend to give higher SET ratings than their male peers (e.g., Bachen et al., 1999; Badri et al., 2006; Darby, 2006a; Santhanam & Hicks, 2002). Therefore, it is hypothesized that, on average, female students in the present study will rate the vignettes significantly more favorably compared to their male peers.

One rationale of the present study, however, was to go beyond the descriptive level at which prior findings have been reported in the literature. Therefore, the goal was to find the reason or mechanism behind the observed gender difference in SET ratings. For example, it might be possible that the gender differences in SET ratings might be caused by systematic differences between men and women's reported levels of specific personality traits and interests (i.e. the individual difference variables serve as a mediator of the gender difference in SET ratings). Gender differences in interests and personality are well documented in the empirical literature. For example, men reported significantly higher interest in things, while women scored higher towards the people pole of Prediger's (1982) people-things dimension (Lippa, 1998; Su, Rounds, & Armstrong, 2009); the magnitude of this difference was about one standard deviation (Su et al., 2009). Across the six RIASEC types, men scored higher than women on realistic (d = 0.84) and investigative (d = 0.26) interests. Conversely, women reported higher interest than men in artistic (d = 0.35), social (d = 0.68) and conventional (d = 0.33) activities (Su et al., 2009). The largest gender differences on the Big Five personality traits have been consistently found for neuroticism and agreeableness, with women reporting higher levels of each trait by up to half a standard deviation (Costa, Terracciano, & McCrae, 2001; Feingold, 1994; Lippa, 2010; Schmitt, Realo, Voracek, & Allik, 2008).

CHAPTER 2: STUDENT EVALUATION OF TEACHING – REVIEW OF THE LITERATURE

Introduction and Overview

Student Evaluation of Teaching

The systematic evaluation of teaching performance has become standard across universities in the United States over the past decade (Clayson, 2009; P. M. Simpson & Siguaw, 2000; Sproule, 2000). Of the various procedures in place, Student Evaluation of Teaching (SET) has emerged as the most important, sometimes even the sole measure for assessing the teaching effectiveness of college teachers (Wilson, 1998). The SET process is typically implemented as follows (Algozzine et al., 2004; Richardson, 2005; Sproule, 2000): Students are asked at the end of the academic semester to rate the teaching effectiveness of their instructor and the quality of the course by anonymously completing a self-report questionnaire. The items included in such evaluation forms can refer to various dimensions of perceived teaching effectiveness (e.g., ability of the instructor to communicate clearly), characteristics of the course and educational context (e.g., class level), and student demographic variables (e.g., gender, year in school, etc). Items can be global (e.g., "Instructor quality overall") or refer to specific aspects of the instructor or course. Some questions are closed-ended, and students select a response option from a Likert-type scale, hereby indicating their degree of agreement with the respective statement. Other questions are open-ended (e.g., "What did you like most about this course?"), and students create their own response. The completed questionnaires are analyzed by designated entities within the university, and the results are returned to the department administration and the instructor for

review. A typical SET report contains descriptive statistics (mean, standard deviation, range, modal response, etc) for the items scored on a scale, as well as the qualitative feedback provided by the students.

SET data are primarily used for the following four purposes (Algozzine et al., 2004; Hobson & Talbot, 2001; Marsh & Dunkin, 1992): a) as diagnostic feedback to instructors about their teaching effectiveness (formative function); b) as input factor for administrative decisions (summative function, e.g., hiring, tenure and promotion, salary raises, and awards); c) as information for students to guide them in their selection of courses and instructors; d) as process or outcome variable in research on teaching.

Since their first appearance in the mid 1920s, SET procedures have been continuously used, researched and fiercely debated. Proponents of the SET process believe that SETs are a reliable and valid indicator of teaching effectiveness and quality of instruction, and that they provide meaningful feedback that helps instructors improve their teaching (e.g., Aleamoni, 1999; Davis, 2009; Marsh & Roche, 1997; McKeachie, 1997). Others argue that SETs are invalid, fraught with bias, a major contributor to grade inflation, and that they are routinely misused by administrators in high-stakes personnel decisions (e.g., Crumbley & Reichelt, 2009; Goldman, 1993; Gray & Bergmann, 2003; Sproule, 2000). The ensuing debate, often emotionally charged, is complicated by the complexity of the issue, the inconsistency and ambiguity of research findings, and the potentially negative consequences that improper use of SET results can have on instructors and the university system as a whole. Virtually every aspect of the SET procedure has been contended: a) the challenge of how to define and measure teaching effectiveness; b) the extent to which SET scores are related to meaningful criteria such as student learning, and to extraneous factors that should not play a role in

teaching effectiveness; c) the question of how, if at all, SET scores can be used appropriately in the service of instructional quality assurance.

Messick's Concept of Validity as Organizing Principle of the SET Controversy

The body of SET literature that has accumulated over the decades is extensive, and the different lines of inquiry frequently appear disjointed from one another when viewed from a superficial perspective. Upon closer examination, however, a common thread emerges that connects the various research areas, which is the central question of meaningfulness and utility of the SET process. In other words, the major SET debates can be represented as different aspects within the larger concept of validity. Therefore, the purpose of the present literature review is to provide an overview of the SET controversy anchored within Messick's (1989, 1995) conceptualization of validity as organizing principle.

Messick (1995, p. 741) views validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment". According to Messick, validity is a unified concept that is comprised of two interrelated aspects: a) the empirical evaluation of score meaning based on the scientific method (evidential validity, this includes the integration of evidence from multiple lines of inquiry, such as construct-related considerations, and convergent and discriminant evidence), and b) the actual and potential social consequences of assessment use (consequential validity). It is possible to frame the various SET controversies in terms of the different facets of validity as specified by Messick. The review will be organized based on four broad themes, each representing one of the lines of validity evidence.

The first line of evidence relates to the definition and content of the construct of teaching effectiveness. Subsumed under the line of construct-related considerations is the challenge of finding a consensus on how to define teaching effectiveness, including the question of dimensionality of this construct. Further, issues in questionnaire construction based on the varying definitions of teaching effectiveness will be discussed, both with regard to content and item selection, as well as the response format of the scales used.

The second line of inquiry concerns empirical research on the convergent validity of SET measures. Various attempts have been made to validate SET measures by demonstrating that SET scores are related in meaningful ways to other criteria such as student learning. Analogue to the definition of teaching effectiveness, there is little consensus on how to best conceptualize and measure student learning. The controversial correlation between SET scores and students' expected grade will be discussed in this context.

Discriminant evidence for validity, the third line of inquiry, can be obtained by investigating a construct's relation to extraneous variables to which it should not be related based on logic and theoretical rationale. Thus, discriminant evidence of SET validity can be obtained by demonstrating that SET scores are unrelated to variables that should not have an effect on instructor teaching effectiveness. A large portion of the literature has been devoted to examining the influence of a variety of background factors related to students (e.g., gender and prior subject interest), the instructor (e.g., attractiveness and personality), the educational context (e.g., academic discipline and class size), or psychosocial dynamics that play out in the class room (e.g., the Halo effect).

The final line of validity evidence concerns the social and political consequences of SET use (consequential validity). A major concern voiced by both scholars and instructors

alike is the potential misuse of SET data by university administrators in personnel decisions. Therefore, the final part of this review will focus on the attitudes and perceptions that instructors and students have regarding the SET process, and policy implications that have been highlighted in the literature. The review will conclude with a summary and critique of the evidence for the validity of the SET process.

Construct Validity

What is effective teaching? The first step towards establishing any meaningful assessment procedure is to clearly define the construct of interest. The first part of the present review will address the challenge of finding consensus on how to define and measure teaching effectiveness. The following questions within this line of inquiry will be emphasized: Why is it so difficult to agree on what SETs are supposed to measure? Is teaching effectiveness unidimensional or multidimensional? How are SET measures constructed and what are their psychometric shortcomings?

Defining and Operationalizing Teaching Effectiveness

The assumption underlying SET measures is that their content reflects characteristics that are relevant to the teaching process, and that can be accurately assessed by students (d'Apollonia & Abrami, 1997). More than 25 criteria have been used over time to define teaching effectiveness (d'Apollonia & Abrami, 1997; Dowell & Neal, 1982; Elliot, 1950). This lack of consensus has been traced back to multiple factors: First, definitions of teaching effectiveness have emphasized either the *process* (i.e., the specific teaching behaviors, such as providing feedback to students) or the *learning outcomes* that these behaviors promote in students (i.e., skills acquisition, performance on tests) (Abrami, d'Apollonia, & Rosenfield,

1996; Barnes et al., 2008; d'Apollonia & Abrami, 1997). A quote by Buck (1998, p. 1224) illustrates this difference in perspective:

> "Equating teaching behaviors or styles with effectiveness is ubiquitous. This can be demonstrated by asking faculty members or a class of students "What do you think best characterizes or defines effective teaching?" Rarely will the reply be "whether or not the students came to master the course objectives by the end of the semester." Rather, you will likely receive responses referring to how clearly an instructor speaks, course organization, method of testing, and the enthusiasm of the instructor. On the contrary, if you ask these same people "How would you know if a course on CPR was effective?" the most common response would likely include statements regarding whether the students who took the course could save lives or at least demonstrate life-saving skills on the practice dummy."

The perspective has shifted somewhat from process-oriented to outcome-oriented measures with the more recent advent of student-centered teaching philosophies (Edstrom, 2008; Frick, Chadha, Watson, & Zlatkovska, 2010).

A second reason for the lack of consensus in the definition of teaching effectiveness stems from differences in how students and instructors weigh and interpret aspects of teaching. For example, Feldman (1988) has summarized the results of 31 studies, which reported characteristics of effective instructors as viewed by both faculty and students. He found that, although there was general agreement between the two groups (the average correlation between faculty and student responses was .71), some systematic differences were found. Students compared to faculty placed more importance on the teacher being available and helpful, and having an engaging presentation style. Faculty, on the other hand,

emphasized that an effective teacher should be intellectually challenging, able to motivate students, and to set high standards.

Further studies have shown that students often interpret SET survey items differently than intended by the creators of the measure. For example, Dujari (1993) demonstrated that many of the university Freshmen enrolled in academic skills classes were not able to understand basic vocabulary included in the SET form. The lack of understanding of education-related terminology has also been uncovered in a qualitative study, in which 24 students from a medical program were engaged in think-aloud interviews while completing their course evaluation measure (Billings-Gagliardi, Barrett, & Mazor, 2004). The data revealed that students often interpreted items in idiosyncratic and sometimes contradictory ways. For example, 29% of the students were uncertain about the meaning of "independent learning". Likewise, 42% of students did not know the meaning of the item "integration with other disciplines" or they doubted their ability to make an accurate judgment. Several students indicated that they always responded to items with unclear meaning by selecting the second-highest response option on the scale. Sometimes students even interpreted an item counter to its intended positive meaning. For example, one student commented on how he feels about independent learning (Billings-Gagliardi et al., 2004, p. 1066): "A lot of independent learning means the course isn't providing needed material. So if I say *a lot*, it's a negative". Not surprisingly, it has been shown that global items (e.g., overall rating of the instructor) are more susceptible to ambiguous interpretation by students (Kolitch & Dean, 1998). Fritschner (2000) reported that faculty and students also markedly differed on the interpretation of educational concepts. For example, his research showed that faculty interpreted "class participation" in terms of note taking, staying awake in class, and

completion of assignments on time. Students on the other hand viewed class participation as being actively involved in discussions and other class activities.

Dimensionality of Teaching Effectiveness

Is teaching effectiveness a unidimensional construct or are there multiple dimensions that represent different aspects of teaching and learning? Models that have been proposed range from one single dimension (Abrami & d'Apollonia, 1991; d'Apollonia & Abrami, 1997) up to nine (Marsh, 1984, 1987; Marsh & Dunkin, 1992), with various structures in between (e.g., two factors (E. H. Cohen, 2005), three factors (Covert & Mason, 1974), or five factors (Barth, 2008)). Both conceptual and methodological problems have been identified in the literature to explain this lack of consensus. First, as already discussed, scholars disagree on how to define and interpret teaching effectiveness in a more general sense. Second, SET measure development is most often purely inductive rather than theory-driven (Barnes et al., 2008; d'Apollonia & Abrami, 1997; Hobson & Talbot, 2001; Marsh & Roche, 1997). Although some standardized measures are available for purchase (see Richardson, 2005, for a review), the majority of SET measures are developed on site by administrators or instructors with little to no training in psychometric theory and educational assessment (Hobson & Talbot, 2001; Marsh & Roche, 1997). Items are often obtained by surveying and tweaking existing measures (Barnes et al., 2008; Keeley, Smith, & Buskist, 2006), or by asking students and instructors for their input (Barnes et al., 2008; Cunningham & MacGregor, 2006; Marsh & Roche, 1997). The dimensionality of the resulting SET measure is then assessed by factor analysis.

D'Apollonia and Abrami (1997) provide a detailed critique of the post-hoc use of factor analysis for establishing the dimensionality of teaching effectiveness. They argue that

the factors reflect the specific items in each SET questionnaire rather than the underlying theoretical construct of teaching effectiveness. Therefore, it would be expected for the factor structure to vary across different SET measures. In addition, they contend that researchers apply different decision rules when interpreting their data (e.g., retention of factors, rotation of the factor structure, etc.). The authors reanalyzed the factor structure of seven SET measures, including Marsh's (1984, 1987) nine-dimensional Student Evaluations of Educational Quality (SEEQ) instrument. The SEEQ's multidimensional structure has previously been cross-validated in over 30 studies (Marsh, 1987). However, Abrami and d'Apollonia (1991) showed in their reanalysis, that a principal-components solution (with 31 out of 35 items loading .60 or higher on the first principal component) was interpretable and accounted for the same amount of variance in the SEEQ scores as the 9-factor solution. Likewise, one global component explained about 30% of the SET score variance across the seven measures they analyzed. Therefore, they proposed that teaching effectiveness should be viewed as one global dimension of General Instructional Skill.

The issue of teaching effectiveness dimensionality not only has theoretical implications, but also extends to the practical realm. It has been argued that the intended use of the data should determine whether global or faceted scores should be reported. Administrators often find it convenient to rely on a single performance indicator when making personnel decisions (McKeachie, 1997). On the other hand instructors tend to prefer feedback on specific aspects of their teaching in order to implement changes (Marsh & Roche, 1997). As a compromise, it has been suggested that a weighted average might be used that places more emphasis on aspects deemed most relevant to the specific application context (Marsh, 1994, 1995).

Psychometric Shortcomings of SET Questionnaires

As previously described, even psychometrically sound items are often interpreted by students in multiple and unintended ways. This issue becomes even more of a concern when the instrument itself is flawed in item wording or scaling of response options (e.g., Sedlmeier, 2006; Tagomori & Bishop, 1995; Tierney & Simon, 2004). Tagomori and Bishop (1995) evaluated 4,028 items included in 200 SET questionnaires for adequacy in wording, response format, and relevance to teaching. They found that 79% of all items were flawed in one or more of these ways. The average questionnaire included 17.7 flawed items out of a total of 20.1 items.

With regard to item wording, Tagomori and Bishop (1995) found that about 55% of all items were ambiguous, unclear, or subjective in wording, and 25% of items were irrelevant to teaching. Ambiguous items were those that could be understood in multiple ways, or that included two or more characteristics of teaching behavior simultaneously (e.g., "How clear were the goals, aims, and requirements of the course?"). Unclear items were items stated in too general and imprecise ways (e.g., "The total experience under the control of this person was very worthwhile"). Subjective items assumed that students are able to infer the feelings or opinions of other students or the instructor (e.g., "Main points of lectures were clearly understood by students in class"). Items irrelevant to the instructor's teaching behavior included examples such as "How well are you able to take notes?".

The choice of scale type and response options has also been shown to greatly influence SET ratings. Specific problems that have been identified include (Darby, 2008; Franklin, 2001; Sedlmeier, 2006; Tagomori & Bishop, 1995; Tierney & Simon, 2004): a) Positive or negative skew in the response options, resulting in more positive than negative

choices, and vice versa; b) Ordering of response alternatives in a check list of items; c) Use of bipolar vs. unipolar scales; and d) anchoring of responses (e.g., relative ratings in comparison to other instructors vs. absolute ratings). A result of these differences in scaling is that it becomes possible to move SET ratings upward or downward just by choosing a particular scale type or response format. Therefore, it is critical to take scaling issues into account when ratings are compared across different departments or institutions (Sedlmeier, 2006).

Summary: Construct-Related Validity

The first step towards establishing a valid assessment procedure is the definition of the construct to be measured. Several threats to the construct-related aspect of validity have been identified in the SET literature: So far, no consensus has been reached on how to best define and measure the construct of teaching effectiveness. The lack of consensus is the result of both conceptual disagreements and methodological shortcomings. Conceptual problems include the confusion about whether to focus on teaching processes or outcomes as defining criterion, and the lack of educational theories that could guide SET construct validation. Methodological issues include the reliance on "dust bowl" empiricism in the development of SET questionnaires, and the use of factor analysis to assess the dimensionality of the construct of teaching effectiveness. In addition, many SET instruments currently in use exhibit psychometric flaws such as inadequate wording of items and selection of scale formats.

Convergent Validity

A second means of supporting the validity of test procedures is to obtain convergent evidence, i.e., to demonstrate empirically that test scores are related in logically consistent

and theoretically meaningful ways to specific criteria. The second part of the present review will discuss convergent evidence for the validation of the SET process. Three major questions will be addressed: Are SET scores related to student learning? Do SET scores correlate with performance ratings by other groups (e.g., instructor self-ratings, peer/administrator ratings, and alumni ratings)? Do SET ratings change with instructor teaching experience?

Relation of SET Scores to Student Learning

If SET questionnaires are a valid measure of teaching effectiveness, SET scores should be positively related to learning outcomes (Bain, 2004; Clayson, 2009; d'Apollonia & Abrami, 1997; Marsh & Roche, 1997). As with the definition of teaching effectiveness, however, there are multiple perspectives on what constitutes learning and how it should be measured (see Clayson, 2009, for an overview). Both objective (e.g., test scores, actual grades obtained by the students) and subjective (e.g., students' own perception of how much they have learned, and what grade they expect to receive at the end of the semester) have been used as criteria to validate SET scores.

Relation of SET scores to objective learning criteria. The bulk of research on the relation between SET scores and objective learning criteria has been conducted through multisection studies (e.g., Clayson, 2009; P. A. Cohen, 1981, 1982; d'Apollonia & Abrami, 1997; Greenwald & Gillmore, 1997). In multisection studies, different instructors teach one of several sections of the same course, and students are randomly assigned to sections. Since all sections use the same syllabus, textbook, and final examinations, the assumption is that differences in mean test performance across sections is a function of the teaching effectiveness of the instructor. Countless multisection studies have been conducted over the

last decades, and several meta-analyses are available (e.g., Clayson, 2009; P. A. Cohen, 1981, 1982, 1987; d'Apollonia & Abrami, 1996; McCallum, 1984). The results across these studies have been fairly consistent: overall instructor ratings seem to show moderate correlations with objective test performance (uncorrected validity coefficients were typically in the range of .30 to .40, meaning that SET scores accounted for 9% to 16% of the variance in test scores across sections). Even though the results have been consistent across studies, scholars are divided in terms of how to interpret these numbers. Some have argued that the obtained effect sizes support the conclusion that SET scores reflect learning (e.g., P. A. Cohen, 1981; d'Apollonia & Abrami, 1997; Marsh & Roche, 1997), while others maintain that the effect is not sufficient to provide convergent evidence for SET score validity (e.g., Dowell & Neal, 1982; McCallum, 1984; Pounder, 2008). Abrami, Cohen, and d'Apollonia (1988) contend that the researchers' varying interpretations were due to critical differences in their meta-analysis procedures (e.g., differences in the inclusion criteria for studies, whether corrections to the validity coefficient were made to account for unreliable measures, and use of statistical techniques (see also d'Apollonia & Abrami, 1996, 1997)).

Multisection studies have also been criticized for other reasons. Scholars have pointed out that most multisection studies were conducted on large introductory courses that relied on the traditional lecture approach in conjunction with tests that mainly required rote memorization of facts. However, many alternative course formats (e.g., small interactive seminars, online courses) have been developed over time. It is unclear how the results of these studies generalize to alternate contexts and performance indicators of higher-level learning outcomes (Abrami & d'Apollonia, 1990; J. S. Armstrong, 1998; Clayson, 2009; d'Apollonia & Abrami, 1997; McKeachie, 1997).

Relation of SET scores to students' subjective perception of learning and performance. A large part of the SET literature has been devoted to the relation between SET scores and students' subjective perception of their learning and performance. If SET measures are valid indicators of teaching effectiveness, students who report a high degree of learning should also give higher SET ratings. Two major debates have dominated this area of the literature; these are the assumption that students have the ability to accurately assess their own learning, and the grading-leniency hypothesis.

Are students able to accurately assess how much they have learned from their instructor? This is the basic question that needs to be answered in order to draw meaningful inferences from research on the relation between SET scores and students' own perception of their learning. Whereas some scholars have maintained the stance that students themselves are the best judges of how much they have learned (e.g., Aleamoni, 1999; Cruse, 1987; Davis, 2009; Machina, 1987; Marsh & Roche, 1997), many others have challenged this assumption.

