5-14-2003

The psychometrics of student evaluations of instructors and courses at Eastern Michigan University

Thomas P. Proctor

THE PSYCHOMETRICS OF STUDENT EVALUATIONS OF INSTRUCTORS AND COURSES AT EASTERN MICHIGAN UNIVERSITY

By

Thomas P. Proctor

Thesis

Submitted to the Department of Psychology

Eastern Michigan University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology with a concentration in General Experimental

Thesis Committee:

John R. Knapp, Ph.D., Chairperson

Dennis Delprato, Ph.D.

David C. S. Richard, Ph.D.

May 14, 2003

Ypsilanti, Michigan

THE PSYCHOMETRICS OF STUDENT EVALUATIONS OF INSTRUCTORS AND COURSES AT EASTERN MICHIGAN UNIVERSITY

by

Thomas P. Proctor

APPROVAL

John R. Knapp, Ph.D. Date

Chairperson

Dennis Delprato, Ph.D. Date

Committee Member

David C. S. Richard, Ph.D. Date

Committee Member

Kenneth W. Rusiniak, Ph.D. Date

Department Head

Robert Holkeboer, Ph.D. Date

Associate Vice President for Graduate Studies and Research

Dedication

In memory of my mom, May 12, 1943 to June 28, 2003, who always instilled the importance of education in her children.

Acknowledgement

I would like to thank Dr. Knapp and the academic affairs division of Eastern Michigan University for providing me with the data used in this study. I would also like to thank Dr. Knapp for his assistance in compiling the data into a usable file from the numerous original VMS data files. I would also like to thank my thesis committee for their guidance on this project.

Finally, I would also like to thank Chris Golla for his assistance in generating the random samples used in the second study. And I would like to thank Yelena Gutman for her assistance in typing and organizing my literature review notes.

Abstract

The study applied Classical Test Theory to the student evaluation forms for 25 departments to assess reliability. The study also applied Generalizability Theory to assess the reliability of the Psychology Department evaluation form. Regression analysis on the Psychology Department form assessed the effect of absolute expected grades on student ratings of teaching effectiveness and courses. The results show that the reliability of the 25 department forms is very high, exceeding .80 for each form. Generalizability theory indicates the Psychology Department form to be reliable for assessment of student ratings of the effectiveness of teaching but not necessarily of courses. Results suggest at least five items from five or more courses would be preferable to obtain reliable results of student ratings of teaching effectiveness. Regression analysis shows absolute expected grade did not significantly account for any variance in overall student ratings of teaching effectiveness or overall course ratings.

Table of Contents

List of Tables

Table 1

Notational Conventions

Abbreviation/Symbol	Definition
x	Crossed with
:	Nested within
$E\rho_{\delta}^{2}$ or ρ^{2}	Generalizability Coefficient
ρ_{Δ}^{2} or Φ	Index of dependability
α	Alpha coefficient in classical test theory; an effect in generalizability theory
τ	Object of measurement
ν_{α}	Score effect for α
X_{α}	An observable score for α
μ_{α}	Population or universe mean score forα
$\sigma^{2}(\tau)$	Universe score variance
$\sigma^{2}(\delta)$	Relative error variance
$\sigma^{2}(\Delta)$	Absolute error variance
$\sigma^{2}(\alpha)$	Random effects variance component for α
E(*MS*)	Expected mean square
n′	D study sample size for facet

Introduction

Student evaluations of teacher effectiveness are being used for numerous reasons throughout universities and colleges in the United States. Many detractors question the use of such ratings for determining personnel decisions such as promotion, retention, merit raises, and tenure. Detractors argue that such ratings are invalid, unreliable, or plagued with bias. The scant research performed on the current student ratings of teachers and courses in use at Eastern Michigan University has not answered the questions raised.