For example, research has shown that students often do not have the meta-cognitive skills to assess their learning, and that they make consistent errors when estimating their performance (e.g., Browne, Hoag, Myers, & Hiers, 1997; Clayson, 2009; Sproule, 2000). Another challenge of the assumption that students can accurately judge how much they have learned is based on research on the influence of instructor presentation style (Abrami, Leventhal, & Perry, 1982; Naftulin, Ware Jr., & Donelly, 1973; Shevlin, Banyard, Davies, & Griffiths, 2000; R. G. Williams & Ware Jr., 1977; W. M. Williams & Ceci, 1997). The basic paradigm has the following design: Students listen to a lecture that is being delivered under different experimental conditions. The first independent variable is the instructor's

presentation style, which is either dull or "seductive" (i.e., full of dramatic speech, use of gestures, frequent modulation of voice, entertaining elements, etc.). The second independent variable is the level of coherence of the presentation content (coherent vs. incoherent). Students are randomly assigned to attend the lecture in one of the four conditions. After the lecture, students rate how much they think they have learned, and the quality of the instructor on different dimensions of teaching effectiveness (e.g., knowledge, enthusiasm, etc.). Students then complete an objective test of the material covered in the lecture. Typical results show that, irrespective of presentation style, students in the coherent content condition objectively retain significantly more information than students in the incoherent content condition. The students' subjective rating of how much they have learned, however, has shown to be moderated by presentation style. Under the dull presentation condition, students in the coherent content condition subjectively report more learning than those in the incoherent condition, which is consistent with the objective test results. Under the seductive presentation condition, however, students, irrespective of whether they had attended a coherent or incoherent lecture, reported the same (high) amount of learning (even though students in the incoherent condition objectively had learned less). The same pattern is obtained for the quality of instructor ratings. In sum, when a lecture is delivered in a highly entertaining presentation style, students can be seduced into thinking that they have learned a lot, even though the objective test performance does not reflect this. Obtained effect sizes can be quite large; mean differences in SET ratings between the dull and seductive presentation conditions have been found to exceed one standard deviation in some cases (e.g., W. M. Williams & Ceci, 1997). The influence of a seductive presentation style on SET ratings is also called the "Dr. Fox effect", named after the fictitious character played by a professional

actor in the first installment of this paradigm (Naftulin et al., 1973). In sum, based on research on students' meta-cognitive abilities as well as the influence of instructor presentation style, there is evidence that students might not always be the best judges of how much they have learned.

To validate SET ratings based on students' subjective perception of learning, the relation between students' expected grade and rating of the instructor has been studied extensively. Currently, there is consensus that a moderate (r = .20 to .40) correlation between expected grades and SET ratings exists (Clayson, Frost, & Sheffet, 2006; Gillmore & Greenwald, 1999; Greenwald, 2002; Marsh & Roche, 1997; Wachtel, 1998). However, there is no agreement on how this relation should be interpreted. Two general hypotheses have been proposed to account for this effect (see Clayson, 2009; Greenwald & Gillmore, 1997; McKeachie, 1997, for a detailed summary and critique of the different positions), namely the validity hypothesis and the grading-leniency hypothesis. According to the validity hypothesis, the observed grade-SET relation is meaningful and interpretable (e.g., Howard, 1984; Marsh & Roche, 1997; McKeachie, 1997; Ory, 2001; Spooren & Mortelmans, 2006). Third variables that are assumed to be instrumental in student learning (e.g., student motivation or interest in the subject matter) are seen as the underlying cause of this relation. Again, the assumption is that students who have learned more can expect to obtain higher grades, and the higher amount of learning is then reflected in higher SET scores.

According to the grading-leniency hypothesis, the grade-SET relation is a spurious relation that is not based on student learning (e.g., Clayson et al., 2006; Gillmore & Greenwald, 1999; Greenwald & Gillmore, 1997; McKeachie, 1997; McPherson, 2006). Instead, the hypothesis specifies that teachers can "buy" higher SET ratings from students by

grading leniently. Therefore, the observed grade-SET relation would constitute bias and threaten the validity of SET ratings.

The empirical investigation of the competing hypotheses has yielded no clear evidence for either perspective, and both explanations might have merit (e.g., Marsh & Roche, 1997; McKeachie, 1997). However, several lines of evidence suggest that SET ratings can be influenced by lenient grading practices. First, experimental field research has shown that the manipulation of expected grades (students were led to believe that they could expect a certain grade that was randomly assigned to them) had an influence on SET ratings (Chacko, 1983; Holmes, 1972; Powell, 1977; Vasta & Sarmiento, 1979; Worthington & Wong, 1979). Second, the use of SET practices has been shown to contribute to grade inflation over time (Eiszler, 2002; Greenwald & Gillmore, 1998; McKeachie, 1997; McPherson, 2006). Third, Clayson and colleagues (2006) have conducted a longitudinal study over the course of a semester during which ratings were conducted at several time points. They found that instructors who gave good grades were rewarded with good SET ratings shortly after students obtained grade-related feedback. Conversely, instructors who gave poor grades were punished with low ratings. This effect was independent of any instructor or preexisting student characteristic. These and other findings have been interpreted as support for the hypothesis that lowering standards and giving undeserved high grades can bias SET ratings. To address this concern, it has been proposed to take grading standards into account when using SET scores in personnel decisions, e.g., through statistical corrections to remove grading leniency effects (e.g., Gillmore & Greenwald, 1999; Greenwald & Gillmore, 1997).

Relation of SET Scores to Instructor Self-Ratings and Peer Ratings of Teaching Effectiveness

A second way of obtaining convergent evidence to support SET validity is the correlation of SET scores with ratings by others who are presumed to be able to make judgments about the teaching effectiveness of the instructor. The assumption is that strong correlations between SET scores and independent ratings by others would show that SET ratings indeed capture teaching effectiveness. Researchers have compared SET ratings with teacher self-ratings, administrator and peer ratings obtained during class observation, and student alumni ratings.

Correlation of SET scores with instructor self-ratings. The correlations found for SET ratings with instructor self-ratings have been shown to be in the small to medium range. For example, Feldman (1989), based on a meta-analysis of 19 studies, reported a mean correlation of r = .29 for the overall instructor rating; mean correlations for specific SET components ranged between r = .15 to .42. Marsh and colleagues (Marsh, 1987; Roche & Marsh, 2000) found median correlations of r = .32 and r = .20 for the overall instructor rating, and median correlations for specific dimensions in the .40 to .50 range. In sum, the correlations between student ratings and instructor self-ratings seem to be in the medium range, and correlations tend to be higher for specific SET dimensions than for the overall instructor rating.

Correlation of SET scores with peer and administrator ratings. Ratings by peers or administrators who observe the instructor during a class period have been correlated with the ratings provided by students. Research has shown that these correlations tend to be small or non-existent (e.g., Centra, 1979; Koon & Murray, 1995; Marsh, 1987; Marsh & Roche,

1997). In addition, there seems to be little agreement between different peers who observe the same class (Marsh & Roche, 1997).

Correlation of SET scores with alumni ratings. Ratings of current students also have been compared to ratings of alumni who have taken a course from the same instructor in the past. Based on the available evidence there appears to be good agreement between current and former students in their view of an instructor. Feldman (1989) reported in his meta-analysis a mean correlation coefficient of $r = .69$ across six studies. Marsh (1977) asked recent graduates to nominate their best and worst teachers. The two instructor groups were then compared on multiple dimensions to ratings from the instructors' current students. The time span between the current student ratings and the alumni nominations was between one and four years. The discriminant analysis yielded a canonical correlation of .82 between the retrospective nominations by the alumni and the SET ratings by current students, and 92% of instructors were correctly classified into the two teacher groups based on the current student ratings.

Stability of SET Ratings over Time

Teaching effectiveness is assumed to increase with teaching experience. As instructors obtain and implement feedback and feel more comfortable in their role as teacher, they are presumed to become more effective (Marsh, 2007; Murray, Jelley, & Renaud, 1996). Therefore, a positive correlation between SET ratings and instructor's teaching experience can be interpreted as convergent evidence for SET validity. Cross-sectional research has yielded mixed results. Some studies have shown that more experienced faculty and those of higher rank have obtained higher SET ratings than those of lesser rank or with less experience (e.g., McPherson, 2006; S. P. Smith & Kinney, 1992). In addition, professors tend

to be rated higher than graduate teaching assistants (Ory, 2001). A review by Feldman (1983), however, yielded inconclusive results. Other research (e.g., Franklin, 2001; Langbein, 1994) has suggested that the relation between SET ratings and experience might be curvilinear, meaning that instructors with a moderate amount of experience are rated higher than those with either very little experience or those who have been teaching for a long time.

A similar picture emerges for longitudinal research in which the ratings of the same instructors were tracked over several years. While Marsh and colleagues (e.g., Marsh, 2007; Marsh & Hocevar, 1991; Marsh & Roche, 1997) found no systematic changes in ratings for a group of instructors over 13 years, a positive relation between SET ratings and experience was found in a study tracking a group of instructors over 21 years (Murray et al., 1996).

Findings from a study by Clayson (1999) suggest that some characteristics (e.g., knowledge, class organization, fairness of grading practices) change over time as the instructor gains more experience, while other teaching-relevant characteristics (personality traits and interpersonal skills such as being responsive and caring) remain stable over time. Therefore, the extent to which a correlation of SET ratings with experience can be found might depend on whether the SET items tap primarily into changeable or stable instructor characteristics.

Summary: Convergent Validity

Research on convergent validity of SET ratings has primarily focused on three areas: The relation of SET ratings to a) objective and subjective indicators of student learning, b) independent ratings by others such as peers and alumni, and c) instructor teaching experience. A multitude of studies have shown that SET ratings show a moderate relation to

objective (e.g., test scores), and subjective (e.g., students' expected grades) indicators of learning. However, there is no consensus in the literature as to how to interpret this relation. One contentious issue is the question of whether the magnitude of the obtained effect size is high enough to support SET score validity. In addition, some scholars have pointed out that students cannot always be trusted in making accurate judgments regarding their learning. Others content that SET scores can be biased by extraneous factors such as instructor presentation style and grading practices. Therefore, the question of how student learning is reflected in SET ratings remains.

With regard to the question of whether student ratings converge with instructor ratings by others, the evidence is mixed. Correspondence seems to be highest between ratings of current students and retrospective alumni ratings of the same instructor. Likewise, the instructors' own perception of their effectiveness as a teacher seems to be somewhat congruent with the ratings given by students. However, ratings by colleagues or administrators as part of classroom observations have found to be unreliable und not significantly related to student ratings. Finally, results on whether SET ratings increase with instructor experience are equally inconclusive; no consistent trend has been documented in the literature.

Discriminant Validity and Bias

Evidence of discriminant validity for SET ratings can be obtained by confirming empirically that no correlation exists between SET scores and variables that should not relate to teaching effectiveness based on logic and theoretical rationale. Conversely, a bias variable is a variable that correlates with SET scores despite being unrelated to teaching effectiveness. Therefore, bias poses a threat to the validity of SET ratings as a measure of teaching

effectiveness. A major part of SET research has been devoted to the evaluation of potential sources of bias in SET ratings. The variables that have been examined for their influence on SET scores can be conceptually arranged into four broad categories: a) Variables related to the educational context (e.g., academic discipline); b) Instructor-related variables (e.g., gender); c) Student-related variables (e.g., prior subject interest); and d) Variables related to specific psycho-social dynamics that play out in the class room (e.g., the importance of first impressions).

Variables Related to the Educational Context

The educational context in which an instructor teaches has been shown to impact SET scores. This is problematic since most of these variables are beyond the control of the instructor (e.g., class size, course format, time at which the class is held, etc.). Therefore, in order to ensure fairness (especially when instructors' SET ratings are compared against each other within a department or the university as a whole), it is important to assess whether they constitute a source of bias in SET ratings. The variables that have been examined empirically include academic discipline, class size, class level, whether the course is required or elective, and class time. In addition, factors related to SET administration (student anonymity, instructor presence during the SET administration, instructions given, and timing of SET administration) have been investigated.

Academic discipline. Research has shown that SET ratings tend to systematically vary across academic disciplines (e.g., Basow & Montgomery, 2005; Cashin, 1990; d'Apollonia & Abrami, 1997; Davis, 2009; Franklin, 2001; Marsh & Roche, 1997; Santhanam & Hicks, 2002). In general, instructors teaching in the arts and humanities are rated higher than those in the biological and social sciences, followed by business, computer

science, and mathematics/engineering; instructors in the physical sciences tend to obtain the lowest ratings.

Class size. SET ratings tend to be higher in small classes compared to large lecture classes (e.g., Feldman, 1984; Koh & Tan, 1997; Liaw & Goh, 2003; McPherson, 2006; Toby, 1993). However, some studies have found that the relation between class size and SET ratings might be curvilinear, with small and very large classes rated highest (Feldman, 1984; Fernandez, Mateo, & Muniz, 1998; Wachtel, 1998).

Course characteristics. Classes taught at a higher level tend to receive higher ratings than lower-level classes, and instructors of graduate level classes fare better than teachers of undergraduate classes (e.g., Bausell & Bausell, 1979; Feldman, 1978; Gaffuri, Wrench, Karr, & Kopp, 1982; Ory, 2001; Santhanam & Hicks, 2002). Elective courses and those within the students' major are rated higher than required courses (e.g., Costin, Greenough, & Menges, 1971; Darby, 2006b; Feldman, 1978; Marsh, 1980).

Class time. The day and time during which a class is scheduled has been shown to have a small and inconsistent effect on SET ratings (Cronin & Capie, 1986; DeBerg & Wilson, 1990; Husbands & Fosh, 1993; Pounder, 2007). For example, classes held early in the morning have received higher ratings than those held later in the day in one study (Badri et al., 2006). Other research has shown that classes scheduled during later days of the week were rated more favorably than those held earlier in the week (Koh & Tan, 1997).

Factors related to SET administration. Certain variables related to SET form administration at the end of the semester have shown to affect SET ratings. Administration of the survey during a regular class period yielded higher ratings than administration on the day of the final exam (Ory, 2001). The instructions given to students also can make a difference.

When students were informed that SET scores would be used in personnel decisions, the instructors received higher ratings than when students were not made aware of this fact (Ory, 2001). When the instructor remained in the room during SET administration, the ratings were higher than when the instructor was absent (Marsh & Roche, 1997; Ory, 2001; Wachtel, 1998). Students who were required to sign their name on the SET form gave higher ratings than students who were allowed to remain anonymous (Marsh & Roche, 1997; Ory, 2001; Wachtel, 1998).

Influence of Instructor-Related Variables on SET Ratings

Various personal characteristics of the instructor have been evaluated for their potential influence on SET ratings. Similar to contextual variables, most of these factors are beyond the instructor's control (e.g., gender, race, personality traits, etc.). Thus, in order to ensure fairness when using SET scores for summative purposes, it is important to examine whether these instructor-related variables have a systematic impact on SET ratings. The variables that have been studied empirically include instructor gender, race, sexual orientation, personality traits, and attractiveness.

Instructor gender. Research on the relation between instructor gender and SET scores has yielded mixed and complex findings: Some studies have found higher global SET scores for male instructors compared to female instructors (e.g., Sidanius & Crane, 1989; B. P. Smith, 2009), some have found the reverse pattern of results (e.g., Basow & Montgomery, 2005; Whitworth et al., 2002), and others have found no systematic difference in SET evaluations based on instructor gender (e.g., Feldman, 1992, 1993; G. Smith & Anderson, 2005). However, more fine-grained analyses have revealed an interesting pattern. When female teachers received higher ratings than men, it was usually on dimensions that captured

the interpersonal relations between instructor and students: Generally, women were praised for being caring, empathetic, approachable, and for fostering a good relational climate in the class room (e.g., Bachen et al., 1999; Basow & Montgomery, 2005; Basow et al., 2006; Bennett, 1982; Centra & Gaubatz, 2000; Kierstead et al., 1988). Men, on the other hand received higher ratings on dimensions such as course planning, competence and knowledge, and organization skills (e.g., Basow et al., 2006; B. P. Smith, 2009). In addition, men have been rated higher than women in physical science disciplines (e.g., Basow & Silberg, 1987; Potvin et al., 2009).

The picture becomes even more complicated when one considers the effect of moderator variables such as student gender, gender role dynamics, and academic discipline. There seems to be a complex interaction between student gender and the gender of the instructor. Although some studies have shown that female students compared to male students tend to indiscriminately give higher ratings in general (e.g., Bachen et al., 1999; Badri et al., 2006; Darby, 2006a; Santhanam & Hicks, 2002), other research indicates a same sex preference for instructors. In several studies, female students gave higher ratings to female instructors while male students preferred male instructors (e.g., Das & Das, 2001; Lueck & et al., 1993; Ory, 2001). However, others found that the same sex preference was limited to female students, while men did not indicate any instructor gender preference (e.g., Bachen et al., 1999; Centra & Gaubatz, 2000).

There is also ample evidence that gender role dynamics between students and instructors can affect SET ratings. Research has shown that female instructors who do not conform to a traditional feminine gender role (i.e. being nurturing, deferring, nice, and relational) tend to be perceived negatively by both male and female students (Bachen et al.,

1999; Basow & Montgomery, 2005; Basow et al., 2006; Bennett, 1982; Martin, 1984). The same might also be true for male instructors who do not behave in traditional masculine ways (Swaffield, 1996). Although these effects occur to some extent across all four possible gender pairings, the influence of gender role stereotypes seems to be most pronounced for the male student / female instructor pairing, with the result that women teachers are held to a higher standard by their male students (Basow et al., 2006; Martin, 1984; Pounder, 2007).

Lastly, the extent of gender and gender role interactions might also be dependent on the academic discipline. For example, Basow and Montgomery (2005) have found that female instructors in the humanities and social sciences were rated higher than male instructors on interpersonal SET dimensions, but the effect was reversed for instructors in the physical sciences (male instructors were rated higher than female instructors on interpersonal characteristics). Overall, however, no consistent effects have been shown across studies with regard to academic discipline as moderator of the gender-SET relation (Bachen et al., 1999; Basow & Montgomery, 2005; Centra & Gaubatz, 2000).

In sum, there is evidence to support a link between instructor gender and SET ratings. However, the extent and the direction of this effect depend on the specific pairing of instructor and student gender, and complex interpersonal dynamics based on gender stereotypes. In addition, there might be a multitude of other factors (e.g., academic discipline) that moderates the gender-SET relation. Another open question concerns the interpretation of the gender-SET link. There is debate about whether this effect is an indication of bias, resulting in discrimination based on gender, or whether there are true gender differences in teaching styles that are accurately reflected in the ratings (Bachen et al., 1999; Centra & Gaubatz, 2000; Statham, Richardson, & Cook, 1991).

Instructor race. There is empirical evidence that ethnic minority instructors have been rated somewhat less favorably than their Caucasian peers on both global and specific SET criteria (Anderson & Smith, 2005; Glascock & Ruggiero, 2006; McPherson & Jewell, 2007; B. P. Smith, 2009; G. Smith & Anderson, 2005). For example, studies have shown that Caucasian instructors obtained slightly higher ratings than Latino/a instructors with regard to competence and caring (Glascock & Ruggiero, 2006). Further, compared to Caucasian women professors, Latina professors were rated higher on interpersonal warmth when their teaching style was lenient and rated lower on this dimension when their teaching style was strict (Anderson & Smith, 2005; G. Smith & Anderson, 2005). Another study found that, although SET ratings for a course with controversial content did not differ by gender or ethnicity, students perceived the course material as more controversial when women and ethnic minority instructors had taught the course (Ludwig & Meacham, 1997).

Instructor sexual orientation. Little research has been conducted on whether the instructor's disclosed sexual orientation has an effect on SET ratings. While one study found no effect (Liddle, 1997), another experimentally controlled study found that a male instructor was rated lower in credibility when he disclosed his homosexual orientation (Russ, Simonds, & Hunt, 2002). In addition, students who attended the lecture in the disclosure condition reported that they had learned less than the students who were in the non-disclosure group with the same instructor.

Instructor personality traits. Research has shown that the instructor's personality can have a substantial effect on SET ratings (e.g., Clayson, 1999; Clayson & Sheffet, 2006; Feldman, 1986; Jackson et al., 1999; Jenkins & Downs, 2001; Pounder, 2007; Shevlin et al., 2000; Sprinkle, 2008; Sweeney, Morrison, Jarratt, & Heffernan, 2009). For example, Clayson

and Sheffet (2006) found that the Big Five traits, as judged by students, independently accounted for between 12% (extraversion) and 55% (conscientiousness) of the variance in SET ratings. Taken together, the five factors accounted for almost 70% of the variance in student ratings. This effect was not significantly modified by any other variable, leading the authors to conclude that students tend to equate perceived instructor personality with teaching effectiveness. Similar effect sizes have been reported by Shevlin et al. (2000), who found that instructor charisma accounted for 37% and 69% of the variance in ratings of instructor ability and overall evaluation of the course, respectively. In addition, as already discussed in the section on the Dr. Fox effect, instructor presentation style (which is partly determined by personality traits, such as being outgoing, positive, and enthusiastic) has resulted in a large effect on SET ratings (e.g., Wachtel, 1998; W. M. Williams & Ceci, 1997). Based on these studies, the most salient traits were those thought to be instrumental in establishing rapport between instructors and students (e.g., being caring, sensitive, approachable, warm, engaging, etc.).