Background

Student evaluations of teaching have become a staple of higher education over the past 30 to 40 years (Braskamp & Ory, 1994; Cashin, 1999; Centra, 1993; McKeachie, 1990; Miller, 1974; Ory, 2000; Seldin, 1999; Williams & Ceci, 1997). Miller (1974) states that finance, governance, and accountability are the three major reasons to evaluate faculty. Centra (1993) points out that legislators, parents, and students want evidence that institutions provide the best possible education. Administrators also use student ratings to make inferences regarding decisions on retention, tenure, and merit pay raises (Ory, 2000). Frances and Gruber (1981) recommend that because of these reasons, academic departments and institutions should research their student evaluation process.

During the 1960s, student groups obtained course evaluations and published the results (Centra, 1993; Williams & Ceci, 1997). Their actions were in part a protest against campus policies that protected irrelevant curricula and uninspired teachers (Centra, 1993). Students began to see themselves as consumers (Centra, 1993). As a result, teaching evaluations in the 1960s responded to student demands for public accountability (Ory, 2000). In the 1970s, the purpose of performing evaluations was to obtain student feedback that could be used to assist teachers improve and develop; in the 1980s and 1990s, teaching evaluation was driven by administrative rather than faculty or student needs (Ory). Most recently, a resurgence of interest in improving undergraduate education, demands for accountability, and the demand by the legal system for improved evaluations drive the process (Ory).

Teacher assessment has always been controversial and is one of the most sensitive issues on campuses (Braskamp & Ory, 1994). Those who discourage the use of student evaluations claim that students cannot make consistent evaluations, only qualified researchers can judge teaching, student ratings constitute a popularity contest, ratings are biased, and grades bias ratings (Cashin, 1999). Critics also argue that student evaluations fail to reflect long-term effects of instruction and that students will only appreciate more demanding teachers years later (Centra, 1993).

The legal implication of student evaluations is a more recent discussion appearing in the literature. Haskell (1997) has reviewed legal decisions regarding the use of student evaluations in personnel decisions. He argues that the university administration, in effect, uses student evaluations as a tool to ensure that instructors teach material that is acceptable to the university and not what the instructor deems as acceptable. For this reason Haskell believes that the use of student evaluations impinges upon academic freedom even though the courts have not supported this argument. Ory (2000) points out that decisions about faculty tenure, promotion, and merit raises are frequently being challenged in courts and that administrators frequently rely upon student evaluations to support decisions in the event of litigation (Ory).

Seldin (1998) reports that in a survey of 598 academic deans, 97.5% reported the evaluation of classroom teaching was the most important measure of faculty performance. He also states that in 1998, 76% of the colleges surveyed use some form of systematic ratings, completed by students, colleagues, or administrators (Seldin, 1999). Unfortunately, only 14% of those colleges surveyed engage in any research on the surveys (Seldin, 1999).

The first published research on student evaluation of teaching was done by Herman Remmers (Centra, 1993) in a series of reports on the Purdue Rating Scale for Instructors (Remmers, 1930, 1934). In the series of reports, Remmers addressed questions about validity, the correlation of student grades to ratings of teachers, and the reliability of ratings. Remmers (1930) found significant correlations between student grades and student ratings of instructors. Remmers (1934) also found the Purdue Rating Scale to be a reliable measure of teaching performance. Since then, numerous studies have been conducted in attempts to clarify his findings.

Considerable research has examined factors that could affect students' ratings of teacher effectiveness. Some of the factors that previous research has examined are instructor characteristics, such as teacher personality, teacher rank, and teacher gender; student characteristics, such as achievement and grades (both expected and actual) and student gender; and course characteristics, such as class level, class size, required course or an elective, time of day the course meets, and course workload (Cashin, 1995; Centra, 1993; Miller, 1974; Gigliotti & Buchtel, 1990; Gilmore, Swerdlik & Beehr, 1980; Greenwald & Gillmore, 1997a, 1997b; Ronco, 1999; Williams & Ceci, 1997). The research shows that most of the variables account for little variance in student ratings, with class size, workload, and expected grade presenting the largest factors influencing student ratings (Gilmore et al., 1980; Greenwald & Gillmore, 1997a, 1997b).

Validity and Potential Contaminants

Validity is the extent to which a measure truly measures what the researcher intends it to (Neale & Liebert, 1986). For student ratings to be valid, they must be both reliable and relevant (Aubrecht, 1979). Various research investigations have addressed validity questions.

Aubrecht (1979) found that student ratings strongly correlated with colleague and administrator ratings. Cohen (1981) performed a meta-analysis of 41 studies of multi-section courses and concluded that student ratings are valid for rating teacher effectiveness. Cohen (1986) performed another meta-analysis of 47 studies and found that student ratings were correlated positively to student learning, suggesting validity of student ratings. Cashin (1999) also concluded that student ratings are valid for most uses in rating teacher effectiveness, based on numerous Individual Development Educational Assessment (IDEA) studies, a widely used ratings system developed by researchers at Kansas State University.

Ronco (1999) reports that several potential contaminant variables have been considered related to student ratings, among them student motivation, expected grades, course level, academic discipline, and workload. Class size has been found to have some relationship, but Ronco says it should not be considered a bias because teachers are usually more effective in smaller classes. Faculty age, gender, race, research productivity, or student age, gender, class level or GPA have not been found influential on student ratings of teacher effectiveness (Cashin, 1995; Centra, 1993; Marsh, 1987). However, two recent studies have shown that instructor enthusiasm and grading leniency are potential contaminants, which question the validity of student ratings of teacher effectiveness.

Instructor Enthusiasm. Williams and Ceci (1997) performed a study in which one of the authors taught a course in the fall term and then again in the spring term. Between the terms, the instructor attended a seminar on how to improve enthusiasm. The students in the courses were compared for demographic variables and no significant difference between the students was found (Williams & Ceci, 1997). Significant rating differences of the instructor were found between the two terms on several items (Williams & Ceci). The professor recorded the lectures during the fall course and reviewed these recordings prior to each class lecture in the spring term (Williams & Ceci). The rating form asked the students to rate the instructor, the course and what they learned (Williams & Ceci). The students in the spring semester believed they had learned markedly more compared to the students in the fall course even though the final grades, used as an external indicator, were nearly identical for both courses (Williams & Ceci). These authors conclude that factors unrelated to teaching effectiveness exert a sizable influence on ratings and that student evaluations can be reliable but not valid (Williams & Ceci).

Williams and Ceci do not mention the possibility that a more enthusiastic teacher may have lower absenteeism of students and thus receive a higher response rate. Did the professor in the study actually increase his clarity because he reviewed previous lectures and unknowingly alter the delivery of the lecture in some systematic manner that could account for the difference? Were final grades calculated differently based on different assignments or variations in course tests? Perhaps there are other systematic differences. Clearly additional studies and replication must be done prior to considering student ratings as reliable but not valid.

Grading Leniency. Grading generates the most suspicion about the validity of student ratings of teacher effectiveness (Marsh, 1984). Numerous studies have attempted to answer this question. Remmers (1930) examined the relationship between student grades and faculty ratings and found that the students with higher scores on achievement tests rated teachers higher.

Holmes (1972) found that when students’ grades disconfirmed the grade they expected, the students rated the instructor lower. Another study found that students assigned a lower grade rated the instructor as less effective (Worthington & Wong, 1979). Vasta and Sarmiento (1979) randomly assigned grades to students on examinations and found that the expected grade was significantly positively correlated to instructor evaluations. Additional studies have shown that expected grades correlated positively to students’ ratings of instructor effectiveness (Hudson, 1989; Stapleton & Murkison, 2001; Stumpf & Freedman, 1979).

Chacko (1983) performed a study in which two sections of a course received a rating form prior to the midterm and then the week after the students learned their grades on the test. The pre-midterm ratings were not significantly different from one another. The control group’s grades were adjusted based on questions missed by 75% or more of the class and resulted in some students getting “As.” The treatment group got no adjustment based on common questions missed by the class and no students in this group received “As” (Chacko). The post midterm ratings between the two groups differed significantly on several dimensions. Particularly, the treatment group rated the instructor lower in effectiveness and harsher on grading (Chacko). It was concluded that grading policy is a potential contaminant of student ratings.