Physical attractiveness. Physical attractiveness has been shown to bias SET ratings of both male and female instructors. In general, more attractive instructors obtained higher student ratings than their less attractive colleagues (Freng & Webber, 2009; Gurung & Vespia, 2007; Klein & Rosar, 2006; Riniolo, Johnson, Sherman, & Misso, 2006). The effect appears to be substantial: For example, Riniolo and colleagues (2006) compared the ratings of professors listed on the public website www.ratemyprofessors.com matched by department and gender. The group of professors who were rated as "hot" by students had obtained SET scores that were on average 0.8 scale points (on a 5-point scale) higher than those of the less attractive group of professors.

Influence of Student-Related Variables on SET Ratings

Student-related factors with no systematic relation to SET scores include student age, grade point average, year in college, and academic ability (e.g., Abrami, 2001; Badri et al., 2006; Braskamp & Ory, 1994; Centra, 2003; Davis, 2009; Marsh & Dunkin, 1992; Ory, 2001). However, some student factors have been found to influence SET ratings; these are student gender, subject interest, nationality, and personality/social style. Student gender has already been discussed in the previous section on instructor-related variables; the following summary will focus on the remaining three variables.

Subject interest. Research has consistently shown a positive correlation between students' interest in the course content and their SET ratings (e.g., Granzin & Painter, 1973; Greimel-Fuhrmann & Geyer, 2003; Marsh & Roche, 1997; Nasser & Glassman, 1997). Marsh and Roche (1997) pointed out, however, that it is not clear whether students' interest existed before taking the course, or whether interest was instilled by the instructor.

Student nationality. One study to date has examined the influence of student nationality on SET ratings (Dolnicar & Grun, 2009). The authors reported that international students had different response styles on SET ratings than domestic students, and the former gave on average harsher evaluations.

Student personality and social style. Little research has been conducted that focuses specifically on the role of students' personality in SET ratings. Munz and Munz (1997) correlated both students' mood trait (positive and negative affectivity) and mood state at the time of SET administration with their SET ratings. The authors found no link between trait variables and SET ratings, but there was a modest positive correlation between mood state at the time of evaluation and SET scores.

Schlee (2005) studied the influence of social style on instructor preferences in a sample of different groups of business majors, namely the people-oriented majors marketing and management in comparison to quantitative areas such as economics, accounting, and finance. The analyses indicated that students in different majors had different social styles: The proportion of students with an expressive social style (characterized by descriptors such as excitable, enthusiastic, stimulating, dramatic, and friendly) was significantly higher in people-oriented majors than in quantitative majors. Conversely, students in quantitative areas were more likely to belong to the "driver" social style (strong-willed, independent, tough, dominating, harsh, and efficient) or to the analytical social style (orderly, industrious, persistent, serious, exacting, and critical). With regard to instructor preference ratings, there was an interaction between the social styles of students and instructors: Students had a tendency to rate instructors higher when they matched their own social style. For example, students with expressive styles appreciated instructors who were responsive and caring, but these characteristics were perceived negatively (as weak and acquiescing) by students with an analytical social style.

Psychosocial Dynamics and SET Ratings

SET ratings also have shown to be influenced by various psychosocial dynamics that play out between students and instructors in the classroom. The three phenomena that have received the most research attention are the Halo effect, the importance of first impressions, and presentation style. Presentation style as potentially biasing variable has already been discussed in the context of students' ability to judge their learning, so the following discussion will focus on the Halo effect and the influence of first impressions on SET ratings.

SET ratings and the Halo effect. The Halo effect, first described by Thorndike (1920), is a cognitive bias that refers to the "tendency to let our assessment of an individual on one trait influence our evaluation of that person on other specific traits" (Blum & Naylor, 1968, p. 200). In the context of SET ratings, the Halo effect refers to the tendency of students to judge their instructor on one characteristic (e.g., whether they like the instructor), which then generalizes to other (unrelated) characteristics (e.g., being knowledgeable about the subject matter). Research has shown that SET ratings are susceptible to the Halo effect. Variables that can influence SET ratings in this way include instructor attractiveness (Freng & Webber, 2009; Gurung & Vespia, 2007; Klein & Rosar, 2006; Riniolo et al., 2006), charisma (Shevlin et al., 2000), dress style (Sebastian & Bristow, 2008), and the perceived personality of the instructor (e.g., Clayson & Sheffet, 2006). Research on multidimensional SET forms has shown that student ratings are similar across all scales, even when they are not logically related (d'Apollonia & Abrami, 1997; Darby, 2007a; Greenwald & Gillmore, 1997; W. M. Williams & Ceci, 1997). For example, Greenwald and Gillmore (1997) conducted a study in which they added three unrelated items to the usual SET form; these were legibility of the instructor's handwriting, audibility of the instructor's voice, and quality of the classroom facilities to aid instruction. Since these variables should affect all students equally, no significant differences between students' ratings of these items were expected. Nonetheless, students' judgment of these items turned out to be influenced by their overall rating of the instructor: students who gave high ratings to the instructor also gave more favorable ratings to the three experimental items.

A study on the Dr. Fox paradigm (W. M. Williams & Ceci, 1997) showed, that, when the instructor delivered his lectures in two different presentation styles (as usual in one

semester and "seductive" in the following semester), the higher ratings in the seductive condition generalized to all other aspects of the course (e.g., the textbook used, assignments given, grading criteria). For example, the quality of the textbook was rated nearly one scale point (on a 5-point scale) higher in the seductive presentation condition than in the control condition.

The importance of first impressions. People's judgment of others is often guided by their first impression. This effect has been repeatedly demonstrated with SET ratings (Ambady & Rosenthal, 1993; Buchert, Laws, Apperson, & Bregman, 2008; Clayson & Sheffet, 2006; Ortinau & Bush, 1987; Tom, Tong, & Hesse, 2010). These studies suggested that students form a lasting impression of the instructor within the first two weeks of class, or even in as little as 30 seconds, which was then reflected in the end-of-semester SET ratings.

Summary: Discriminant Validity and Bias

Sources of potential bias of SET ratings include the educational context, personal characteristics of the instructor and the students, and psychosocial effects. There seems to be consensus in the literature that correlations between these variables and SET ratings exist, and that many of them are large enough to be practically relevant. The issue that remains, however, is the question of whether the correlation of these variables with SET ratings indicates bias (i.e. the relation between the two variables cannot be justified by logical or theoretical considerations), or whether they are meaningful mediators of student learning (e.g., d'Apollonia & Abrami, 1997; Marsh & Roche, 1997; McKeachie, 1997; Pounder, 2007). The inverse relation between class size and SET ratings provides an example of this dilemma: If smaller classes were more conducive to learning (e.g., through more instructor-student interaction, assignments and assessments that require higher-level thinking), class

size would be expected to have a meaningful influence on both student learning and SET ratings. On the other hand, if the same teaching strategies were applied equally to large and small classes, the finding of a significant correlation between class size and SET ratings would be an indication of bias. The same argument has been made for other variables: If certain personality traits of an instructor (e.g., being warm and empathetic) led to more productive instructor-student interactions (e.g., students willing to approach the instructor for help during office hours), these personality traits would not be evidence of bias in a statistical sense, but valid indicators of teaching effectiveness. However, due to the preponderance of non-experimental field research in the SET literature, it is not possible in most cases to distinguish between the two hypotheses.

Consequential Validity

The preceding three sections focused on the evidential basis for the validity of SET ratings, which is concerned with the inferences that can be drawn from empirical research. However, even if the evidential aspect of SET validity were supported, SET ratings could still be used in inappropriate ways, which could lead to adverse social consequences. This is the consequential aspect of validity.

The manner in which SET data are being used and interpreted has been one of the most contentious issues in the literature, and many scholars and instructors have voiced strong opinions in either direction. Therefore, the fourth and final section of this review will address three main question related to the consequences of SET use: What are the attitudes and practices of instructors, students, and administrators regarding SET use at their institution? What are some of the negative outcomes of the reliance on SET data that have

been documented? What are some recommendations that have been voiced in the literature to prevent the misuse of SET data?

Attitudes and Practices Regarding SET Procedures

Several studies (both qualitative and quantitative) have addressed the question of how the SET process is experienced by those affected by it. A survey of this literature paints a picture of ambivalence that mirrors the confusion and lack of clarity that has dominated the empirical research on SET validity. In the following section, attitudes and practices of three groups affected by the SET process will be reviewed; these are the instructors who are being rated, the students who provide the ratings, and the administrators who rely on SET ratings in personnel decisions.

Instructor attitudes and practices. Research indicates that there is little consensus among instructors at a particular institution regarding their attitude towards SET ratings. However, some common themes can be traced across the different studies. Many instructors concede that it is the students' democratic right to have a say in their education, and to voice their opinions about the classes they take (K. Smith & Welicker-Pollak, 2008). Instructors also seem to be more supportive of SET data when they are used for teacher development rather than for summative purposes (Beran & Rokosh, 2009; Yao & Grady, 2005). Although instructors tend to generally acknowledge the potential usefulness of SET data to improve their teaching, few instructors report that they have changed their courses based on SET feedback (Aleamoni, 1999; Beran & Rokosh, 2009; Edstrom, 2008; Nasser & Fresko, 2002; Yao & Grady, 2005). Reasons cited for the reluctance to make use of SET feedback included lack of trust in students' judgment, time demands, and not considering the suggestions as

useful or pedagogically sound (Aleamoni, 1987; Edstrom, 2008; P. M. Simpson & Siguaw, 2000; K. Smith & Welicker-Pollak, 2008; Yao & Grady, 2005).

Instructors seem to be particularly critical of the use of SET ratings for summative purposes. The majority of instructors surveyed were concerned about the use of their ratings by university administration, and many reported feelings of anxiety and powerlessness (Beran & Rokosh, 2009; P. M. Simpson & Siguaw, 2000; Yao & Grady, 2005), and a lack of clarity about expected teaching behaviors and evaluative standards (Johnson, 2000). Other general concerns expressed by instructors concern the reliability and validity of the SET forms, feeling pressured to enter into a popularity contest, and the potential for abuse of the SET process by students as a means of waging revenge (Aleamoni, 1999; Balam & Shannon, 2010; P. M. Simpson & Siguaw, 2000).

Some instructors also admitted to engaging in impression management in order to raise their SET ratings (Crumbley & Reichelt, 2009; Emery, 1995; Pounder, 2007; P. M. Simpson & Siguaw, 2000). These behaviors included lenient grading and lowering of standards, handing out food bribes, administering the SET questionnaire at a strategic point in time (e.g., before a tough exam), putting the students in a good mood right before handing out the SET form, or throwing a party for the students. In a study on a sample of 447 accounting instructors (Crumbley & Reichelt, 2009), 53% of those surveyed said that they knew of other professors who have resorted to such strategies. Empirical research has shown that some of these strategies do indeed work, which poses a threat to the validity of the ratings. For example, Youmans and Jee (2007) demonstrated that handing out chocolate at the time of the evaluation increased the average rating by about 0.2 points on a 5-point scale compared to the control group, and research reported by Fortunato and Mincy (2003) showed

that a positive mood induction right before SET administration significantly increased the ratings.

Student attitudes. How students experience the SET process has also been investigated, albeit to a lesser extent. The available research shows that the majority of students seem to value the opportunity to rate their instructors (Chen & Hoshower, 2003; Greimel-Fuhrmann & Geyer, 2003; Soyjka, Gupta, & Deeter-Schmelz, 2002). Students reported that they are primarily motivated to participate in the evaluation process based on the expectation that their input will lead to positive changes in the course (Chen & Hoshower, 2003). However, students have also criticized aspects of the SET process. For example, students were concerned that instructors do not take their feedback seriously, and they voiced frustration about not seeing any changes. In addition, they wished they would have more opportunities for input early in the semester rather than at the end (Wachtel, 1998).

Administrator attitudes. Very little research has been conducted on how administrators view the SET process and its role in personnel decisions. However, administrators, on average, seem to have a more positive attitude towards the use of SET data than course instructors (Aleamoni, 1999; McMartin & Rich, 1979; Sproule, 2000). Administrators tend to favor single global ratings of teaching effectiveness, since they limit the amount of data that needs to be taken into account. Further, they consider it important to give students a sense of involvement in personnel decisions, and to give them information on which they can base the selection of their courses. Finally, administrators often view SET ratings as a measure of student satisfaction.

Misuse of SET Data and its Consequences

Instructors are frequently concerned about how SET data are used, especially in the context of high-stakes decisions such as tenure and promotion. The literature is replete with complaints about how the SET process has been implemented and the negative consequences that resulted from it. The literature on SET misuse broadly falls into two categories: issues concerning data interpretation, and the effect that SET misuse can have on work satisfaction and performance of instructors.

Data interpretation. One major complaint voiced by instructors is the administrator's lack of knowledge concerning the psychometrically sound interpretation of SET data (e.g., Algozzine et al., 2004; Franklin, 2001; McKeachie, 1997; Wachtel, 1998). Franklin (2001), based on a multi-institutional study, reported that over 50% of the committee members using SET data were not able to answer basic questions about the meaning of statistics (e.g., means and standard deviations) typically included in SET reports. Questionable practices of SET interpretation reported in the literature include: a) Reduction of multidimensional SET data to a single meaningless indicator of performance by averaging across characteristics that should be evaluated separately (Ruskai, 1996); b) creation of fine-grained categories of performance that are not warranted by the imprecision of the instrument (Algozzine et al., 2004; McKeachie, 1997; Ory & Ryan, 2001); c) Arbitrary specification of cutoff scores to denote different categories of performance (Franklin, 2001); d) Use of raw mean scores to compare instructors of different courses against each other without taking into account extraneous factors (e.g., class size, academic discipline, etc) that are linked to SET ratings (e.g., Algozzine et al., 2004; Badri et al., 2006; Dolnicar & Grun, 2009; R. D. Simpson, 1995); and d) Failure to take into account small sample sizes, low response rates,

and spread and shape of the score distribution (Darby, 2007b; Franklin, 2001; Wolfer & Johnson, 2003).

Consequences of SET misuse. At many institutions, SET data are the only measure of instructor teaching effectiveness (e.g., d'Apollonia & Abrami, 1997; Franklin, 2001; Pounder, 2007). Therefore, due to the high stakes involved, inappropriate use of SET data has been shown to contribute to instructor anxiety, low morale, and job dissatisfaction (Ory & Ryan, 2001; Wachtel, 1998; Yao & Grady, 2005). Some instructors also react to the pressure by engaging in impression management (see section on instructor attitudes), by lowering standards and inflating grades, or by disregarding SET feedback altogether (e.g., Crumbley & Reichelt, 2009; Redding, 1998; Wachtel, 1998).

Recommendations for the appropriate use of SET data. Fortunately, many scholars, administrators, and instructors have recognized the problems associated with the use of SET data, and they have provided recommendations to alleviate some of the issues of contention. Most of these recommendations have focused on the consequential validity aspect, since policies are easier to change than the underlying conceptual and psychometric problems. The following suggestions have been empirically shown to improve SET validity a) Use of properly constructed and validated questionnaires to gather SET data (Richardson, 2005; Sedlmeier, 2006); b) Implementation of psychometrically sound and standardized evaluation procedures across academic units or the institution as a whole based on input from all affected parties (e.g., Edstrom, 2008; Hoyt & Cashin, 1977; Pounder, 2008; Ruffer, McMahon, & Rogers, 2001) c) Development of ethical guidelines for SET use (McCormack, 2005); d) Providing training to students on how to give meaningful feedback (Svinicki, 2001) and training of administrators and committee members in proper SET data interpretation

(e.g., Algozzine et al., 2004); e) Offering consultation to instructors to help them improve their teaching (Ballantyne, Borthwick, & Packer, 2000; Marsh & Roche, 1993; Stevens & Aleamoni, 1985); f) Control of biasing variables in SET ratings over which the instructor has no influence (d'Apollonia & Abrami, 1997; Franklin, 2001); g) Use of multiple means of evaluation, such as standardized exam scores, portfolios, and narrative reflections of instructors on their own teaching (J. S. Armstrong, 1998; Davis, 2009; Franklin, 2001; Peterson & Stevens, 1998; Zayani, 2001).

Summary: Consequential Validity

Both instructors and students have somewhat ambivalent feelings towards the SET process. Students generally are motivated to give feedback with the goal of improving the quality of the course. However, they also express frustration about the extent to which instructors have implemented changes based on their feedback. Conversely, instructors often report concern about the use of SET scores by administrators and the potential consequences for their career. Some instructors deal with this pressure by displaying a cynical attitude towards the SET process, and by discounting the feedback they receive. Others might resort to impression management strategies (e.g., easy grading) to raise their SET scores in the hope of being able to "game the system".

Research has shown that many of the concerns voiced by faculty and students are justified. Administrators are often not aware of the complexity of the issue, and they frequently lack the training necessary to make meaningful inferences based on SET data. However, researchers and those directly affected by the SET process have started to recognize and address questionable SET practices with the goal of enhancing the validity and fairness of the process.

Critical Evaluation of the SET Literature and Recommendations for Future Research

Summary and Critique

The goal of this review was to summarize research on the validity of the student evaluation of teaching process. Four major aspects of validity have been discussed: construct validity, convergent validity, discriminant validity and the related issue of bias, and consequential validity. Considering the evidence, one can conclude that there is only limited support for the validity of SET procedures in their current form. The overarching theme across the different areas of research is a lack of consensus and clarity. There is no agreement on how to define teaching effectiveness, the construct that SETs claim to measure. There is a paucity of theoretical frameworks that could guide the development of SET measures and provide a foundation for answering the question of the dimensionality of teaching effectiveness. Many SET surveys have never been psychometrically evaluated, and they too frequently exhibit major flaws in item wording and scale construction.

If SET ratings reflect teaching effectiveness, they should be related to how much students learn. The available research, however, has only yielded a moderate correlation between SET scores and objective and subjective indicators of student learning. In addition, there is debate about the extent to which students are able to accurately assess their learning and make judgments about the quality of instruction they receive. Further, a myriad of extraneous variables have been found to systematically relate to SET scores, which indicates that SET ratings might be biased and susceptible to manipulation. However, in many cases it is not possible to decide whether the effect of third variables indicates bias, or whether the relations can be meaningfully interpreted as mediators of student learning. Finally, there is ample evidence that SET data have been used inappropriately by university committees in

high-stakes personnel decisions, which has led to negative consequences regarding instructors' integrity and acceptance of the SET process.

There are several methodological and conceptual problems that have limited the inferences that can be drawn from the available SET research. First, a large part of SET research has been non-experimental in design, involving post-hoc analyses of actual SET data gathered at universities as part of the standard SET process. Although this type of research has high ecological validity, the conclusions derived from this body of research are limited; in many cases there was no control for confounding variables in order to rule out alternative explanations for the findings. This type of problem is ubiquitous in the SET literature, especially in studies on discriminant validity and bias. For example, in order to determine whether a difference in SET ratings for male and female instructors indicates bias based on the influence of gender stereotypes or whether this finding correctly reflects a true difference in teaching styles, it is necessary to control for specific teaching behaviors across instructors.

Second, there is considerable inconsistency across different studies investigating the same topic area, both with regard to the direction and the magnitude of the effect. As discussed in the section on the influence of instructor gender on SET ratings, some studies have yielded higher SET scores for men while others found higher ratings for women, or no gender effect at all. There are a number of possible methodological reasons for the pronounced inconsistency of findings across studies on SET ratings. Since most SET research is conducted in field settings, there are many contextual factors that render it difficult to make meaningful comparisons. Studies greatly vary in terms of the educational context in which they have been conducted, including student population (e.g., Freshmen vs.

graduate students), type of institution (e.g., community college vs. private liberal arts college), geographical location and country, time period (SET research has been systematically conducted since the 1970s), academic discipline, and SET measures used. Further, most variables thought to be linked to SET ratings have been examined in isolation without taking into account other variables that might substantially alter the relation between the target variables. Thus, the inconsistency of the results across studies suggests the existence of a multitude of mediator or moderator variables that, if taken into account, can help to interpret the seemingly contradictory findings. So far, the only instance in which several variables have been explored in conjunction with each other concerns the link between gender and SET ratings. Here, the explicit inclusion of student gender and gender role stereotypes has helped to elucidate the dynamics behind the apparent contradictions in the research findings.

Recommendations for Future Research on SET Validity

A number of recommendations can be made to address the aforementioned shortcomings of the SET research base. First, there is a need for theory-driven research. Most SET research has been atheoretical, inductive, and merely descriptive. Drawing on existing educational and psychological theories will help to better define and operationalize the construct of teaching effectiveness, and to address the question of dimensionality. In addition, a sound theoretical framework can guide the development of SET measures, and it allows the formulation of hypotheses regarding the relation of various variables that then can be tested empirically.

Second, there is the need for experimentally controlled research. By manipulating variables of interest, and controlling for possible extraneous variables it will be possible to

gain a better understanding of the causal relation between variables and rule out alternative explanations.

Third, there is a need for the comprehensive evaluation of sets of variables that goes beyond the descriptive level. By assessing multiple variables simultaneously, it will be possible to detect possible moderators and mediators that systematically alter the relation between the target variables.