Greenwald and Gillmore (1997a) have proposed five theories to account for the positive relationship between grades and student ratings: (a) teaching effectiveness influences both; (b) students' general academic motivation influences both; (c) students' course specific motivation influences both; (d) students infer both course quality and their own abilities from grades; and (e) high ratings are given by students in appreciation of lenient grading. The first three theories explain the grades-ratings correlation by virtue of a third variable influencing the relationship (Greenwald & Gillmore, 1997a). The last two theories assume that grades do have a causal relationship with student ratings of teacher effectiveness (Greenwald & Gillmore, 1997a).

Greenwald and Gillmore (1997a) found that the data obtained at the University of Washington repeatedly showed a relationship between expected grades and student ratings. However, absolute expected grades, 0.0 to 4.0, had less relationship to student ratings than relative expected grades, the relationship of the expected grade to the students' average grade in other courses (Greenwald & Gillmore, 1997a). They also found a substantial negative relationship between expected grades and course workload (Greenwald & Gillmore, 1997a, 1997b). Based on structural modeling of the data, Greenwald and Gillmore (1997a, 1997b) determined that only the last theory accounts for the positive relationship between grades and student ratings.

These findings indicate that ratings fail to discriminate between the rating of the instructor, the students' satisfaction with their grade, or course workload (Greenwald & Gillmore, 1997a, 1997b). Some may take these results as a reason to dispose of student ratings altogether. Since no other cost effective alternatives currently exist, several reasons warrant not abandoning ratings. Student ratings still contain important information about student beliefs, and the evidence of convergent validity cannot be dismissed (Greenwald & Gillmore, 1997a). Greenwald and Gillmore (1997a) do suggest that it is important to measure course workload as well as expected grade because the data then can be used to remove any grading leniency as a contaminant.

In a more recent study (Chang, 2000), regression analysis determined what variables account for variance in predicting student ratings of teacher effectiveness. The sample included all student responses within the education department for one academic year and excluded forms that did not have responses for the key variables. In this study, expected grade did not account for a significant amount of variance in student ratings, although grading standard did account for an additional 1% of the variance.

In the Chang (2000) study, expected grade was the final course grade students expected. Greenwald and Gillmore (1997a) found that relative expected grade influenced student ratings more than absolute expected grades. This may account for why the study did not find that expected grade contributed to any significant variance.

Reliability

The reliability of student ratings is one of the most researched topics regarding teacher effectiveness (Aubrecht, 1979). Psychology and education generally use classical test theory to address issues of measurement (Brennan, 1992, 2001). In classical theory, reliability is the proportion of observed score variance attributed to "true" score differences (Brennan, 1992, 2001; Kane, Gillmore, & Crooks, 1976).

Reliability within the student ratings literature generally refers to internal consistency or inter-rater agreement (Cashin, 1995). The stability of ratings over time and agreement among raters are also important gauges for measuring reliability (Centra, 1993). In most cases, coefficient alpha is used to assess reliability (Sun, Valiga, Gao, & ACT, 1997). Another procedure used for reliability indices is inter-rater correlations within the classical test theory (Sun & Valiga, 1997). However, classical test theory receives criticism for not being able to deal with multiple sources of random error and being cumbersome and confusing when partitioning variance into more than two components (Gillmore, Kane, & Naccarato, 1976; Sun & Valiga, 1997). Tinsley and Weiss (2000) state that reliability, as defined by classical test theory, is not a generalizable property of measurements but descriptive of a set of data.

Generalizability theory is an extensive conceptual framework and powerful statistical procedure to address measurement issues (Brennan, 2001). It liberalizes the classical test theory by allowing the researcher additional methods to investigate the multiple sources of error, as well as to partition the error into more specific components than classical test theory allows (Brennan, 1992, 2001; Marcoulides, 2000). Specifically, generalizability theory extends the classical test theory in allowing applied researchers to generalize about a person’s behavior (Marcoulides, 2000). For these reasons, Sun et al. (1997) state that generalizability theory is more appropriate, effective, and efficient for providing meaningful reliability of student ratings.

In generalizability theory, a universe is the condition of measurement, and the population is the object of measurement (Brennan, 1992, 2001; Marcoulides, 2000). For example, questions and student raters would define a universe within student ratings. Teachers, or courses, would be examples of a population. A facet is a set of similar conditions within the acceptable universe (i.e., items or raters) (Brennan, 1992, 2001; Marcoulides, 2000). The universe score replaces the true score within classical test theory and places emphasis on the idea that there are many universes to which a researcher can generalize (Gillmore et al., 1976).

Marcoulides (2000) states that one of the advantages of using generalizability theory versus classical test theory is that the researcher can distinguish between two types of error variance. The first type of error variance is referred to as relative error variance, also called δ-type error, that is considered when the researcher wants to make decisions about individual differences between objects of measurements (Marcoulides, 2000). Relative error is the difference between the person’s observed deviation score and his universe deviation score (Brennan, 1992, 2001).

The second type, absolute error variance, or Δ-type error, is used when the researcher wants to know whether a person can perform at a pre-specified level, or the researcher is interested in rank ordering and differences (Marcoulides, 2000). Absolute error is the difference between a person's observed score and his universe score (Brennan, 1992, 2001). Generalizability theory places a great deal of importance on variance components because the magnitude provides information about potential sources of error that influence the measure (Marcoulides, 2000).

Two types of studies can be performed within the framework of generalizability theory. The first, called a generalizability (G) study, refers to the initial study of a measurement procedure (Marcoulides, 2000). A G study obtains estimates of variance components for the universe of admissible observations (Brennan, 1992, 2001). The second type of study, the decision (D) study (Marcoulides, 2000), emphasizes the estimation, use, and interpretation of variance components (Brennan, 1992, 2001). Gillmore et al. (1976) point out that the G study is distinguished from the D study, which is an extension of the Spearman-Brown prophecy formula.

Within the generalizability theory framework, several coefficients can be calculated. The first is the generalizability coefficient and is defined as:

$$E\rho^2_\delta = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \text{ (Brennan, 1992; Marcoulides, 2000).} \qquad (1)$$

In the equation, $\sigma^2(\tau)$ is equal to universe score variance, and $\sigma^2(\delta)$ is equal to the relative error variance. The generalizability coefficient is analogous to the reliability coefficient in classical test theory (Brennan, 1992, 2001; Gillmore et al., 1976; Marcoulides, 2000).

The second coefficient, index of dependability, is defined as:

$$\rho^2_\Delta = \Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \text{ (Brennan, 1992; Marcoulides, 2000).} \quad (2)$$

In the equation, $\sigma^2(\Delta)$ is equal to the absolute error variance. Brennan (1992) points out that the index Φ is used when scores prompt interpretations based on domain-referenced or criterion-referenced situations such as cut-off scores.

Gillmore et al. (1976) and Gillmore, Kane, and Naccarato (1978) applied generalizability theory to the University of Washington instructional assessment program, which involved several different forms, each containing several common items. Because the main purpose of the instrument was to provide evaluative information about courses and instructors, class means are more appropriate than individual student responses (Gillmore et al., 1976). Because all students answer the same questions but different students rate each course, the design of the study nests students in the class crossed with items (Gillmore et al., 1976). The design was balanced by randomly selecting 13 students from 14 randomly selected classes within three general areas of study (Gillmore et al., 1976, 1978). To assess stability Gillmore et al. (1976, 1978) used two data sets and analyzed them separately.

Gillmore et al. (1976, 1978) used generalizability theory to define and interpret five generalizability coefficients: (a) $\rho^2(I)$, generalizing over items, which is equivalent to internal consistency in classical test theory; (b) $\rho^2(S)$, generalizing over students, corresponding to stability; (c) $\rho^2(S, I)$, generalizing over both students and items; (d) $\rho^2(T, S, I)$, generalizing over all courses a teacher might teach, all students who might enroll in courses, and all items within the domain that could be used to rate the teacher; and (e) $\rho^2(C, S, I)$, generalizing over all teachers that a course may be taught by, all students who might enroll with a specific teacher, and all items within the domain that could be used to rate the course. The coefficients listed in (c), (d), and (e) may have been missed by the use of classical test theory. The third coefficient $\rho^2(S, I)$ was favored for assessing the dependability of rating for general instruction (Gillmore et al., 1976, 1978). Gillmore et al. suggest that $\rho^2(S)$ is appropriate for assessing instructional problems. However, $\rho^2(I)$ would be useful if the researcher wanted to be confined to a particular set of items for assessing student ratings of teaching effectiveness (Gillmore et al., 1976, 1978). They (1978) suggest that $\rho^2(T, S, I)$ is used to assess the effect of the course and that $\rho^2(C, S, I)$ is used to assess the effect of the teacher.