Finally, more research needs to be conducted on alternative means of evaluating teaching effectiveness. One reason for the continued use of SET ratings in their current form is the lack of alternatives to the traditional SET process.

In summary, despite the substantial body of research on student evaluations of teaching, the evidence for the validity of the current practices is weak. However, it is unlikely that use of SET ratings to evaluate instructor teaching effectiveness will be discontinued in the near future. Therefore, researchers need to continue their efforts to improve the validity of the process, and users of SET data need to be aware of their limitations.

CHAPTER 3: METHODS

Participants

Participants (n = 610) were students from a large Midwestern university who received experimental credit applied to their introductory psychology courses as compensation for their participation. The participant sample was representative of the range of academic majors offered at the university. The sample included 372 women and 238 men with a mean age of 19.8 years (SD = 2.6 years). With respect to year in college, 44.6% of students were freshmen, 27.2% sophomores, 18.0% juniors, 9.5% seniors, and 0.5% graduate students; 0.2% did not indicate their year in college. The participants reported the following racial-ethnic identities: White (84.8%), African American (2.0%), Latino/Hispanic (1.6%), Asian American (3.6%), Native American (0.3%), or other (e.g., biracial, 3.0%); 4.8% of students were international students.

Measures

The measures used in this study are included in the Appendix. The four instructor vignettes and the SET rating scale can be found in Appendix A. All remaining measures are given in the order in which they are presented to participants in Appendix B.

Instructor Vignette and Rating Scale

Instructor vignette. The vignette that students were asked to rate described the academic and professional background of a hypothetical instructor. The vignette included information that students typically have access to (e.g., through the department website), such as academic credentials, courses taught, research interests, and membership in professional associations. The two variables that were manipulated were instructor gender ("Dr. Robert Smith" vs. "Dr. Roberta Smith") and academic specialty (counseling

psychology vs. statistics/research methods). The wording of the description was adjusted accordingly (e.g., replacing "American Psychological Association" by "American Statistical Association").

The overall instructor description was formulated in fairly general terms. This was based on the rationale that people tend to rely more on their biases and stereotypes to guide their evaluation of others when the context in which a judgment is made is ambiguous (e.g., Bargh, Chen, & Burrows, 1996; Operario & Fiske, 2004; Plous, 2003). Hence, the experimental variables (instructor gender and academic specialty) were expected to become more salient in students' evaluation of the instructor when all other instructor characteristics remained relatively nondescript, hereby enhancing the magnitude of the expected effects.

Instructor rating scale. Students were asked to read the vignette and to imagine that they are considering taking a course with this instructor in the respective academic discipline (counseling or research methods). Students were then prompted to indicate to what extent they would expect the instructor to perform on eight dimensions of teaching effectiveness. These were knowledge about the course subject, organization/preparedness, availability for help outside of class, enthusiasm about the course subject, effectiveness in communicating course objectives and requirements, creation of a respectful and comfortable classroom environment, fairness/accommodation of individual differences, and interest in helping students learn. The rating scale was a 7-point Likert scale ranging from 1 = *not at all* to 7 = *extremely*.

The eight SET dimensions were chosen to represent the content of typical SET forms administered at universities in the US. Specific items were selected based on the available literature on teaching effectiveness (Bain, 2004; Barnes et al., 2008; Marsh & Dunkin, 1992),

and a survey of existing SET instruments used at various institutions. Both stereotypically male (e.g., knowledge and organization) and stereotypically female (e.g., creation of a respectful and comfortable classroom environment) items were included in the present study. The estimated internal consistency of the 8-item scale was high, with $\alpha = .903$, 95% CI [.891, .914]. Therefore, a single score representing the SET rating, namely the average score across all eight items, was used in all further analyses.

Basic Interest Markers

The Basic Interest Markers (BIM; Liao, Armstrong, & Rounds, 2008) measure domain-specific vocational interests. The BIMs are available for free in the public domain, and they have been developed specifically for research on college students. Overall, there are 31 scales with a total of 338 items. Three of these 31 scales were used in the present study to assess interest in the two academic specialties manipulated in the instructor vignette. Two BIM scales (social service and social science) were chosen to represent the counseling psychology course. Since no scale was available to measure interest in statistics, the mathematics scale was adapted to measure interest in statistics by making minor wording changes. The statistics and social science scales each have ten items, and the social service scale has 12 items. Each item consists of a short phrase describing an activity (e.g., "Learn about human behavior" or "Organize a social support group"). Participants indicated the extent to which they would like to do each activity by responding on a 5-point Likert scale ranging from 1 = *strongly dislike* to 5 = *strongly like*; higher scores indicated a higher interest in the activity.

Reliability. Internal consistency estimates based on the norming sample of 545 college students ranged from .85 to .95 across the 31 scales (Liao et al., 2008). The values

reported by Liao et al. for the three scales used in the present study were $\alpha = .95$ for mathematics (changed to statistics in the present study), $\alpha = .91$ for social science, and $\alpha = .93$ for social service. In the present study estimates for these three scales were $\alpha = .942$ (95% CI [.935, .949]) for statistics, $\alpha = .880$ (95% CI [.865, .894]) for social science, and $\alpha = .932$ (95% CI [.924, .940]) for social service, respectively.

Validity. Convergent validity has been demonstrated by correlating the BIMs with matched Basic Interest Scales (BISs) from the 1994 and 2005 editions of the Strong Interest Inventory. The correlation of the mathematics BIM with the mathematics BIS was $r = .60$ (1994 edition) and $r = .68$ (2005 edition), respectively. The social service BIM correlated $r = .61$ with the social service BIS (1994), and $r = .69$ with the counseling and helping BIS of the 2005 edition. The correlation of the social science BIM with the social science BIS in the 2005 edition was $r = .63$.

Further convergent validity evidence for the BIM scales was shown by using discriminant function analyses to predict membership in 12 academic major areas (Liao et al., 2008). The 31 BIM scales accounted for 95.1% of the variance in the 12 major categories, and group membership was correctly predicted 63.4% of the time (chance hit rate was 8.33%).

Alternate Forms Public Domain Interest and Confidence Markers

The activity-based scales from the Alternate Forms Public Domain (AFPD) Interest and Confidence Markers (P. I. Armstrong, Allison, & Rounds, 2008) were used to measure interest and confidence in each of Holland's (1959, 1997b) six vocational types, which are realistic, investigative, artistic, social, enterprising, and conventional (RIASEC). Each RIASEC scale consists of eight items selected from the scales in the Interest Profiler (Lewis

& Rivkin, 1999), resulting in a total of 48 items for each alternate form (Form A and Form B). Each item consists of short descriptions of various occupational activities (e.g., "Work in a biology lab" or "Write books or plays").

Form A was used to assess participants' interest in the six RIASEC domains. Participants indicated the extent to which they liked to do the described activities by using a 5-point Likert scale ranging from 1 = *strongly dislike* to 5 = *strongly like*. Higher scores indicated higher levels of interest. Form B was administered to measure participants' confidence in being able to perform the activities pertinent to each RIASEC domain. Participants rated their confidence on a 5-point Likert scale ranging from 1 = *very low confidence* to 5 = *very high confidence*. Therefore, higher scores represented more confidence in being able to perform the respective activity.

Reliability. Armstrong et al. (2008) reported that the internal consistency estimates for the activity-based AFPD *interest* scales ranged from .79 to .94 (mean α = .88). In the current study, internal consistency coefficients for the AFPD RIASEC interest scales (Form A) were comparably high: For realistic interest α = .928 (95% CI [.919, .936]); for investigative interest α = .895 (95% CI [.882, .907]); for artistic interest α = .861 (95% CI [.844, .877]); for social interest α = .835 (95% CI [.814, .854]); for enterprising interest α = .841 (95% CI [.821, .859]); and for conventional interest α = .905 (95% CI [.893, .916]).

For the six activity-based AFPD *confidence* scales Armstrong et al. (2008) reported coefficient alpha estimates in the range between .85 and .94 (mean α = .89). Based on the sample in the present study internal consistency estimates for the AFPD RIASEC confidence scales (Form B) were as follows: For realistic confidence α = .928 (95% CI [.919, .936]); for investigative confidence α = .922 (95% CI [.912, .930]); for artistic confidence α = .855

(95% CI [.837, .872]); for social confidence α = .895 (95% CI [.882, .907]); for enterprising confidence α = .885 (95% CI [.871, .898]); and for conventional confidence α = .914 (95% CI [.903, .924]).

Validity. Convergent validity of the activity-based AFPD Markers was established by correlating each RIASEC scale with the corresponding scale of the 1994 edition of the Strong Interest Inventory. Correlations between the matching scales ranged from .56 to .72 with a mean correlation of r = .64 (Armstrong et al., 2008). Further, structural analyses of the AFPD Markers supported the order predictions in Holland's (1997) RIASEC model. Armstrong and Vogel (2009) reported that interest-confidence correlations for the RIASEC types measured by the AFPD activity scales ranged from .60 to .72 (mean r = .70). These interest-confidence correlations were consistent with those of established commercial RIASEC interest and confidence measures (Rottinghaus, Larson, & Borgen, 2003), thereby providing validity evidence for the administration format used in the current study.

Gender Attitude Inventory

The Gender Attitude Inventory (GAI; Ashmore, Del Boca, & Bilder, 1995) is a multidimensional inventory derived from an inter-group relations perspective that measures "attitudes toward the multiple objects that organize college students' thoughts and feelings about sex and gender" (Ashmore et al., 1995, p. 753). The GAI has been specifically developed to assess gender attitudes in college students. The 109-item inventory has 14 primary scales. Items from three of these scales (traditional stereotypes (ten items), family roles (11 items), and differential work Roles (nine items) were relevant to the objective of the present study; therefore, only these 30 items were administered to participants.

The items consist of short statements such as "All occupations should be equally accessible to both men and women" or "The wife should have primary responsibility for taking care of the home and children". Participants indicated the extent to which they agree to each statement by responding on a 5-point Likert scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*. About half of the items were positively keyed, the other half negatively keyed. Items were scored so that higher scores indicated a more traditional gender role attitude. The mean score across all 30 items was calculated, yielding a score range of 1 to 5.

Reliability. Internal consistency estimates for the three scales used in the present study have been reported by the authors as follows (Ashmore et al., 1995): α = .83 - .86 for the traditional stereotypes scale; α = .79 - .89 for the family roles scale; and α = .84 - .87 for the differential work roles scale. In the present study the internal consistency estimate for the combined 30-item scale was α = .93 (95% CI [.92, .94]).

Three-week test-retest reliability was assessed in the initial sample by Ashmore et al. (1995). The stability estimates for the three scales used in the present study were reported as r = .83 - .89 for the traditional stereotypes scale, r = .75 - .78 for the family roles scale, and r = .80 - .85 for the differential work roles scale.

Validity. Ashmore et al. (1995) provided several pieces of evidence for the validity of the GAI scales. Based on the inter-group relations theory that underlies the construction of the GAI it was predicted that men, as the beneficiaries of the status quo, would endorse more traditional gender role attitudes than women. This prediction was supported for all three scales used in the present study. The authors further reported convergent validity evidence: all three GAI scales had moderate to high correlations with the Attitudes Toward Women Scale (a global sex-role ideology scale) and moderate correlations with conservatism. In

addition, the finding that the GAI scale scores were not significantly related to a measure of social desirability provided discriminant evidence for the validity of the GAI.

International Personality Item Pool Big-Five Markers

The International Personality Item Pool (IPIP) Big-Five Markers (Goldberg, 1999; Goldberg et al., 2006) are a broad-bandwidth measure of the Big Five personality traits neuroticism, extraversion, openness, agreeableness, and conscientiousness. The IPIP Markers have been developed as a proxy of Costa and McCrae's (1992) Revised NEO Personality Inventory. The measure is intended primarily for research purposes and is available for free in the public domain. In the present study, the 50-item version of the measure was used. Each of the five scales has ten items, which consist of short descriptive phrases such as "Am always prepared" or "Feel little concern for others". About half of the items are positively keyed, the other half are negatively keyed. Participants were asked to rate each statement in terms of how accurately it describes them on a 5-point Likert scale ranging from 1 = *very inaccurate* to 5 = *very accurate*. The items were scored so that higher scores indicated a stronger endorsement of the respective personality trait.

Reliability. Internal consistency estimates for the 50-item measure reported across several validation studies range from α = .74 - .90 (median α = .89) for extraversion, from α = .78 - .85 (median α = .83) for agreeableness, from α = .79 - .89 (median α = .80) for conscientiousness, from α = .80 - .93 (median α = .88) for neuroticism, and from α = .78 - .90 (median α = .85) for openness (Ehrhart, Roesch, Ehrhart, & Kilian, 2008; Goldberg, 1999; Gow, Whiteman, Pattie, & Deary, 2005; Lim & Ployhart, 2006; Zheng et al., 2008).

In the present sample, internal consistency estimates were α = .887 (95% CI [.873, .900]) for extraversion, α = .849 (95% CI [.830, .866]) for agreeableness, α = .794 (95% CI

[.769, .817]) for conscientiousness, α = .866 (95% CI [.849, .881]) for neuroticism, and α = .804 (95% CI [.780, .826]) for openness.

Validity. The hypothesized 5-factor structure of the 50-item measure has been confirmed empirically with samples from the United States (Ehrhart et al., 2008; Lim & Ployhart, 2006), Scotland (Gow et al., 2005), Croatia (Mlacic & Goldberg, 2007), and China (Zheng et al., 2008). Further, the factor structure holds up across gender and different ethnic groups within the United States (Ehrhart et al., 2008), and different age groups (Gow et al., 2005).

Evidence for convergent validity has been obtained by demonstrating that the five broad IPIP domains correlate highly with other inventories of the Big-Five traits. Gow et al. (2005) and Lim and Ployhart (2006), respectively, reported similar correlations between the IPIP measure and the original NEO-PI-R: For extraversion r = .69 and r = .69; for agreeableness r = .49 and r = .50; for conscientiousness r = .76 and r = .72; for neuroticism r = .83 and r = .76; and for openness r = .79 and r = .71. Zheng and colleagues (2008) obtained the following correlations of the IPIP Big-Five Markers with the Big-Five Inventory: For extraversion r = .72; for agreeableness r = .47; for conscientiousness r = .67; for neuroticism r = .70; and for openness r = .59. Further, mean score differences on the five dimensions between men and women (Ehrhardt et al., 2008), and different age groups (Gow et al., 2005), respectively, have been shown to be consistent with theoretical prediction and prior empirical findings.

Procedure

Participants completed the survey through a commercial data collection site online (www.surveymonkey.com). Data collection occurred over the course of two semesters. After giving consent, students were randomly assigned to one of the four vignette conditions based on their month and year of birth (odd vs. even). Each participant first completed the demographic section of the questionnaire, followed by the instructor ratings for their respective vignette. Participants then completed the remaining scales in the following order: Basic Interest Markers (BIM) for statistics, social sciences, and social services, Alternate Forms Public Domain (AFPD) Interest Markers, Gender Attitude Inventory (GAI), AFPD Confidence Markers, and International Personality Item Pool (IPIP) Big Five Markers. Within each measure, items were presented to each participant in a different random order. Participants completed the entire questionnaire in approximately 50 minutes. The study has been approved by Iowa State University's Institutional Review Board (IRB). The consent form and the study posting form can be found in Appendix C and Appendix D, respectively.

CHAPTER 4: RESULTS

The following chapter is conceptually organized based on the three main research questions posed in the introduction. For each research question, the experimental design and statistical analyses will be presented, followed by the results.

Course Type, Instructor Gender, and Student Gender as Moderator

The first analysis was conducted to find out whether there were any systematic differences in students' SET ratings across course type and instructor gender, the two variables manipulated in the instructor vignette. In addition, the replication of the same-sex preference for instructors documented in the prior literature (student gender as moderator of the effect of instructor gender) was of particular interest. There were three independent variables (two levels each) in the analysis performed to address these questions: instructor gender (male vs. female), course type (counseling psychology vs. research methods), and student gender (male vs. female). The first two variables were experimental variables, while student gender was a classification factor. All three independent variables were between-subjects variables, yielding a total of eight conditions. The dependent variable, SET rating, was defined as the average score across the eight items that represented the rating of the perceived quality of the instructor depicted in the vignettes. Descriptive statistics (means and standard deviations) for all eight conditions are shown in Table 1.

In order to test for mean score differences in SET ratings across the levels of the three independent variables, a 2 (course type) x 2 (student gender) x 2 (instructor gender) Analysis of Variance (ANOVA) was performed. The ANOVA yielded a significant main effect of student gender with $F(1, 602) = 19.4$, $p < .001$, $\eta^2 = .03$, 95% CI for the difference in means = [0.46, 0.18]. Female students on average rated the instructor depicted in the vignettes

significantly more favorably (M = 6.00 points, SD = 0.86) than male students (M = 5.64 points, SD = 0.86) irrespective of course type and instructor gender, but the effect was small (J. Cohen, 1988). The main effects of instructor gender and course type were not statistically significant. For instructor gender, $F(1, 602) = 0.13$, $p = .72$, 95% CI for the difference in means = [0.18, -0.10]; for course type: $F(1, 602) = 0.01$, $p = .93$, 95% CI = [0.17, -0.11]. Therefore, there was no evidence that the instructor when described as teaching counseling psychology was rated differently than the instructor teaching research methods. Likewise, the gender of the instructor did not seem to matter in terms of how she or he was perceived by students with regard to teaching effectiveness. In fact, the mean rating scores across the different levels of instructor gender and course type were nearly identical (see Table 1), and the magnitude of the confidence intervals for the estimated mean differences was very narrow.

Further, there was no indication that students showed a same-sex preference in terms of their SET ratings, since the instructor gender by student gender interaction remained insignificant [$F(1, 602) = 0.76$, $p = .38$]. Likewise, none of the remaining interaction terms reached statistical significance [$F(1, 602) = 1.8$, $p = .18$ for the student gender by course type interaction, $F(1, 602) = 0.2$, $p = .64$ for the instructor gender by course type interaction, and $F(1, 602) = 0.5$, $p = .47$ for the 3-way interaction].

Effect of Student Individual Differences on SET ratings

The second major goal of this study was to identify student individual differences that are systematically related to SET ratings. Considering that the SET ratings across the four vignette conditions were virtually identical (the confidence intervals for the difference in means were very narrow and close to zero), ratings across the four conditions were collapsed

into a single category for the remaining analyses. The results will be presented in the following order: First, two sets of multiple regression analyses (all predictors entered simultaneously vs. predictors entered separately by construct) were performed to evaluate the relations between student individual differences and SET ratings. Then, moderation analyses were conducted to test whether the relations between the predictors and SET ratings differed depending on instructor gender, student gender, or course type.

In all analyses, the following student individual difference variables were tested in terms of their relation to students' SET ratings: a) the Basic Interest Markers (three domains, namely social science, social service, and statistics); b) the AFPD RIASEC Interest Markers (six domains, which are realistic, investigative, artistic, social, enterprising, and conventional); c) the AFPD RIASEC Confidence Markers (six domains, which are realistic, investigative, artistic, social, enterprising, and conventional); d) the GAI total score; and e) the IPIP Big-Five Markers (five traits, namely extraversion, agreeableness, conscientiousness, neuroticism, and openness).

Table 2 shows the bivariate correlations between students' SET rating and all individual difference measures. These correlations suggest that several individual differences might be systematically associated with students' SET scores. For both male and female students, the traits that had the strongest correlations with SET scores were agreeableness (r = .29 for both men and women), conscientiousness (r = .21 for men, and r = .20 for women), and gender role attitudes (a more traditional gender role attitude was associated with lower SET scores; r = -.18 for men and r = -.26 for women).

The relation between SET scores and all student individual differences was formally assessed through a series of multiple regression analyses with the SET rating as the criterion,

and the respective individual difference variables as the predictors. The regression analysis was performed in two different ways in order to address the concern of collinearity among the groups of predictors. In particular, the correlations between the six RIASEC interest domains with their respective confidence domains were substantial (*r* ranged between .43 and .76 for the six interest-confidence pairings, see Table 2), suggesting a moderate degree of overlap between these two constructs, which has also been documented elsewhere (P. I. Armstrong & Vogel, 2009; Rottinghaus et al., 2003).

In the first analysis, all predictors were entered simultaneously to predict the criterion. This approach has the advantage that the correlations between the predictors (e.g., interest and confidence) are fully taken into account (J. Cohen, Cohen, West, & Aiken, 2003). Therefore, the partial effects of the constructs involved after correcting for multicollinearity can be seen as a cleaner measure of the respective construct. However, the downside of this approach is the difficulty of interpreting the meaning of the regression coefficients after partialing out the contributions from the overlapping constructs. Therefore, a second set of analyses was run to increase the interpretability of the regression coefficients. Here, predictors were entered separately by construct. For example, all five personality traits were entered as predictors simultaneously in one analysis. Then, a separate analysis was conducted with all six RIASEC interest types as predictors, etc. Although this method makes it easier to conceptually interpret the relations between the predictors and the criterion, the disadvantage of this approach is that it assumes independence of the constructs (e.g., that domain-specific interests and confidence have no shared variance, which overlooks the empirical evidence supporting a moderate relation between these two constructs). Since both approaches have

advantages as well as shortcomings, the results from both types of analyses will be presented, and any discrepancies will be addressed as part of the discussion section.