The results of the Gillmore et al. (1976, 1978) studies indicate that the variability between student responses is much greater than that within students. As a result, a larger sample of students is more important than items to assess reliability of student ratings (Gillmore et al., 1976, 1978). Based on the findings, reliability studies of student ratings of teaching effectiveness should be based on five items with ten or more students (Gillmore et al., 1976, 1978). Gillmore et al (1976, 1978) found that when generalizing across teachers, $\rho^2(C, S, I) = .71$ using 5 courses and 20 students, and when generalizing across courses, $\rho^2(T, S, I) = .33$ using 10 items and 20 students. These findings indicate that 40% of the estimated variance in ratings was due to the teacher effect while only 6% was due to course effect (Gillmore et al., 1978). The interaction effect between courses and teachers accounts for 54% of the ratings variance (Gillmore et al.). These results suggest that the course is not a major factor in determining ratings and that the quality of teaching as rated by students could be improved by assignment of teachers to courses (Gillmore et al.).

Huang, Guo, Druva-Rouch, and Moore (1995) applied the design from the Gillmore et al. (1976, 1978) studies to determine if four different forms created from a cafeteria-style system varied in their dependability. Again, the design was balanced with 15 students randomly selected from each class as long as the class had more than 15 students (Huang et al., 1995).

Findings indicated that three of the four forms were similar in reliability but the fourth one was discernibly smaller than the other forms though all four forms exceeded the .70 reliability level (Huang et al., 1995). Cronbach’s alpha was also calculated for each form, and all forms had alphas of .9 or higher (Huang et al.). The authors noted that Cronbach’s alpha would indicate little difference in the forms, but generalizability theory indicates greater difference (Huang et al.).

In the Huang et al. (1995) study, alphas ranged from .910 to .934. Gillmore et al. (1976, 1978) suggest that based on generalizability theory internal consistency is measured by $\rho^2(I)$. The Huang et al. data show that the range of $\rho^2(I)$ is .936 to .977. Both measures suggest very high levels of internal consistency. The range of $\rho^2(S, I)$ was .737 to .894, all above the generally acceptable limit of .70 for reliability, suggesting all forms are reliable measures of overall teacher/course effectiveness (Haung et al.). This study did not generate $\rho^2(T, S, I)$ nor $\rho^2(C, S, I)$, so the effects reported within the study do not differentiate between the effect of the teacher and the course.

Purpose of the Research

The intent of the present thesis was to examine the reliability of the student rating system currently in use at Eastern Michigan University by applying classical test and generalizability theory. The thesis also attempted to determine if the expected grade reported by students contributed significantly to the overall instructor and course ratings.

Having a reliable measure is a prerequisite for a valid measure. One free of contaminants helps establish discriminant validity of the measure. Centra (1993) states, “Poor evaluation, whether of students or of faculty, renders an unfair judgment and fails to reveal shortcomings in performance” (p. 1). Though this particular research may be primarily descriptive, it will provide the groundwork for future research on the student rating system at Eastern Michigan University.

Research Questions

Study 1. To measure the reliability of items, using classical test theory, on different forms used by different departments at Eastern Michigan University. It was hypothesized that reliability using individual student responses will be lower than that obtained from class means. Further, it was hypothesized that reliability based on class means will differ for each department but the reliability for all forms analyzed will achieve reliability greater than .70.

Study 2. Generalizability theory was applied to measure the reliability of the form used by the Department of Psychology to generalize over teachers and courses. It was hypothesized that the form will yield different reliability across these populations. D studies were conducted to determine the number of students and courses necessary to obtain adequate reliability for generalizing over teachers. D studies were conducted to determine the number of students and teachers needed to obtain adequate reliability to generalize over courses. It was hypothesized that the majority of variability will be attributed to the teacher effect and not the course.

Study 3. The purpose was to determine if expected grades are a possible contaminant in student ratings of teaching effectiveness. It was hypothesized that expected grades would contribute a significant portion of variance to the overall instructor rating but not course ratings.

Data Gathering for All Studies. Eastern Michigan University has been collecting student ratings data since the early 1970s. No additional data gathering was performed. Data files were generated by random sampling and will be discussed more thoroughly in the appropriate sections.

Study 1: Classical Test Theory and Reliability

Methodology

Instrument. The student ratings of teaching effectiveness forms at Eastern Michigan University is a cafeteria-style system. Each form contains two common items, overall rating of teacher effectiveness and overall course rating. These items are evaluated with a 5-point Likert scale: Much Above Average (A), Above Average (B), Average (C), Below Average (D) and Much Below Average (E), coded as 4, 3, 2, 1, and 0 respectively. The remainder of the form varies across departments and within some departments, depending on the course. The additional common departmental items on each form are chosen by the department head or a faculty committee. Each additional item on the form is rated on a 5-point Likert scale: Strongly Agree (SA), Agree (A), Undecided (U), Disagree (D) and Strongly Disagree (SD), coded as 4, 3, 2, 1, and 0 respectively. See Appendices A, B, and C for a sample form, a list of questions available for use, and a list of questions used by each department included in the following studies.

Sample. The data came from the 25 departments that used a common set of questions during the Fall 2001 and Winter 2002 semesters. Only complete response sets on the analyzed variables were included; missing data were removed by utilizing listwise deletion.

Design. An analysis was performed on each form currently in use for all departments that use a common set of questions at the University. Reliability analysis, following classical test theory, determined the reliability of each form. The two overall questions were not included in the analyses. One analysis used individual student responses to items as the unit of analysis, and a second analysis used the class mean as the unit of analysis. These analyses evaluated the ratings across students and across classes. Analyses were performed on two different semesters to assess stability.

Results

Cronbach's alpha was calculated for the 25 departments in which the entire department used a common set of questions, for both individual student responses and class means, for Fall 2001 and Winter 2002 academic terms. Overall, internal consistency was high for all forms, ranging from .81 to .99, across both individual student responses and class means. Table 2 presents the results of the reliability analysis.

Review of Table 2 indicates that for each department Cronbach's alpha is not only high but also very stable when comparing the coefficients across terms for both individual student responses and class means. In all cases, the class mean results exceeded the results based on individual responses. These results also held true for the Human,

Table 2

Cronbach's Alpha for each Department's Evaluation Form

	F01	F01	W02	W02
	Student	Class Mean	Student	Class Mean
Department	Response		Response	
Art (10 items)	.89 $N_S=1283$	.92 $N_C=84$	.90 $N_S=1230$	.93 $N_C=78$
Biology (10 items)	.91 $N_S=567$	.96 $N_C=35$	.91 $N_S=593$	.96 $N_c=38$
Chemistry (6 items)	.87 $N_S=1973$	.93 $N_C=109$	.88 $N_S=1477$	.94 $N_C=86$
Economics (8 items)	.89 $N_S=728$	.93 $N_C=37$	.89 $N_S=696$	.94 $N_C=35$
English Language & Literature (6 items)	.86 $N_S=3281$	.92 $N_C=240$	.87 $N_S=4785$	.92 $N_C=249$
Foreign Languages (9 items)	.93 $N_S=850$	.96 $N_C=79$	.93 $N_S=809$	.95 $N_C=80$
History & Philosophy (6 items)	.88 $N_S=2982$	.94 $N_C=118$	.88 $N_S=2438$	.94 $N_C=95$