Multiple Regression with all Predictors Entered Simultaneously

In the first multiple regression analysis, all predictors were entered simultaneously to predict SET ratings. Regression coefficients and inferential statistics are shown in Table 3. The overall regression model was significant with $F(21, 588) = 7.7$, $p < .001$, $R^2 = .22$ (95%CI = [.137, .254), meaning that student individual differences explained 22% of the variance in SET ratings (between 13.7% and 25.4% as applied to the population, when taking into account the limits of the confidence interval at a level of confidence set at 95%); this corresponds to a medium to large effect (Cohen, 1988). Six out of the 21 independent variables entered significantly (at $p < .01$) contributed to the prediction of SET ratings (see Table 3): Investigative interest, conventional confidence, conscientiousness, and agreeableness were positively related to SET ratings, while investigative confidence was negatively related to the criterion. In addition, a more traditional gender role attitude was associated with lower SET ratings.

Multiple Regression with Predictors Entered Separately by Construct

In the second set of multiple regressions, the predictors were entered separately by construct. Therefore, five analyses were performed overall with the following groups of variables as predictors: 1) The three BIM scales; 2) the six RIASEC interest types; 3) the six RIASEC confidence types; 4) the five personality traits; and 5) the GAI score. Statistics for the overall regression models as well as the regression coefficients are displayed in Table 4. All regression models were statistically significant at $p < .001$. The respective constructs explained between 3.4% (95% CI = [0.9%, 6.4%], BIM scales) and 12.4% (95%CI = [7.3%,

17.0%], Big Five personality traits) of the variance in SET ratings. The predictors that contributed significantly (p <.01) to these results were the following: Social and conventional confidence, agreeableness, and conscientiousness were positively related to SET ratings, while interest in statistics and realistic activities, as well as realistic and investigative confidence showed a negative association with students' SET scores. Likewise, a more traditional gender role attitude was associated with lower SET ratings.

Instructor Gender, Student Gender, and Course Type as Possible Moderators of the Individual Difference - SET Score Relation

The results of the regression analyses described above indicated that multiple individual difference variables are systematically linked to students' SET ratings. Additional analyses were conducted to test whether the direction and magnitude of the observed relations between the predictors and the SET criterion were different for male students vs. female students, male instructors vs. female instructors, and counseling psychology vs. research methods. For example, it might be possible that interest in statistics would be correlated with ratings of the two vignettes featuring the statistics instructor, but not with the ratings of the counseling instructor vignettes. Likewise, students' gender role attitudes might be salient for the female instructor but not for the male instructor. In these cases, the three dichotomous variables instructor gender, student gender, and course type would serve as moderators of the individual difference – SET score relation. The hypothesis of differential correlations across conditions was formally tested by specifying a regression model that contained both the effect-coded independent variables and their respective interaction terms as predictors of SET ratings. The interaction terms were created by computing the products of the centered continuous predictors with the respective moderator variable. All continuous

predictors were centered in order to reduce multicollinearity among the main effects and the interaction terms (J. Cohen et al., 2003; Frazier, Tix, & Barron, 2004). The basic regression model was specified as follows:

y = a + b1*x1 + b2*x2 + **b3*x1*x2** + e

y = predicted SET score; a = constant; b1-b3 = regression coefficients; x1 = individual difference variable (e.g., interest in statistics, gender role attitude, etc.); x2 = moderator (instructor gender, student gender, or course type; e = error term; the term printed in bold font is the interaction term. Therefore, this model allowed to test both the main effects and the possibility of differential individual difference – SET relations across student gender and vignette condition as shown by a significant interaction term.

Due to the exploratory nature of these analyses, all possible combinations of the three moderators with each individual difference variable were tested. As a result, the interaction term was not statistically significant ($p > .05$) in any of the regression analyses. Therefore, there is no evidence that the magnitude or direction of the correlation between the individual difference variables and students' SET rating is different for male or female students, male or female instructors, or the type of course taught.

Gender Difference in SET Ratings: Mediating Effect of Individual Differences

The ANOVA performed to test the first research hypothesis yielded a significant but small gender difference in SET ratings [$F(1, 602) = 19.4$, $p < .001$, $\eta^2 = .03$], with female students giving higher mean SET ratings than male students. As shown in Table 5, there were also significant gender differences on most of the individual difference variables. This suggests that the gender effect for SET ratings might be explained by gender differences on individual difference variables linked to SET ratings. To test this hypothesis, a series of

Analyses of Covariance (ANCOVAs) was performed. In the ANCOVA, student gender was the independent variable, and the mean SET rating the dependent variable; the individual difference variables were treated as the covariates. First, separate analyses by construct were performed (e.g., all personality traits entered together in one analysis, all RIASEC interests entered together in the next analysis, etc.) to assess the relative contributions of each construct. This procedure was followed by an additional analysis in which all covariates were entered simultaneously to obtain the overall effect the covariates on the magnitude of the gender difference. A mediation effect occurs if the magnitude of the gender effect when obtained without controlling for student individual differences is greatly reduced (partial mediation) or eliminated (full mediation).

The results can be summarized as follows: All student individual differences mediated the gender effect in SET ratings, albeit to varying degrees. The strongest effect was obtained when controlling for the six RIASEC interests. Here, the gender effect was reduced below statistical significance [$F(1, 596) = 1.2$, $p = .29$, $\eta^2 = .002$]. Therefore, the gender difference in SET ratings was fully explained by gender differences in RIASEC interests. For the three specific interest domains (social science, social service, and statistics), the reduction of the gender effect was less pronounced [$F(1, 599) = 6.9$, $p = .01$, $\eta^2 = .01$]. Nonetheless, the gender differences in these three interest areas partially explained the gender effect with regard to SET ratings. When controlling for the six RIASEC confidence variables and the Big Five personality traits, respectively, the magnitude of the gender effect dropped to $F(1, 596) = 5.2$, $p = .02$, $\eta^2 = .009$ for confidence and $F(1, 597) = 4.6$, $p = .03$, $\eta^2 = .008$ for personality. Further, the gender difference in gender role attitudes also affected the SET gender effect, reducing its magnitude to $F(1, 607) = 6.2$, $p = .013$, $\eta^2 = .01$. In the additional

analysis in which all covariates were entered simultaneously, the main effect of gender was eliminated [$F(1, 587) = 0.95$, $p = .33$, $\eta^2 = .002$]. Therefore, one can conclude that gender differences in domain-specific interests and confidence, personality traits, and gender role attitudes can fully explain the small but significant difference in the mean ratings between male and female students.

CHAPTER 5: DISCUSSION

The present experimental study evaluated the role of course type, instructor and student gender, and student individual differences in the context of SET ratings. Students rated hypothetical instructor descriptions based on eight common dimensions of teaching effectiveness, and completed self-report measures of vocational interests and confidence, personality, and gender role attitudes. Three main questions were addressed: 1) Do students rate instructors differently depending on the gender of the instructor and the type of course? 2) What are the salient traits that predict students' SET ratings? 3) Can mean differences between male and female students on SET ratings be explained based on individual difference variables? The following discussion will focus on these three questions as well as the implications of the results for policies on SET use in higher education. The chapter concludes with a discussion of limitations of this study and directions for future research.

Course Type and Instructor Gender

The two variables that were experimentally manipulated in the instructor vignettes were course type (counseling vs. research methods) and instructor gender. Given the findings from the literature on students' perception of these two courses, the hypothesis was that students would give significantly higher ratings to the counseling instructor than to the instructor teaching research methods (Conners et al., 1998; Early, 2007; Manning et al., 2006; Vittengl et al., 2004). In addition, prior research on SET bias indicates that instructor gender might play a role in how students rate teacher effectiveness, but the results have remained inconclusive (e.g., Basow & Montgomery, 2005; G. Smith & Anderson, 2005; P. Smith, 2009).

The results from this study showed that there were no systematic differences in SET ratings across the levels of instructor gender or course type, and there were no significant interactions between course type, instructor gender and student gender. There are several possible explanations for these null findings: First, at least with regard to instructor gender, the results might be an accurate representation of the role of instructor gender in the context of SET ratings. In the prior empirical literature, instructor gender did not have any consistent effects on SET ratings (e.g., Basow & Montgomery, 2005; G. Smith & Anderson, 2005; P. Smith, 2009), and the present study corroborates these findings. Therefore, despite the anecdotal evidence that female instructors might be penalized in SET ratings, it is possible that there are no systematic differences in students' perception of the teaching effectiveness of men compared to women.

Second, the failure to find an effect of course type or instructor gender on SET ratings might be due to methodological artifacts. Since hypothetical descriptions of an instructor were used as stimuli rather than an actual instructor with whom students engaged in the teaching and learning process, students might have responded differently to the hypothetical scenarios.

Third, concerning the null effect of course type, it is possible that the two courses described in the vignettes (counseling and research methods) were perceived as similar by students, considering that both courses are within the same discipline. Although the prior literature indicates that mean SET scores indeed vary by academic discipline, no research could be located in which different subareas within the same discipline had been contrasted. Therefore, although the two course types included in this study were chosen because psychology students tend to have a more positive attitude towards human service courses

compared to quantitative subjects, the two courses still both fall within the discipline of psychology. Since the majority of participants were not psychology majors (they took an introductory psychology course to fulfill an academic requirement) they might not have shared the psychology majors' attitudes towards these two course types as being quite distinct.

Finally, the failure to replicate the difference in SET ratings across course types might be an indication that the common explanations for the observed differences in SET scores across academic disciplines (e.g., actual differences in teaching skills, rigor in grading, or appeal of the discipline to the majority of students) might not be sufficient to explain the phenomenon. Another possibility is that the differences in SET ratings across academic disciplines result from the types of students that compose the roster of the courses in various disciplines. Students with a particular combination of interests and personality traits tend to be attracted to some majors but not others as the prior literature indicates (Gasser et al., 2004; Larson, Wei, Wu, Borgen, & Bailey, 2007; Larson, Wu, Bailey, Borgen, & Gasser, 2010; Larson, Wu, Bailey, Gasser et al., 2010). Therefore, if the traits in question are differentially linked to SET ratings, SET scores across disciplines are expected to vary based on what types of students are over- or under-represented in these courses. The conclusions from the present study would support this hypothesis as discussed throughout the remainder of the discussion section.

The Link between Student Individual Differences and SET Ratings

The relation between student individual differences and SET ratings was assessed in two different ways. In the first regression analysis, all predictors were entered simultaneously to predict the criterion. In the second regression, the predictors were entered separately by

construct. Both methods have distinct advantages and drawbacks. By entering all predictors in one single analysis, the overlap of variance between constructs can be taken into account, but the resulting multicollinearity between the predictors can distort the direction and magnitude of the regression coefficients, which can make them difficult to interpret. This problem was therefore addressed by also conducting separate analyses by construct. This second approach, however, assumes no overlap between constructs, which disregards the empirical evidence that there is a moderate amount of shared variance between them (e.g., domain-specific interests and confidence, see Armstrong & Vogel, 2009; Rottinghaus et al., 2003). With these limitations in mind, the following section will focus on the interpretation of the findings from both sets of analyses, including a discussion of the communalities and discrepancies in findings.

Consistent Individual Difference Effects

The present study found that several student individual differences were systematically linked to SET ratings. The variables that had consistent effects across the two analysis approaches were investigative confidence (negative association with SET ratings as indicated by the regression coefficient), conventional confidence, agreeableness, and conscientiousness; the latter three had positive regression coefficients. Further, a more traditional gender attitude contributed to lower SET ratings in both analyses. Out of these five variables, the two personality traits (agreeableness and conscientiousness) and gender role attitudes appeared to have the largest effect, judged by the magnitude of both the bivariate SET-individual difference correlations (see Table 2) and the regression coefficients (Tables 3 and 4).

Agreeableness. The idea that agreeable students tend to give higher SET ratings makes intuitive sense. These students are individuals who are interested in helping others, who are considerate of others' needs, and empathetic towards their feelings. They are able to emphasize with the negative feelings an instructor might have when obtaining low SET ratings. Thus, they might be reluctant to give low ratings that would hurt the instructor's feelings. In addition, agreeable students tend to value interpersonal relationships, and are more likely than science/engineering oriented students to have interest in career paths related to teaching, counseling, and education (Larson et al., 2007; Larson, Wu, Bailey, Borgen et al., 2010; Larson, Wu, Bailey, Gasser et al., 2010). Therefore, these students might be more likely to intrinsically value the teaching and learning process. This positive attitude might then be reflected in how they perceive the instructor, giving way to higher SET ratings.

As a side note it should be noted that extraversion, although positively correlated with social interests and agreeableness (see Table 2), was not significantly related to SET scores. People who are very extraverted also enjoy interacting with other people. However, these interpersonal relationships do not necessarily occur in the context characterized by social interest (helping, nurturing, teaching), but also in a business context, where interactions between people are often focused on leading and persuading others (social potency). Therefore, the indication that extraversion is not a unique feature of people who are highly interested in relating to others in a nurturing and empathic, way might explain the absence of a significant extraversion – SET correlation.

Conscientiousness and conventional confidence. Both conscientiousness and conventional confidence were positively associated with SET scores. Moreover, the two predictors showed a small but significant correlation with each other (r = .19 for male student

and $r = .18$ for female students, $p < .01$ for both bivariate correlations). This is in agreement with the prior literature on individual difference integration, which documented the existence of shared variance between conscientiousness and the conventional domain (Ackerman & Heggestad, 1997; P. I. Armstrong, Day, McVay, & Rounds, 2008). Therefore, the effects of these two traits on SET ratings will be discussed together.

One possible explanation for the positive relation between SET ratings on the one hand, and conscientiousness and conventional confidence on the other hand, might be based on how these students approach the learning process and the classroom environment. Students who are conscientious and have confidence in their ability to complete conventional tasks can be characterized as being organized, they tend to like order, get tasks done right away, and pay attention to details and schedules. These are all qualities that help to succeed in an academic context. It is possible that these students have a more positive learning experience, better learning outcomes, and stronger rapport with their instructor, which is then reflected in their SET ratings.

Support for this idea (at least in the case of conscientiousness) comes from the pattern of correlations of conscientiousness, Grade Point Average (GPA), and motivation to complete their degree as estimated in the present study: The higher the students' level of conscientiousness, the higher was their GPA ($r = .17$, $p < .01$), and the higher their motivation for degree completion ($r = .25$, $p < .01$); the latter was assessed based on a single item ("How motivated are you to complete your bachelor's degree?") scored on a 4-point scale ranging from 1 = *I am unmotivated to complete my bachelor's degree* to 4 = *I am very motivated to complete my bachelor's degree*.

It should be noted, however, that, unlike conventional confidence, interest in this domain was not significantly related to SET ratings based on both the bivariate correlations (Table 2) and the outcomes of the regression analyses (Tables 3 and 4). Given the moderate to strong correlation between conventional interest and conventional confidence (r = .43 for male students and r = .63 for female students, both p < .01), it is not clear why only confidence but not interest in this domain showed a significant association with SET ratings. This remains a question for future research.

Gender role attitudes and investigative confidence. Students with a more traditional gender role attitude gave lower ratings to the instructor depicted in the vignette. Although gender role stereotypes have been somewhat addressed in the prior SET literature (Bachen et al., 1999; Basow & Montgomery, 2005; Basow et al., 2006; Bennett, 1982; Martin, 1984), this primarily occurred in the context of the interaction between instructor and students, and there is some evidence that female instructors might be penalized for acting contrary to traditionally female gender roles. These studies, however, did not examine the influence on students' gender role attitudes on SET ratings outside of this context.

One possible conceptual explanation for the finding that those students with traditional gender role attitudes tended to give lower SET ratings in the present study can be derived from the relation between this construct and other individual difference domains. For example, people who are interested in realistic activities and who are less interpersonally oriented tend to be drawn to academic majors and occupations in engineering and science fields (Larson et al., 2007; Larson, Wu, Bailey, Borgen et al., 2010; Larson, Wu, Bailey, Gasser et al., 2010). Further, other research has shown that men and women who choose to enter such occupations tend to endorse a more traditional and conservative view with regard

to gender roles (Dodson & Borders, 2006; Hirschi, 2010; Leaper & Van, 2008; Mahalik, Perry, Coonerty-Femiano, Catraio, & Land, 2006; Oswald, 2008; Tokar & Jome, 1998). The findings from the present study are consistent with this body of literature: For example, more traditional students reported higher realistic interests, less openness to new experiences, and lower levels of agreeableness (Table 2); this pattern of results was present for both genders.

Therefore, it might be possible to interpret the SET – gender role relation within the larger context of a more traditional worldview that manifests itself across multiple individual difference domains. Some support for this idea can be derived from the finding of a negative relation between SET ratings and investigative confidence. High levels of investigative confidence are typically found in individuals who are attracted to the field of science and engineering (Larson et al., 2007; Larson, Wu, Bailey, Borgen et al., 2010; Larson, Wu, Bailey, Gasser et al., 2010). Therefore, the finding that both a more traditional gender role attitude and high confidence in a field, which tends to attract individuals who are more traditional in their worldview are associated with lower SET ratings, is conceptually congruent.

Additional Tentative Individual Difference Effects

There were several variables that significantly ($p < .01$) predicted SET ratings in only one of the two regression models (see Tables 3 and 4). While investigative interest was positively related to SET scores when all predictors were entered together, this variable was not salient in the analysis that solely focused on the RIASEC interests as predictors. On the other hand, four variables were significant predictors in the analyses conducted separately by construct, but not in the combined analysis; these were interest in statistics, realistic interest and confidence (all three had negative regression coefficients), and social confidence (which

was positively linked to the criterion). The inconsistency in findings can be largely attributed to the different emphasis placed on accounting for collinearity between the predictors in the two types of analyses. In the first analysis, the shared portion of the variance between constructs was removed, while the second set of analyses assumed independence of constructs. Therefore, the findings from the regression analyses should be interpreted with caution, and more weight should be given to the findings that were consistent across the two analysis approaches.

Another effect that is potentially linked to the issue of multicollinearity in the construct-combined regression is the inversion of the sign of the regression coefficient for interest and confidence in the same domain. For example, within the investigative domain, interest had a significant positive relation with SET ratings, while confidence appears to be negatively related to the criterion. It is not clear how this can be interpreted conceptually, and it is likely that this phenomenon is an artifact of the removal of shared variance in the regression. This effect warrants further investigation.

In sum, there seem to be several student individual difference variables that are systematically related to SET ratings. The interpretation of the findings is complicated due to the limitations of the statistical analyses in terms of the potential conceptual overlap between construct domains. Nonetheless, the general pattern of the individual difference - SET relations appears to be meaningful and largely consistent with the existing literature on individual difference integration.

Student Gender and SET Ratings

Previous studies have found that female students on average tend to give significantly higher SET ratings than their male peers (e.g., Bachen et al., 1999; Badri et al., 2006; Darby,

2006a; Santhanam & Hicks, 2002). This result was replicated in the present study, although the effect was small. Prior research on this topic has remained primarily at the descriptive level. The present research adds to these findings by providing a plausible explanation of the mean difference in SET ratings. There is robust evidence of gender differences in interests and personality traits, with women typically showing significantly higher interest in people-oriented occupations than men (Lippa, 1998; Su et al., 2009), and reporting higher levels of agreeableness and neuroticism (Costa et al., 2001; Feingold, 1994; Lippa, 2010; Schmitt et al., 2008). The results from this study were in line with these prior findings. Since some of these variables (e.g., agreeableness) were also systematically related to SET ratings, the hypothesis was that the gender differences in SET ratings could be explained by gender differences on traits that correlate with the SET scores. This hypothesis was supported; the gender effect regarding SET ratings was eliminated when statistically controlling for students' individual differences. Therefore, it is not gender per se that is responsible for the mean difference in SET scores, but the differential endorsement of traits that correlate with the ratings.

Bias, Validity, and Policy Implications

The findings from the present research suggest that student variables unrelated to teaching effectiveness (the construct assumed to be measured by SET ratings) nonetheless are systematically related to SET scores. Since students rated fictitious instructors described in a vignette, they did not actually engage in the teaching and learning process with the instructor. Therefore, the individual difference – SET correlations can be regarded as evidence of bias, which poses a threat to the validity of the ratings. This is particularly a concern since the largest obtained effects were in the practically meaningful range. For

example, everything else being equal, the difference in predicted SET ratings between students high (95th percentile) in agreeableness and those low (5th percentile) on this trait was in the magnitude of one full scale point on a 7-point scale.

The observed relation between student background and their SET ratings has implications for the policies that govern the use of SET scores, especially for summative purposes. In practice, students are not randomly assigned to courses and academic majors, but they tend to gravitate towards those that are in line with their interests, personality, and abilities (Larson et al., 2007; Larson, Wu, Bailey, Borgen et al., 2010; Larson, Wu, Bailey, Gasser et al., 2010). Therefore, an instructor of a course that attracts mainly students high in agreeableness (e.g., child development) could expect higher SET ratings than an equally effective instructor whose course consists mainly of students high in investigative confidence (e.g., engineering). Likewise, if there are courses (e.g., introductory statistics) that have designated sections for students based on academic discipline (e.g., social sciences, engineering, etc.) that are taught by different instructors, the SET rating will be confounded by the student composition of the respective section. Here, it can be expected that more agreeable students are overrepresented in the social sciences section, while the engineering section will have a larger proportion of students that score lower on this trait. If the instructor of the social science section obtains higher ratings than the instructor of the engineering section of the same course, does this mean that the social science instructor is the more effective teacher? Since there is a competing explanation for these ratings (the two instructors are equally effective, but the social science instructor has an unfair advantage based on course composition), it is difficult to make accurate judgments about what the ratings mean. Therefore, when instructors are compared relative to each other across different

academic disciplines (e.g., when determining eligibility for promotion), the fairness of the process might be undermined.

To increase the fairness of the SET evaluation process, the following suggestions could be implemented. First, instructors should not be compared relative to one another when they teach in different academic disciplines, especially when these courses attract different types of students in terms of personality and interests. In addition, student characteristics could be taken into account when making judgments about the teaching effectiveness of the instructor (e.g., gender balance, academic major, etc.). Second, student individual differences are a systematic error component that factors into SET scores. This can be viewed as a problem of rater agreement. It might be possible to statistically remove rater differences due to student variables. For example, one could choose the strongest and most reliable individual difference predictors of SET scores and administer a short scale measuring these constructs along with the SET questionnaire. For each student, the SET score could then be statistically corrected based on how they score on the respective trait.

Limitations and Future Directions

This study has many strengths and unique features (e.g., experimental design, multiple variables examined simultaneously), but the following limitations should be noted. First, although the use of vignettes in place of actual instructors made it possible to distinguish the effects of bias and teaching effectiveness, there are downsides to this paradigm. For example, the vignette design necessitated the assumption that students' rating behavior would be similar to how they usually respond in the classroom under normal SET administration conditions. As is the case with any laboratory research, it is important to generalize the findings to the actual field setting. Future research could replicate key aspects

of this study in the actual classroom, e.g., by assessing students on relevant individual differences at the beginning of the semester, which can be then used to predict their SET ratings at the end of the semester. This would also allow determining the proportion of variance in SET ratings explained by student individual differences as compared to other factors.

Second, as already discussed, only two types of courses (which were both within psychology) were included in the vignettes. A future study could investigate whether the null findings hold up when courses from a variety of academic disciplines (e.g., electrical engineering vs. child development) are compared with each other.

Third, many of the variables included in the study had not been examined in the prior literature, and the study was exploratory in nature. Therefore, the individual difference domains were chosen to be relatively broad in order to obtain a preliminary overview. Future research could look at specific facets of the most salient traits that are linked to SET ratings (e.g., the different facets of agreeableness or conscientiousness) in order to find the most salient predictors of SET ratings.

Finally, the content and format of the SET rating scale might have affected the results. Students were asked to rate the vignettes on a 7-point scale. However, the scale commonly used at the students' university is a 5-point scale. In addition, not all of the eight SET items used in this study were part of the standard SET questionnaire administered at the university. Although unlikely, these small changes in item format and scaling might have prompted students to respond differently than during routine SET administration in their courses. Therefore, future research could look at the influence of different rating scale formats on students' response style.

Beyond these more proximal concerns, future SET research could address the following broader questions. First, it would be important to know how student individual differences interact with instructor characteristics in their mutual impact on SET scores. So far, only one study has assessed both the instructor and the students concurrently on the same individual difference variable, namely social style (Schlee, 2005). These findings suggest that there is an interaction between the social styles of instructor and student; matching social styles between instructor and student resulted in higher SET ratings. Future research could expand on these findings by assessing a wide variety of individual differences.

A second question concerns the nature of the SET rating process. It is not well understood how students actually make their evaluations. Prior research has shown that students often have difficulties interpreting SET questions, and that they do not necessarily define teaching effectiveness in the same way as their instructors or administrators (e.g., Billings-Gagliardi et al., 2004; Fritschner, 2000; Kolitch & Dean, 1998). Therefore, more research is needed (e.g., in the form of think-aloud interviews and other qualitative research formats) that would clarify students' thought processes and rating behavior when completing SET questionnaires.

Another open question concerns the magnitude of the observed individual difference bias in SET ratings and the types of questions that are most susceptible to it. In general, research has shown that items that are global, ambiguous, and vague are more likely to be affected by bias (e.g., Bargh et al., 1996; Operario & Fiske, 2004; Plous, 2003). Future research should attempt to address this issue and find the types of items that are least prone to influence by student individual differences.

Finally, very little is known at this point about how administrators currently use SET information. In order to increase the fairness of the process, it is important to know the level of training that administrators have received with regard to validity issues, their awareness of the state of the SET literature, and their decision- making protocols, both formal and informal.

Conclusion

This experimental study examined the contributions of course type, instructor gender, student gender, and student individual differences in terms of their biasing effect on SET ratings. As the main result, student individual differences were systematically related to students' ratings of a set of instructor vignettes. In addition, student individual differences explained the commonly found mean difference in ratings given by male and female students. Students tend to choose their courses and majors partly based on their individual difference profile, a process the instructor has no control over. Therefore, differences in the composition of a course in terms of student background and trait constellation can be seen as a potential threat to the validity of the ratings. This in turn challenges the fairness of instructor evaluation practices in the context of high-stakes personnel decisions in higher education.

REFERENCES

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research, 109*, 59-87.

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58*, 151-179.

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice*. San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness - Generalizability of "N=1" research: Comment on Marsh (1991). *Journal of Educational Psychology, 83*, 411-415.

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 11, pp. 213-264). New York: Agathon Press.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research, 52*(3), 446-464.

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*(2), 227-257.

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219-245.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning, 31*, 25-31.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153-166.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., et al. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134-141.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431-441.

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences, 27*(2), 184-201.

Armstrong, J. S. (1998). Are student ratings of instruction useful? *American Psychologist, 53*(11), 1223-1224.

Armstrong, P. I., Allison, W., & Rounds, J. (2008). Development and initial validation of brief public domain RIASEC marker scales. *Journal of Vocational Behavior, 73*(2), 287-299.

Armstrong, P. I., Day, S. X., McVay, J. P., & Rounds, J. (2008). Holland's RIASEC model as an integrative framework for individual differences. *Journal of Counseling Psychology, 55*(1), 1-18.

Armstrong, P. I., & Vogel, D. L. (2009). Interpreting the interest-efficacy association from a RIASEC perspective. *Journal of Counseling Psychology, 56*(3), 392-407.

Ashmore, R. D., Del Boca, F. K., & Bilder, S. M. (1995). Construction and validation of the Gender Attitude Inventory, a structured inventory to assess multiple dimensions of gender attitudes. *Sex Roles, 32*, 11-12.

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*(3), 193-210.

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management, 20*(1), 43-59.

Bain, K. (2004). *What the best college teachers do*. Cambridge, MA: Harvard University Press.

Balam, E. M., & Shannon, D. M. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment & Evaluation in Higher Education, 35*(2), 209-221.

Ballantyne, R., Borthwick, J., & Packer, J. (2000). Beyond student evaluation of teaching: Identifying and addressing academic staff development needs. *Assessment & Evaluation in Higher Education, 25*(3), 221-236.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230-244.

Barnes, D. C., Engelland, B. T., Matherine, C. F., Martin, W. C., Orgeron, C. P., Ring, J. K., et al. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal, 42*(1), 199-213.

Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business, 84*(1), 40-46.

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*(2), 91-106.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*(1), 25-35.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*, 308-314.

Bausell, R. B., & Bausell, C. R. (1979). Student ratings and various instructional variables from a within-instructor perspective. *Research in Higher Education, 11*, 167-177.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology, 74*, 170-179.

Beran, T. N., & Rokosh, J. L. (2009). Instructors' perspectives on the utility of student ratings of instruction. *Instructional Science, 37*(2), 171-184.

Billings-Gagliardi, S., Barrett, S. V., & Mazor, K. M. (2004). Interpreting course evaluation results: Insights from thinkaloud interviews with medical students. *Medical Education, 38*(10), 1061-1070.

Blum, M. I., & Naylor, J. C. (1968). *Industrial psychology: Its theoretical and social foundations*. New York: Harper & Row.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Browne, M. N., Hoag, J. H., Myers, M. L., & Hiers, W. J. (1997). Student evaluation of teaching as if critical thinking really mattered. *Journal of General Education, 46*(3), 192-206.

Buchert, S., Laws, E. L., Apperson, J. M., & Bregman, N. J. (2008). First impressions and professor reputation: Influence on student evaluations of instruction. *Social Psychology of Education, 11*(4), 397-408.

Buck, D. (1998). Student evaluations of teaching measure the intervention, not the effect. *American Psychologist, 53*(11), 1224-1226.

Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning, 43*, 113-121.

Centra, J. A. (1979). *Determining faculty effectiveness*. San Francisco: Jossey-Bass.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495-518.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71*(1), 17-33.

Chacko, T. I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly, 8*(2), 19-25.

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*(1), 71-88.

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education, 21*(1), 68-75.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16-30.

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education, 5*(1), 52-85.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28*(2), 149-160.

Cohen, E. H. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education, 30*(2), 123-136.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281-309.

Cohen, P. A. (1982). Validity of student ratings in psychology courses: A research synthesis. *Teaching of Psychology, 9*(2), 78-82.

Cohen, P. A. (1987). *A critical analysis and reanalysis of the multisection validity meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association.

Conners, F. A., McCown, S. M., & Roskos-Ewoldsen, B. (1998). Unique challenges in teaching undergraduate statistics. *Teaching of Psychology, 25*(1), 40-42.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*, 5-13.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322-331.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41*, 511-535.

Covert, R. W., & Mason, E. J. (1974). Factorial validity of a student evaluation of teaching instrument. *Educational and Psychological Measurement, 34*(4), 903-905.

Cronin, L., & Capie, W. (1986). *The influence of daily variation in teacher performance on the reliability and validity of assessment data*. Paper presented at the Annual Meeting of the American Educational Research Association.

Crumbley, D. L., & Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior: Student evaluation of teaching control data. *Quality Assurance in Education: An International Perspective, 17*(4), 377-392.

Cruse, D. B. (1987). Student evaluations of the university professor: Caveat professor. *Higher Education, 16*, 723-737.

Cunningham, J. B., & MacGregor, J. N. (2006). The echo approach in developing items for student evaluation of teaching performance. *Teaching of Psychology, 33*(2), 96-100.

d'Apollonia, S., & Abrami, P. C. (1996). *Variables moderating the validity of student ratings of instruction: A meta-analysis*. Paper presented at the 77th annual meeting of the American Educational Research Association.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208.

Darby, J. A. (2006a). Evaluating courses: An examination of the impact of student gender. *Educational Studies, 32*(2), 187-199.

Darby, J. A. (2006b). The effects of the elective or required status of courses on student evaluations. *Journal of Vocational Education and Training, 58*(1), 19-29.

Darby, J. A. (2007a). Are course evaluations subject to a Halo Effect? *Research in Education, 77*, 46-55.

Darby, J. A. (2007b). Open-ended course evaluations: A response rate problem? *Journal of European Industrial Training, 31*(5), 402-412.

Darby, J. A. (2008). Course evaluations: A tendency to respond "favourably" on scales? *Quality Assurance in Education: An International Perspective, 16*(1), 7-18.

Das, M., & Das, H. (2001). Business students' perceptions of best university professors: Does gender role matter? *Sex Roles, 45*(9-10), 665-676.

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco, CA: Jossey-Bass.

DeBerg, C. I., & Wilson, J. R. (1990). An empirical investigation of the potential confounding variables in student evaluation of teaching. *Journal of Accounting Education, 8*(1), 37-62.

Dodson, T. A., & Borders, L. D. (2006). Men in traditional and nontraditional careers: Gender role attitudes, gender role conflict, and job satisfaction. *The Career Development Quarterly, 54*(4), 283-296.

Dolnicar, S., & Grun, B. (2009). Response style contamination of student evaluation data. *Journal of Marketing Education, 31*(2), 160-172.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education, 53*(1), 51-62.

Dujari, A. (1993). Vocabulary comprehension of evaluation form: Its influence on student rating of faculty. *Report: ED364164.* .

Early, M. A. (2007). Students' expectations of introductory statistics teachers. *Statistics Education Research Journal, 6*(1), 51-66.

Edstrom, K. (2008). Doing course evaluation as if learning matters most. *Higher Education Research & Development, 27*(2), 95-106.

Ehrhart, K. H., Roesch, S. C., Ehrhart, M. G., & Kilian, B. (2008). A test of the factor structure equivalence of the 50-item IPIP Five-factor model measure across gender and ethnic groups. *Journal of Personality Assessment, 90*(5), 507-516.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483-501.

Elliot, D. N. (1950). Characteristics and relationships of various criteria of college and university training. *Purdue University Studies in Higher Education, 70*, 5-61.

Emery, C. R. (1995). *Student evaluations of faculty performance*. Clemson, SC: Clemson University.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429-456.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers and courses: What we know and what we don't. *Research in Higher Education, 9*, 199-242.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*(1), 3-124.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: a closer look. *Research in Higher Education, 21*(1), 45-116.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education, 24*(2), 139-213.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education, 28*(4), 291-344.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*(2), 137-174.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I - Evidence from the social laboratory and experiments *Research in Higher Education, 33*, 317-375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II - Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*, 151-211.

Fernandez, J., Mateo, M. A., & Muniz, J. (1998). Is there a relationship between class size and student ratings of teaching quality? *Educational and Psychological Measurement, 58*(4), 596-604.

Fortunato, V. J., & Mincy, M. D. (2003). The interactive effects of dispositional affectivity, sex, and a positive mood induction on student evaluations of teachers. *Journal of Applied Social Psychology, 33*(9), 1945-1972.

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning, 87*, 85-100.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51*(1), 115-134.

Freng, S., & Webber, D. (2009). Turning up the heat on online teaching evaluations: Does "hotness" matter? *Teaching of Psychology, 36*(3), 189-193.

Frick, T. W., Chadha, R., Watson, C., & Zlatkovska, E. (2010). Improving course evaluations to improve instruction and complex learning in higher education. *Educational Technology Research and Development, 58*(2), 115-136.

Fritschner, L. M. (2000). Inside the undergraduate college classroom: Faculty and students differ on the meaning of student participation. *Journal of Higher Education, 71*, 342-362.

Gaffuri, A., Wrench, D., Karr, C., & Kopp, R. (1982). Exploring some pitfalls in student evaluation of teaching. *Teaching of Psychology, 9*(4), 229-230.

Gasser, C. E., Larson, L. M., & Borgen, F. H. (2004). Contributions of personality and interests to explaining the educational aspirations of college students. *Journal of Career Assessment, 12*(4), 347-365.

Gillmore, G. M., & Greenwald, A. G. (1999). Using statistical adjustment to reduce biases in student ratings. *American Psychologist, 54*(7), 518-519.

Glascock, J., & Ruggiero, T. E. (2006). The relationship of ethnicity and sex to professor credibility at a culturally diverse university. *Communication education, 55*(2), 197-207.

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.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96.

Goldman, L. (1993). On the erosion of education and the eroding foundations of teacher education (or why we should not take student evaluation of faculty seriously). *Teacher Education Quarterly, 20*(2), 57-64.

Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's 'IPIP' Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences, 39*(2), 317-329.

Granzin, K. L., & Painter, J. J. (1973). A new explanation for students' course evaluation tendencies. *American Educational Research Journal, 10*(2), 115-124.

Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe, 89*(5), 44-46.
Greenwald, A. G. (2002). Constructs in student ratings of instructors. In H. I. Braun (Ed.), *The role of constructs in psychological and educational measurement* (pp. 277-297). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.
Greenwald, A. G., & Gillmore, G. M. (1998). How useful are student ratings? Reactions to comments on the current issues section. *American Psychologist, 53*(11), 1228-1229.
Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality--Analysis of relevant factors based on empirical evaluation research. *Assessment & Evaluation in Higher Education, 28*(3), 229-238.
Gurung, R. A. R., & Vespia, K. M. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34*(1), 5-10.
Heppner, P. P., Wampold, B. E., & Kivlighan, D. M. (2008). *Research design in counseling* (3rd ed.). Belmont, CA: Thomson.
Hirschi, A. (2010). Vocational interests and career goals: Development and relations to personality in middle adolescence. *Journal of Career Assessment, 18*(3), 223-238.
Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49*(1), 26-31.
Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6*(1), 35-45.
Holland, J. L. (1997a). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.
Holland, J. L. (1997b). *Making vocational choices: A theory of vocational personalities and work environments (3rd ed.)*. Odessa, FL: Psychological Assessment Resources.
Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology, 63*, 130-133.
Howard, G. S. (1984). Thoughts on assumptions in "Exploring some pitfalls in student evaluation of teaching.". *Teaching of Psychology, 11*(3), 184-185.
Hoyt, D. P., & Cashin, W. E. (1977). Development of the IDEA system. In *IDEA Technical Report No. 1*. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.
Husbands, C. T., & Fosh, P. (1993). Students' evaluation of teaching in higher education: Experiences from four European countries and some implications of the practice. *Assessment and Evaluation in Higher Education, 18*(2), 95-114.
Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*(4), 580-596.
Jenkins, S. J., & Downs, E. (2001). Relationship between faculty personality and student evaluation of courses. *College Student Journal, 35*(4), 636-640.
Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5*(4), 419-434.
Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*(2), 84-91.

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology, 80*, 342-344.

Klein, M., & Rosar, U. (2006). The eye is also listening! *Zeitschrift fuer Soziologie, 35*(4), 305-316.

Koh, C. H., & Tan, T. M. (1997). Empirical investigation of the factors affecting SET results *International Journal of Educational Management, 11*(4), 170-178.

Kolitch, E., & Dean, A. V. (1998). Item 22, "overall, [the Instructor] was an effective teacher": Multiple meanings and confounding influences. *Journal on Excellence in College Teaching, 9*(2), 119-140.

Koon, J., & Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education, 66*(1), 61-81.

Langbein, L. (1994). The validity of student evaluations of teaching. *Political Science and Politics, September*, 545-553.

Larson, L. M., & Borgen, F. H. (2002). Convergence of vocational interests and personality: Examples in an adolescent gifted sample. *Journal of Vocational Behavior, 60*(1), 91-112.

Larson, L. M., Rottinghaus, P. J., & Borgen, F. H. (2002). Meta-analyses of Big Six interests and Big Five personality factors. *Journal of Vocational Behavior, 61*(2), 217-239.

Larson, L. M., Wei, M., Wu, T. F., Borgen, F. H., & Bailey, D. C. (2007). Discriminating among educational majors and career aspirations in Taiwanese undergraduates: The contribution of personality and self-efficacy. *Journal of Counseling Psychology, 54*, 395-408.

Larson, L. M., Wu, T. F., Bailey, D. C., Borgen, F. H., & Gasser, C. E. (2010). Male and female college students' educational majors: The contribution of basic vocational confidence and interests. *Journal of Career Assessment, 18*, 16-33.

Larson, L. M., Wu, T. F., Bailey, D. C., Gasser, C. E., Bonitz, V. S., & Borgen, F. H. (2010). The role of personality in the selection of a major: with and without vocational self-efficacy and interests. *Journal of Vocational Behavior, 76*, 211-222.

Lauer, J. B., Rajecki, D. W., & Minke, K. A. (2006). Statistics and methodology courses: Interdepartmental variability in undergraduate majors' first enrollments. *Teaching of Psychology, 33*(1), 24-30.

Leaper, C., & Van, S. R. (2008). Masculinity ideology, covert sexism, and perceived gender typicality in relation to young men's academic motivation and choices in college. *Psychology of Men & Masculinity, 9*(3), 139-153.

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*(1), 79-122.

Lewis, P., & Rivkin, D. (1999). *Development of the O*NET Interest Profiler*. Raleigh, NC: National Center for O*NET Development.

Liao, H.-Y., Armstrong, P. I., & Rounds, J. (2008). Development and initial validation of public domain Basic Interest Markers. *Journal of Vocational Behavior, 73*(1), 159-183.

Liaw, S. H., & Goh, K. L. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management, 17*(1), 37-43.

Liddle, B. J. (1997). Coming out in class: Disclosure of sexual orientation and teaching evaluations. *Teaching of Psychology, 24*(1), 32-35.

Lim, B.-C., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's International Personality Item Pool: A multitrait-multimethod examination. *Organizational Research Methods, 9*(1), 29-54.

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests: The importance of the people-things dimension. *Journal of Personality and Social Psychology, 74*(4), 996-1009.

Lippa, R. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior, 39*, 619-636.

Ludwig, J. M., & Meacham, J. A. (1997). Teaching controversial courses: Student evaluations of instructors and content. *Educational Research Quarterly, 21*(1), 27-38.

Lueck, T. L., & et al. (1993). The interaction effects of gender on teaching evaluations. *Journalism Educator, 48*(3), 46-54.

Machina, K. (1987). Evaluating student evaluations *Academe, 73*(3), 19-22.

Mahalik, J. R., Perry, J. C., Coonerty-Femiano, A., Catraio, C., & Land, L. N. (2006). Examining conformity to masculinity norms as a function of RIASEC vocational interests. *Journal of Career Assessment, 14*(2), 203-213.

Manning, K., Zachar, P., Ray, G. E., & LoBello, S. (2006). Research methods courses and the scientist and practitioner interests of psychology majors. *Teaching of Psychology, 33*(3), 194-196.

Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal, 14*(4), 441-447.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal, 17*(1), 219-237.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-764.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-388.

Marsh, H. W. (1994). Weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 86*, 631-648.

Marsh, H. W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 87*(4), 666-679.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775-790.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: a multidimensional perspective. In J. C. Smart (Ed.), *Higher education: handbook of theory and research* (Vol. 8). New York: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*(4), 303-314.

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30*(1), 217-251.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

Martin, E. (1984). Power and authority in the classroom: Sexist stereotypes in teaching evaluations. *Journal of Women in Culture and Society, 9*, 482-492.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education, 21*, 150-158.

McCormack, C. (2005). Reconceptualizing student evaluation of teaching: An ethical framework for changing times. *Assessment & Evaluation in Higher Education, 30*(5), 463-476.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

McMartin, J. A., & Rich, H. E. (1979). Faculty attitudes toward student evaluation of teaching. *Research in Higher Education, 11*(2), 137-152.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education, 37*(1), 3-20.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 88*(3), 868-881.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education.

Messick, S. (1995). Validity of psychological assessment - Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Mlacic, B., & Goldberg, L. R. (2007). An analysis of a cross-cultural personality inventory: The IPIP Big-Five factor markers in Croatia. *Journal of Personality Assessment, 88*, 168-177.

Munz, D. C., & Munz, H. E. (1997). Student mood and teaching evaluations. *Journal of Social Behavior and Personality, 12*(1), 233-242.

Murray, H. G., Jelley, R. B., & Renaud, R. D. (1996). Longitudinal trends in student instructional ratings: Does evaluation of teaching lead to improvement of teaching? *Report: ED417664.* .

Naftulin, D. H., Ware Jr., J. E., & Donelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187-198.

Nasser, F., & Glassman, D. (1997). Student evaluation of university teaching: Structure and relationship with student characteristics. *Report: ED407390*.

Operario, D., & Fiske, S. T. (2004). Stereotypes: Content, structures, processes, and context. In M. B. Brewer & M. Hewstone (Eds.), *Social cognition. Perspectives on social psychology* (pp. 120-141). Malden: Blackwell Publishing.

Ortinau, D. J., & Bush, R. P. (1987). The propensity of college students to modify course expectations and its impact on course performance information. *Journal of Marketing Education, 9*, 42-52.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning, 87*, 3-15.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P. C. Abrami & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* San Francisco: Jossey-Bass.

Oswald, D. L. (2008). Gender stereotypes and women's reports of liking and ability in traditionally masculine and feminine occupations. *Psychology of Women Quarterly, 32*(2), 196-203.

Peterson, K. D., & Stevens, D. (1998). Variable data sources in teacher evaluation. *Journal of Research and Development in Education, 31*(3), 123-132.

Plous, S. (2003). The psychology of prejudice, stereotyping, and discrimination: An overview. In S. Plous (Ed.), *Understanding prejudice and discrimination* (pp. 3-48). New York: McGraw-Hill.

Potvin, G., Hazari, Z., Tai, R. H., & Sadler, P. M. (2009). Unraveling bias from student evaluations of their high school science teachers. *Science Education, 93*(5), 827-845.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education: An International Perspective, 15*(2), 178-191.

Pounder, J. S. (2008). Transformational classroom leadership: A novel approach to evaluating classroom performance. *Assessment & Evaluation in Higher Education, 33*(3), 233-243.

Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. *Research in Higher Education, 7*, 193-205.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior, 21*(3), 259-287.

Redding, R. E. (1998). Students' evaluations of teaching fuel grade inflation. *American Psychologist, 53*(11), 1227-1228.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387-415.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology, 133*(1), 19-35.

Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science, 28*(5-6), 439-468.

Rottinghaus, P. J., Larson, L. M., & Borgen, F. H. (2003). The relation of self-efficacy and interests: A meta-analysis of 60 samples. *Journal of Vocational Behavior, 62*(2), 221-236.

Rounds, J., & Tracey, T. J. (1993). Prediger's dimensional representation of Holland's RIASEC circumplex. *Journal of Applied Psychology, 78*(6), 875-890.

Ruffer, R. L., McMahon, A. M., & Rogers, J. R. (2001). Revising a student evaluation of teaching form: A campus-wide transformation process. *Report: ED452773.* .

Ruskai, M. B. (1996). Evaluating student evaluations. *Notices of the American Mathematical Society, 44*(3), 308.

Russ, T. L., Simonds, C. J., & Hunt, S. K. (2002). Coming out in the classroom...an occupational hazard: The influence of sexual orientation on teacher credibility and perceived student learning. *Communication education, 51*(3), 311-324.

Santhanam, E., & Hicks, O. (2002). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education, 7*(1), 17-31.

Schlee, R. P. (2005). Social styles of students and professors: Do students' social styles influence their preferences for professors? *Journal of Marketing Education, 27*(2), 130-142.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168-182.

Sebastian, R. J., & Bristow, D. (2008). Formal or informal? The impact of style of dress and forms of address on business students' perceptions of professors. *Journal of Education for Business, 83*(4), 196-201.

Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction, 16*(5), 401-415.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405.

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology, 19*(2), 174-197.

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*(3), 199-213.

Simpson, R. D. (1995). Uses and misuses of student evaluations of teaching effectiveness. *Innovative Higher Education, 20*(1), 3-5.

Smith, B. P. (2009). Student ratings of teaching effectiveness for faculty groups based on race and gender. *Education, 129*(4), 615-624.

Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos & Education, 4*(2), 115-136.

Smith, K., & Welicker-Pollak, M. (2008). What can they say about my teaching? Teacher educators' attitudes to standardised student evaluation of teaching. *European Journal of Teacher Education, 31*(2), 203-214.

Smith, S. P., & Kinney, D. P. (1992). Age and teaching performance. *Journal of Higher Education, 63*(3), 282-302.

Soyjka, J., Gupta, A. K., & Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching, 50*(2), 44-49.

Spooren, P., & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies, 32*(2), 201-214.

Sprinkle, J. E. (2008). Student Perceptions of Effectiveness: An Examination of the Influence of Student Biases. *College Student Journal, 42*(2), 276-293.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives, 8*(50), Retrieved from http://epaa.asu.edu/ojs/article/view/441.

Staggs, G. D., Larson, L. M., & Borgen, F. H. (2003). Convergence of specific factors in vocational interests and personality. *Journal of Career Assessment, 11*(3), 243-261.

Staggs, G. D., Larson, L. M., & Borgen, F. H. (2007). Convergence of personality and interests: Meta-analysis of the Multidimensional Personality Questionnaire and the Strong Interest Inventory. *Journal of Career Assessment, 15*(4), 423-445.

Statham, A., Richardson, L., & Cook, J. A. (1991). *Gender and university teaching*. Albany, NY: State University of New York.

Stevens, J. J., & Aleamoni, L. M. (1985). The use of evaluative feedback for instructional improvement: A longitudinal perspective. *Instructional Science, 13*(4), 285-304.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135*(6), 859-884.

Svinicki, M. (2001). Encouraging your students to give feedback. *New Directions for Teaching and Learning, 87*, 17-24.

Swaffield, B. C. (1996). What happens when male professors enact feminist pedagogies? *Report: ED397429*.

Sweeney, A. D. P., Morrison, M. D., Jarratt, D., & Heffernan, T. (2009). Modeling the constructs contributing to the effectiveness of marketing lecturers. *Journal of Marketing Education, 31*(3), 190-202.

Tagomori, H. T., & Bishop, L. A. (1995). Student evaluation of teaching: Flaws in the instruments. *Thought & Action, 11*(1), 63-78.

Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology, 4*, 25-29.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2).

Toby, S. (1993). Class size and teaching evaluation. *Journal of Chemical Education, 70*(6), 465-466.

Tokar, D. M., & Jome, L. M. (1998). Masculinity, vocational interests, and career choice traditionality: Evidence for a fully mediated model. *Journal of Counseling Psychology, 45*(4), 424-435.

Tom, G., Tong, S. T., & Hesse, C. (2010). Thick slice and thin slice teaching evaluations. *Social Psychology of Education, 13*(1), 129-136.

Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance *Journal of Educational Psychology, 71*, 207-211.

Vittengl, J. R., Bosley, C. Y., Brescia, S. A., Eckardt, E. A., Neidig, J. M., Shelver, K. S., et al. (2004). Why are some undergraduates more (and others less) interested in psychological research? *Teaching of Psychology, 31*(2), 91-97.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211.

Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors that affect college of business student opinion of teaching and learning. *Journal of Education for Business, 77*(5), 282-289.

Williams, R. G., & Ware Jr., J. E. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings. *American Educational Research Journal, 14*(4), 449-457.

Williams, W. M., & Ceci, S. J. (1997). How'm I doing? *Change, 29*, 12-23.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education, 44*(19), A12-A14.

Wolfer, T. A., & Johnson, M. M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education, 39*(1), 111-121.

Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71*, 764-775.

Yao, Y., & Grady, M. L. (2005). How do faculty make formative use of student evaluation feedback?: A multiple case study. *Journal of Personnel Evaluation in Education, 18*(2), 107-126.

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology, 34*(4), 245-247.

Zayani, M. (2001). The teaching portfolio: Toward an alternative outcomes assessment. *Research and Teaching in Developmental Education, 18*(1), 58-64.

Zheng, L., Goldberg, L. R., Zheng, Y., Zhao, Y., Tang, Y., & Liu, L. (2008). Reliability and concurrent validation of the IPIP Big-Five factor markers in China: Consistencies in factor structure between internet-obtained heterosexual and homosexual samples. *Personality and Individual Differences, 45*(7), 649-654.

Figure 1

Holland's (1997) Model in Juxtaposition with Prediger's (1982) People vs. Things and Data vs. Ideas Dimensions

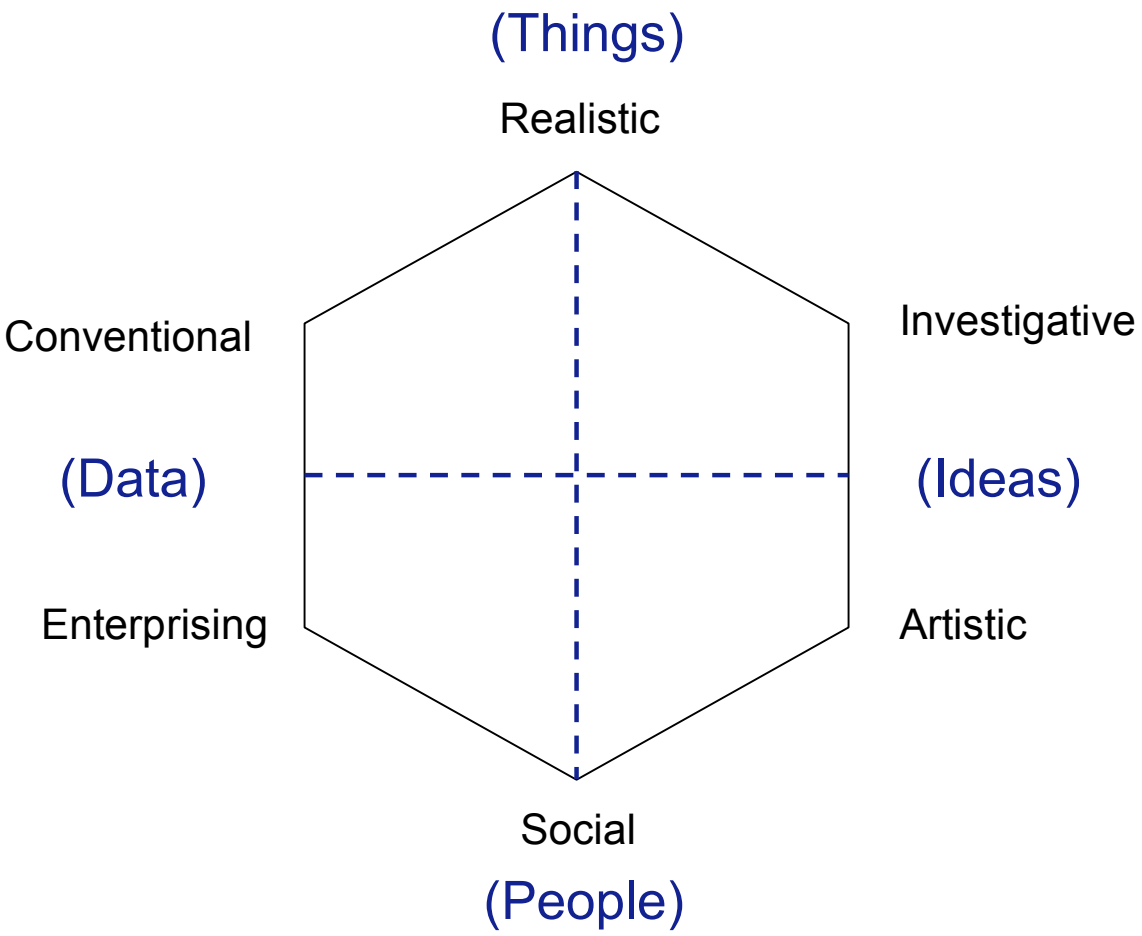

Table 1

Mean Vignette Ratings by Student Gender, Instructor Gender, and Course Type

Student Gender	Instructor Gender	Course Type	*M*	*SD*	*N*
Female	Female	Psychology	6.00	0.83	104
		Research M.	6.00	0.77	87
		Total	6.00	0.80	191
	Male	Psychology	6.00	0.89	93
		Research M.	5.82	0.93	88
		Total	5.91	0.91	181
	Total	Psychology	6.00	0.85	197
		Research M.	5.91	0.86	175
		Total	5.96	0.86	372
Male	Female	Psychology	5.58	0.91	55
		Research M.	5.66	0.78	67
		Total	5.62	0.84	122
	Male	Psychology	5.60	0.83	63
		Research M.	5.72	0.95	53
		Total	5.65	0.88	116
	Total	Psychology	5.59	0.86	118
		Research M.	5.69	0.86	120
		Total	5.64	0.86	238
Total	Female	Psychology	5.85	0.88	159
		Research M.	5.85	0.79	154
		Total	5.85	0.84	313
	Male	Psychology	5.83	0.88	156
		Research M.	5.78	0.94	141
		Total	5.81	0.91	297
	Total	Psychology	5.84	0.88	315
		Research M.	5.83	0.86	295
		Total	5.83	0.87	610

Note. Research M. = Research Methods; Vignette ratings are scored on a 7-point scale, where higher values indicate higher perceived teaching effectiveness of the instructor.

Table 2

Bivariate Correlations between SET Ratings and Student Individual Differences

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. SET rating		.03	.11	-.14	-.25	-.05	-.13	.10	-.04	-.07	-.13	-.11	-.14	.09	-.02	.16	-.08	.08	.13	.29	.20	-.26
2. Int. SSci	.09		.63	.05	.06	.26	.29	.44	.06	-.06	.08	.15	.15	.31	.11	.02	.05	.06	.30	.31	.01	-.19
3. Int. SServ	.04	.56		-.12	-.09	.04	.22	.79	.12	-.10	-.11	-.05	.06	.57	.12	-.08	.01	.20	.18	.42	.12	-.11
4. Int. Stat.	-.03	.02	-.04		.33	.27	.05	-.08	.08	.51	.25	.30	.07	-.10	.11	.39	-.08	-.12	.07	-.18	.03	.08
5. Interest R	-.11	.13	.12	.50		.34	.36	.00	.34	.50	.63	.29	.27	-.08	.19	.19	-.09	-.11	.01	-.30	-.13	.21
6. Interest I	.04	.37	.21	.26	.37		.25	.16	.05	.15	.34	.72	.24	.11	.17	.17	-.04	-.04	.16	-.03	-.04	-.04
7. Interest A	.04	.41	.34	-.02	.25	.35		.24	.37	.09	.29	.18	.76	.18	.37	.01	-.02	.10	.40	.10	-.15	.08
8. Interest S	-.10	.46	.84	-.01	.27	.32	.38		.20	-.02	-.03	.03	.10	.66	.18	.00	-.11	.24	.17	.39	.12	-.05
9. Interest E	-.09	.28	.29	.14	.51	.22	.37	.35		.46	.17	-.09	.21	.10	.53	.17	-.04	.16	-.01	-.09	-.11	.17
10. Interest C	-.11	.06	.09	.58	.65	.25	.13	.13	.51		.28	.08	.02	-.08	.24	.62	.02	-.11	-.04	-.20	.04	.08
11. Conf. R	.03	.17	.16	.27	.60	.32	.23	.28	.26	.26		.49	.45	.14	.41	.39	-.12	-.07	.12	-.20	-.04	.05
12. Conf. I	-.16	.23	.19	.30	.26	.54	.18	.23	.00	.11	.45		.35	.22	.26	.30	-.06	-.14	.15	-.12	.00	.00
13. Conf. A	.00	.33	.29	.01	.17	.21	.71	.30	.25	.04	.37	.40		.27	.46	.12	-.02	.09	.37	.00	-.12	.14
14. Conf. S	-.01	.35	.62	-.02	.08	.18	.27	.65	.17	-.05	.41	.42	.47		.40	.13	-.06	.25	.20	.35	.09	.00
15. Conf. E	.01	.29	.35	-.07	.19	.08	.30	.35	.50	.13	.43	.26	.52	.55		.41	-.13	.26	.24	.06	.01	.10
16. Conf. C	.10	.10	.12	.29	.22	.15	.09	.13	.19	.43	.52	.32	.28	.30	.44		-.03	-.09	.14	-.01	.18	-.10
17. Neurot	-.13	.06	.01	-.05	.04	-.08	-.07	-.04	.12	.10	-.19	-.09	-.03	-.04	-.07	-.11		-.22	-.16	-.17	.19	.00
18. Extrav	.01	.16	.31	-.11	.10	.09	.27	.35	.26	-.09	.25	.15	.33	.40	.52	.08	-.25		.30	.21	.21	.02
19. Open	.16	.29	.01	-.05	-.02	.17	.31	.07	-.02	-.11	.16	.12	.27	.03	.12	.15	-.11	.28		.43	.24	-.23
20. Agree	.29	.20	.31	-.23	-.18	-.04	.22	.35	-.06	-.25	.00	-.09	.11	.24	.12	.08	-.21	.26	.41		.31	-.33
21. Consc	.21	-.11	-.06	.11	.08	-.02	-.03	.04	-.05	.06	.16	-.02	-.02	-.05	-.04	.19	-.21	.00	.13	.25		-.04
22. GAI	-.18	-.07	.00	.10	.25	-.02	-.11	-.01	.17	.21	.05	.12	-.06	-.01	.05	-.05	.16	.05	-.27	-.28	-.10	

Note. Bivariate correlations for female students (n = 372) are presented above the diagonal, and bivariate correlations for male students (n = 238) are presented below the diagonal; SET = Student Evaluation of Teaching; Int. = Interest; SSci. = Social Science; SServ. = Social Service; Stat = Statistics; R = Realistic; I = Investigative; A = Artistic; S = Social; E = Enterprising; C = Conventional; Conf. = confidence; Neurot = Neuroticism; Extrav = Extraversion; Open = Openness; Agree = Agreeableness; Consc = Conscientiousness; GAI = Gender Attitude Inventory; correlations greater than $|r|$ = .10 (for women) and $|r|$ = .12 (for men) are statistically significant at $p < .05$; correlations greater than $|r|$ = .13 (for women) and $|r|$ = .17 (for men) are statistically significant at $p < .01$. Higher scores indicate stronger endorsement of the construct; higher scores on the GAI indicate a more traditional attitude.

Table 3

Student Individual Differences as Predictors of SET Ratings: Regression Coefficients (All Predictors Entered Simultaneously)

Variable	β [95%CI]	$t(588)$	p
Interest			
Social Science	-.02 [-.12, .08]	-0.4	.697
Social Service	.12 [-.04, .27]	1.4	.157
Statistics	-.03 [-.12, .06]	-0.6	.540
Realistic	-.08 [-.22, .07]	-1.0	.298
Investigative	.19 [.09, .30]	3.4	.001
Artistic	-.05 [-.18, .09]	-0.7	.466
Social	-.13 [-.28, .03]	-1.6	.114
Enterprising	.07 [-.05, .18]	1.2	.250
Conventional	-.15 [-.29, -.01]	-2.2	.030
Confidence			
Realistic	-.01 [-.14, .12]	-0.1	.882
Investigative	-.24 [-.36, -.12]	-3.9	<.001
Artistic	-.02 [-.16, .12]	-0.2	.809
Social	.05 [-.08, .18]	0.8	.437
Enterprising	-.06 [-.19, .06]	-1.1	.292
Conventional	.27 [.15, .39]	4.4	<.001
Big Five Personality			
Neuroticism	-.04 [-.12, .04]	-1.0	.343
Extraversion	.00 [-.10, .11]	0.1	.932
Openness	.01 [-.08, .10]	0.3	.797
Agreeableness	.16 [.05, .26]	3.0	.003
Conscientiousness	.11 [.03, .19]	2.8	.006
Gender Attitude Inventory	-.11 [-.19, -.03]	-2.6	.009

Note. $N = 610$; SET = Student Evaluation of Teaching;
For the overall regression model $F(21, 588) = 7.7$, $p < .001$, $R^2 = .22$, 95%CI = [.137, .254], $R^2adj = .19$.

Table 4

Regression Analyses: Student Individual Differences as Predictors of SET Ratings (Predictors Entered Separately by Construct)

Variable	Regression overall				β [95%CI]	*t*	*df*	*p*
	R^2 [95%CI]	R^2adj	*F*	*df*				
Interest (BIM)	.034 [.009, .064]	.029	7.1*	3, 606				
Social Science					.00 [-.12, .13]	0.1	606	.958
Social Service					.12 [.02, .22]	2.4	606	.017
Statistics					-.12 [-.20, -.04]	-2.8	606	.005
Interest (AFPD)	.073 [.031, .109]	.063	7.9*	6, 603				
Realistic					-.27 [-.37, -.17]	-5.1	603	<.001
Investigative					.05 [-.04, .13]	1.0	603	.308
Artistic					-.03 [-.12, .06]	-0.7	603	.468
Social					.10 [.01, .18]	2.3	603	.023
Enterprising					.01 [-.11, .12]	0.2	603	.879
Conventional					.03 [-.08, .14]	0.6	603	.540
Confidence (AFPD)	.090 [.044, .129]	.081	9.9*	6, 603				
Realistic					-.15 [-.25, -.05]	-2.9	603	.004
Investigative					-.16 [-.25, -.07]	-3.5	603	.001
Artistic					-.05 [-.14, .05]	-1.0	603	.329
Social					.15 [.06, .24]	3.4	603	.001
Enterprising					-.06 [-.16, .05]	-1.1	603	.288
Conventional					.23 [.14, .32]	4.9	603	<.001
Big Five Personality	.124 [.073, .170]	.116	17.1*	5, 604				
Neuroticism					-.01 [-.08, .07]	-0.2	604	.864
Extraversion					-.01 [-.08, .07]	-0.2	604	.876
Openness					-.01 [-.09, .07]	-0.3	604	.787
Agreeableness					.29 [.20, .38]	6.7	604	<.001
Conscientious					.14 [.06, .22]	3.4	604	.001
Gender Attitude Inv.	.072 [.037, .116]	.071	47.5*	1, 608	-.27 [-.19, -.35]	-6.9	608	<.001

Note. $N = 610$; SET = Student Evaluation of Teaching; BIM = Basic Interest Markers; AFPD = Alternate Forms Public Domain; Inv. = Inventory.
*statistically significant at $p < .001$.

Table 5

Descriptive and Inferential Statistics for All Student Individual Differences

Variable	Women (n = 372)		Men (n = 238)			
	M	*SD*	*M*	*SD*	*t*(608)	*p*
Interest						
Social Science	3.40	0.67	3.24	0.68	2.9	.004
Social Service	3.68	0.66	3.06	0.72	10.9	< .001
Statistics	2.23	0.86	2.74	0.88	-7.2	< .001
Realistic	1.86	0.69	2.65	0.85	-12.6	< .001
Investigative	2.79	0.89	3.10	0.85	-4.2	< .001
Artistic	2.85	0.87	2.93	0.84	-1.1	.277
Social	3.52	0.66	3.03	0.71	8.6	< .001
Enterprising	2.84	0.76	2.82	0.71	0.2	.812
Conventional	2.39	0.80	2.70	0.75	-4.8	< .001
Confidence						
Realistic	2.15	0.78	3.16	0.86	-14.9	< .001
Investigative	2.26	0.88	2.61	0.93	-4.7	< .001
Artistic	2.44	0.83	2.69	0.83	-3.5	< .001
Social	3.22	0.83	2.83	0.83	5.8	< .001
Enterprising	2.80	0.81	2.90	0.82	-1.3	.178
Conventional	2.98	0.87	3.31	0.81	-4.8	< .001
Big Five Personality						
Neuroticism	2.97	0.66	2.72	0.69	4.5	< .001
Extraversion	3.23	0.72	3.04	0.71	3.3	.001
Openness	3.49	0.53	3.55	0.56	-1.4	.155
Agreeableness	3.99	0.55	3.64	0.56	7.6	< .001
Conscientiousness	3.49	0.57	3.35	0.53	3.1	.002
Gender Attitude Inventory	2.36	0.52	2.73	0.53	-8.5	< .001

Note. All variables are measured on a 5-point scale; higher numbers indicate higher levels of interest, confidence, and a stronger endorsement of the respective Big Five personality trait; higher scores on the Gender Attitude Inventory indicate a more traditional attitude regarding gender roles.
Mean differences are significant at p = .002 when adjusting for the number of comparisons made.

APPENDIX A
Instructor vignettes and rating scale.

Vignette version 1: male, psychology

Dr. Robert Smith is a tenured associate professor in the department of psychology. He received his PhD in psychology in 1992 from a Big-10 university. He has been teaching college-level courses for over 10 years. His teaching load over time has included both graduate and undergraduate courses in different areas of psychology. He is currently teaching an introductory course in mental health counseling. He has taught this class repeatedly over the past years. Apart from teaching, Dr. Smith also has an established research program in the area of counseling psychology, mentoring both graduate and undergraduate students, and he is an active member of the American Psychological Association.

Vignette version 2: female, psychology

Dr. Roberta Smith is a tenured associate professor in the department of psychology. She received her PhD in psychology in 1992 from a Big-10 university. She has been teaching college-level courses for over 10 years. Her teaching load over time has included both graduate and undergraduate courses in different areas of psychology. She is currently teaching an introductory course in mental health counseling. She has taught this class repeatedly over the past years. Apart from teaching, Dr. Smith also has an established research program in the area of counseling psychology, mentoring both graduate and undergraduate students, and she is an active member of the American Psychological Association.

Vignette version 3: male, statistics

Dr. Robert Smith is a tenured associate professor in the department of statistics. He received his PhD in statistics in 1992 from a Big-10 university. He has been teaching college-level courses for over 10 years. His teaching load over time has included both graduate and undergraduate courses in different areas of statistics and research methods. He is currently teaching an introductory course in statistics. He has taught this class repeatedly over the past years. Apart from teaching, Dr. Smith also has an established research program in the area of quantitative methodology, mentoring both graduate and undergraduate students, and he is an active member of the American Statistical Association.

Vignette version 4: female, statistics

Dr. Roberta Smith is a tenured associate professor in the department of statistics. She received her PhD in statistics in 1992 from a Big-10 university. She has been teaching college-level courses for over 10 years. Her teaching load over time has included both graduate and undergraduate courses in different areas of statistics and research methods. She is currently teaching an introductory course in statistics. She has taught this class repeatedly over the past years. Apart from teaching, Dr. Smith also has an established research program in the area of quantitative methodology, mentoring both graduate and undergraduate students, and she is an active member of the American Statistical Association.

Instructions and rating scale (for all four vignette versions):

Please read the following description of a college instructor. Imagine that you are considering taking Dr. Smith's statistics course. Based on the description above, please indicate to what extent you would expect Dr. Smith to…

…be knowledgeable about the subject she is teaching
…be well organized and prepared
…be available for help outside of class
…be enthusiastic about the subject she is teaching
…be effective in communicating course objectives and requirements
…create a respectful and comfortable classroom environment
…be fair and accommodating to all students in the class
…be interested in helping students learn

Note. Rated on a Likert scale from 1 = *not at all* to 7 = *extremely*.

APPENDIX B

Demographic items and measures in the order in which they are administered (items within each measure were presented to participants in random order).

1) Demographic items:

Year and month of birth (used to randomly assign participants to conditions)
Age
Sex
- Male
- Female

Racial-ethnic identity
- Caucasian/White
- African-American
- Hispanic-American
- Asian-American
- Native American
- International Student
- Other (example: bi-racial)

Year in school
- Freshman (< 30 credits)
- Sophomore (30-60 credits)
- Junior (60-90 credits)
- Senior (> 90 credits)
- Graduate student

Current major
GPA
How many times how you changed your major?
Previous majors
How motivated are you to complete your bachelor's degree?
- I am unmotivated to complete my bachelor's degree
- I am somewhat motivated to complete my bachelor's degree
- I am motivated to complete my bachelor's degree
- I am very motivated to complete my bachelor's degree

What are your current educational aspirations?
- Some college/no degree
- Associate degree
- Bachelor's degree
- Master's degree
- Doctorate (Ph.D.)
- Medical degree (M.D.)
- Law degree (J.D.)

How many psychology courses have you taken (including those currently enrolled in)?

How many statistics or research methods courses have you taken (including those currently enrolled in)?

2) BIMs (Basic Interest Markers):

Statistics

1. Solve an algebraic equation
2. Develop statistical formulas
3. Understand applications of calculus
4. Learn about a new branch of statistics
5. Graph an equation
6. Take a course in advance statistics
7. Solve geometric proofs
8. Apply statistical techniques to practical problems
9. Calculate the probability of winning a contest
10. Use mathematical theorems to solve problems

Social Sciences

1. Learn about human behavior
2. Develop a theory about human behavior
3. Investigate cultural practices
4. Conduct social science experiments
5. Study child-rearing problems
6. Compare cultural differences among groups
7. Analyze the effects of discrimination on minority groups
8. Review the interpersonal relationship literature
9. Study class structures of a society
10. Study intersections among people in a group

Social Service

1. Assist people with disabilities to find employment
2. Help families to adopt a child
3. Counsel families in crisis
4. Help the homeless find shelter
5. Help people find community resources
6. Provide childcare services
7. Organize a social support group
8. Volunteer for a community service center
9. Help children from disadvantaged background adjust to school
10. Counsel clients with personal problems
11. Provide services to individuals with disabilities
12. Help people overcome social problems

3) AFPD (Alternate Forms Public Domain) Interest Markers:

Realistic

1. Test the quality of parts before shipment
2. Lay brick or tile
3. Work on an offshore oil-drilling rig
4. Assemble electronic parts
5. Operate a grinding machine in a factory
6. Fix a broken faucet
7. Assemble products in a factory
8. Install flooring in houses

Investigative

1. Study the structure of the human body
2. Study animal behavior
3. Do research on plants or animals
4. Develop a new medical treatment or procedure
5. Conduct biological research
6. Study whales and other types of marine life
7. Work in a biology lab
8. Make a map of the bottom of an ocean

Artistic

1. Conduct a musical choir
2. Direct a play
3. Design artwork for magazines
4. Write a song
5. Write books or plays
6. Play a musical instrument
7. Perform stunts for a movie or television show
8. Design sets for plays

Social

1. Give career guidance to people
2. Do volunteer work at a non-profit organization
3. Help people who have problems with drugs or alcohol
4. Teach an individual an exercise routine
5. Help people with family-related problems
6. Supervise the activities of children at a camp
7. Teach children how to read
8. Help elderly people with their daily activities

Enterprising

1. Sell restaurant franchises to individuals
2. Sell merchandise at a department store
3. Manage the operations of a hotel

4. Operate a beauty salon or barber shop
5. Manage a department within a large company
6. Manage a clothing store
7. Sell houses
8. Run a toy store

Conventional

1. Generate the monthly payroll checks for an office
2. Inventory supplies using a hand-held computer
3. Use a computer program to generate customer bills
4. Maintain employee records
5. Compute and record statistical and other numerical data
6. Operate a calculator
7. Handle customers' bank transactions
8. Keep shipping and receiving records

4) GAI (Gender Attitude Inventory):

Traditional Stereotypes

1. Men are more competitive than women.
2. Men are generally more adventurous than women are.
3. Men are generally more egotistical than women.
4. On the average, men are more arrogant than women.
5. Women are more gentle than men.
6. Men are more independent than women.
7. Men are more sure of they can do than women are.
8. Compared to men, women tend to be gullible.
9. Compared to men, women are more able to devote themselves completely to others.
10. Compared to men, women tend to be weak.

Family Roles

1. It's all right for the woman to have a career and the man to stay at home with the children.
2. I approve of a wife entering the labor force and leaving her husband at home to care for the children.
3. I would not respect a man if he decided to stay at home and take care of his children while his wife worked.
4. The wife should have primary responsibility for taking care of the home and children.
5. A woman should work only if she can do so without interfering with her domestic duties.
6. The husband should have primary responsibility for support of the family.
7. In marriage, the husband should take the lean in decision making.

8. Working women should not be expected to sacrifice their careers for the same of home duties to any greater extent than men.
9. Women should be concerned with their duties of child-rearing and house tending, rather than with desires for professional and business careers.
10. As head of the household, the husband should have more responsibility for the family's financial plans than his wife.
11. Care of children should be shared equally by both spouses.

Differential Work Roles

1. There are almost no jobs which should be closed to women because of physical requirements.
2. Many jobs should be closed to women because of the physical requirements.
3. Men and women are better suited to different kinds of occupations due to physical strength.
4. All occupations should be equally accessible to both men and women.
5. There are many jobs in which men should be given preference over women in being hired and promoted.
6. There are some professions and types of business that are more suitable for men than women.
7. In today's world the idea of "women's work" and "men's work" makes no sense.
8. It is appropriate to divide work into "men's work" and "women's work".
9. A woman's work and a man's work should be fundamentally different.

5) AFPD (Alternate Forms Public Domain) Confidence Markers:

Realistic

1. Perform lawn care services
2. Repair household appliances
3. Build kitchen cabinets
4. Guard money in an armored car
5. Operate a machine on a production line
6. Repair and install locks
7. Set up and operate machines to make products
8. Build a brick walkway

Investigative

1. Study ways to reduce water pollution
2. Study the movement of planets
3. Examine blood samples using a microscope
4. Study genetics
5. Determine the infection rate of a new disease
6. Diagnose and treat sick animals
7. Do laboratory tests to identify diseases
8. Develop a new medicine

Artistic

1. Paint sets for plays
2. Sing in a band
3. Act in a movie
4. Conduct a symphony orchestra
5. Create special effects for movies
6. Compose or arrange music
7. Write reviews of books or plays
8. Draw pictures

Social

1. Work with juveniles on probation
2. Take care of children at a day-care center
3. Teach an elementary school class
4. Work with mentally disabled children
5. Teach disabled people work and living skills
6. Organize field trips for disabled people
7. Teach a high-school class
8. Help conduct a group therapy session

Enterprising

1. Sell newspaper advertisements
2. Sell a soft drink product line to stores and restaurants
3. Give a presentation about a product you are selling
4. Sell hair-care products to stores and salons
5. Negotiate contracts for professional athletes
6. Manage a retail store
7. Start your own business
8. Market a new line of clothing

Conventional

1. Keep inventory records
2. Keep accounts payable/receivable for an office
3. Calculate the wages of employees
4. Develop a spreadsheet using computer software
5. Assist senior level accountants in performing bookkeeping tasks
6. Transfer funds between banks using a computer
7. Enter information into a database
8. Keep records of financial transactions for an organization

6) IPIP (International Personality Item Pool) Big Five Markers:

Neuroticism

1. Get stressed out easily

2. Often feel blue
3. Worry about things
4. Am easily disturbed
5. Get upset easily
6. Get irritated easily
7. Seldom feel blue
8. Am relaxed most of the time
9. Have frequent mood swings
10. Change my mood a lot

Extraversion

1. Have little to say
2. Don't like to draw attention to myself
3. Don't mind being the center of attention
4. Talk to a lot of different people at parties
5. Keep in the background
6. Start conversations
7. Don't talk a lot
8. Am quiet around strangers
9. Am the life of the party
10. Feel comfortable around people

Agreeableness

1. Sympathize with others' feelings
2. Take time out for others
3. Feel others' emotions
4. Make people feel at ease
5. Feel little concern for others
6. Insult people
7. Am interested in people
8. Am not interested in other people's problems
9. Have a soft heart
10. Am not really interested in others

Openness

1. Have a vivid imagination
2. Am not interested in abstract ideas
3. Do not have a good imagination
4. Am quick to understand things
5. Use difficult words
6. Spend time reflecting on things
7. Am full of ideas
8. Have a rich vocabulary
9. Have difficulty understanding abstract ideas
10. Have excellent ideas

Conscientiousness

1. Am always prepared
2. Pay attention to details
3. Get chores done right away
4. Often forget to put things back in their proper place
5. Like order
6. Shirk my duties
7. Follow a schedule
8. Am exacting in my work
9. Leave my belongings around
10. Make a mess of things

APPENDIX C
Consent form.

CONSENT FORM FOR: EDUCATIONAL EXPERIENCE STUDY

This form describes a research project. It has information to help you decide whether or not you wish to participate. Research studies include only people who choose to take part—your participation is completely voluntary. Please discuss any questions you have about the study or about this form with the project staff (see contact information below) before deciding to participate.

Who is conducting this study?
This study is being conducted by Verena Bonitz and Lisa Larson in ISU's department of psychology.

Why am I invited to participate in this study?
You are being asked to take part in this study because you indicated your interest in participating in psychological studies in exchange for experimental credit counting towards a psychology course.

What is the purpose of this study?
The purpose of this study is to learn about different factors that contribute to a student's educational experience in different college courses.

What will I be asked to do?
If you agree to participate, you will complete a series of self-report questions. Participation will require a time commitment of 50 minutes or less, which corresponds to one (1) experimental credit point. You will be asked to complete the questionnaire in an online format. The questionnaire content relates to different aspects of a student's educational experience in a college course, and the factors (such as personality and vocational interests) that contribute to the quality of this experience. You may skip any question that you do not wish to answer or that makes you feel uncomfortable.

What are the possible risks and benefits of my participation?
Risks - There are no foreseeable risks at this time from participating in this study.
Benefits – You may not receive any direct benefit from taking part in this study. However, it is hoped that the information gained in this study will benefit society by helping to better understand the various factors that determine students' educational experiences in a college course.

What measures will be taken to ensure the confidentiality of the data or to protect my privacy?
Records identifying participants will be kept confidential to the extent permitted by applicable laws and regulations and will not be made publicly available. However, federal government regulatory agencies, auditing departments of Iowa State University, and the

Institutional Review Board (a committee that reviews and approves human subject research studies) may inspect and/or copy your records for quality assurance and data analysis. These records may contain private information.
To ensure confidentiality to the extent permitted by law, the following measures will be taken: Each participant will be assigned a unique code that will be used on forms instead of their name. Only the investigators will have access to the data, and data and identifying information will be stored separately in locked filing cabinets. Electronic files will be stored on password-protected computers. If the results are published, your identity will remain confidential.

Will I incur any costs from participating or will I be compensated?
You will not have any costs from participating in this study. You will be compensated for participating in this study with 1 experimental credit point that applies towards your respective psychology course.

What are my rights as a human research participant?
Your participation in this study is completely voluntary and you may refuse to participate or leave the study at any time. If you decide to not participate in the study or leave the study early, it will not result in any penalty or loss of benefits to which you are otherwise entitled. The participation in this study is only one option of obtaining experimental credit; other options are noted on your course syllabus.

Whom can I call if I have questions or problems?
You are encouraged to ask questions at any time during this study.

- For further information about the study contact Verena Bonitz (phone: 515-294-8480; email: vsbonitz@iastate.edu) or Lisa Larson (phone: 515-294-1487; email: lmlarson@iastate.edu).
- If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, (515) 294-4566, IRB@iastate.edu, or Director, (515) 294-3115, Office for Responsible Research, 1138 Pearson Hall, Iowa State University, Ames, Iowa 50011.

Consent and Authorization Provisions
Please print a copy of this informed consent form for your records. By continuing to the next page, you consent to participate in this study.

APPENDIX D
Study posting form.

STUDY POSTING FORM

Ann Schmidt MUST receive a copy of this form before you send an activation request.

PRINCIPAL INVESTIGATOR (*Faculty Supervisor*): Lisa Larson

RESEARCHERS: Verena Bonitz, Lisa Larson

STUDY NAME & NUMBER: Educational Experience

BRIEF ABSTRACT: Students' educational experience in college courses and the factors contributing to it.

STUDY DESCRIPTION (*Must be exactly as approved by IRB*):

This study investigates the influence of different factors such as interests and personality that might play a role in how students experience their course instructors.

ELIGIBILITY REQUIREMENTS: 18 years or older

DURATION (*Minimum 50min.): 50 minutes*

CREDITS: 1

PREPARATION:

IRB APPROVAL CODE:

IRB APPROVAL EXPIRATION:

IS THIS AN ONLINE STUDY? yes

ACKNOWLEDGEMENT

I am heartily thankful to the following individuals whose advice, encouragement, and support throughout the course of my studies have allowed me to refine my interests, and to develop a unique and highly satisfying career path. First and foremost, I would like to thank my advisor Lisa Larson for her guidance on all things academic, for taking a sincere interest in my professional and personal development, and for her trust in my abilities, which allowed me to freely pursue my research and teaching interests.

I am grateful to Douglas Bonett, not only for his advice on statistics, but his thoughtful and honest approach to mentoring throughout the graduate program and the job search. I am thankful to Patrick Armstrong for providing me with many opportunities for collaboration on research projects, for his guidance in his function as a teaching advisor, and the many ways in which he supported me during the job search. I am grateful to David Vogel for his continued support during my enrollment in the counseling program, and whose effort and flexibility made it possible for me to transition to a more compatible career path. I would like to thank Donna Kienzler for her refreshing views and sound advice as part of the PFF experience. I am particularly delighted that she agreed to serve on my dissertation committee.

Further, I am indebted to Karen Scheel for her friendship and emotional support. I would also like to acknowledge Madeleine Henry, who introduced me to the joys of learning classical Greek, and who greatly facilitated the logistics concerning my research project on classical language learning.