Computer Science (6 items)	.86 N_S=1214	.91 N_C=77	.85 N_S=872	.90 N_C=63
Mathematics (9 items)	.81 N_S=2289	.91 N_C=134	.81 N_S=1959	.89 N_C=113
Physics & Astronomy (8 items)	.90 N_S=945	.95 N_C=50	.90 N_S=884	.92 N_C=66
Political Science (15 items)	.95 N_S=1719	.98 N_C=74	.94 N_S=1481	.97 N_C=65
Psychology (13 items)	.91 N_S=2009	.93 N_C=83	.91 N_S=1700	.93 N_C=69
Sociology, Anthropology & Criminology (6 items)	.88 N_S=1719	.94 N_C=74	.87 N_S=1536	.93 N_C=67
Communications & Theatre Arts (6 items)	.87 N_S=3863	.95 N_C=175	.86 N_S=3165	.93 N_C=149
Accounting & Finance (10 items)	.93 N_S=1086	.97 N_C=57	.94 N_S=1607	.98 N_C=80

Marketing (10 items)	.94 N_S=1093	.98 N_C=48	.94 N_S=1049	.98 N_C=50
Teacher Education (9 items)	.93 N_S=2496	.97 N_C=145	.93 N_S=2875	.97 N_C=172
Leadership & Counseling (10 items)	.94 N_S=422	.96 N_C=32	.91 N_S=403	.95 N_C=33
Special Education (9 items)	.93 N_S=1030	.97 N_C=65	.94 N_S=1222	.97 N_C=72
Human, Environmental & Consumer Resources (3 items)	.90 N_S=713	.95 N_C=56	.85 N_S=677	.91 N_C=51
Associated Health (6 items)	.93 N_S=472	.99 N_C=31	.93 N_S=516	.98 N_C=35
Nursing (5 items)	.91 N_S=612	.96 N_C=48	.89 N_S=544	.94 N_C=44

Social Work (14 items)	.95 $N_S=689$	.97 $N_C=44$	.93 $N_S=386$	.94 $N_C=32$
Business & Technology Education (10 items)	.95 $N_S=572$	.98 $N_C=44$	.95 $N_S=489$	.98 $N_C=37$
Industrial Technology (6 items)	.90 $N_S=939$	.95 $N_C=74$	.91 $N_S=684$	.94 $N_C=61$

Note. See Appendix C for list of items used by each department.

Environmental and Consumer Resources Department, which utilizes only three items.

Discussion

Cronbach’s alpha was computed to assess internal consistency of 25 departments that use sets of items on department evaluation forms. All the departmental forms show high levels of reliability on both the individual student responses and class means during both academic terms. In all cases, the alphas for the individual responses are lower than those computed based on class means. However, inspection of the differences indicates they are very small. Inspection of the alpha across terms shows that they remain stable for both individual student responses and class means.

The number of items used by each department varied, with the Department of Human, Environmental and Consumer Resources using the fewest, three items. The most items, 15, were used by the Department of Political Science. Though the difference between these extremes is relatively large, the alphas computed for the two departments are not very different.

Clearly, all forms included in this study have very high internal consistency. It would be difficult to say from the results that any one form is more reliable than any other form. Future research to measure the reliability of the evaluation forms would want to test the hypothesis that any set of randomly selected items would result in sufficient reliability.

The alphas for individual responses are based on sample sizes up into the thousands. In follow-up analysis on three departments, 20 response sets were randomly sampled. Those results were mixed with two department’s alphas being reduced marginally, while the other department’s increased marginally. It seems that the sample size may not greatly affect the results, though more research would be needed to confirm such a conclusion.

This study raises questions about measurement. Even with high internal consistency, the results do not tell us directly what the items are actually measuring. Do these items measure the effectiveness of the instructor, of the course, an interaction of the two, or even possibly an unrelated concept? For this reason, many researchers of student evaluations use generalizability theory to address such questions.

Study 2: Generalizability Theory and Reliability

Methodology

Sample. To examine the teacher effect, pairs of courses were randomly sampled for 14 psychology instructors from the Fall 2001 and Winter 2002 academic terms. To ensure a balanced design, each class selected had at least 15 students. From each class a subset of 14 students was randomly sampled from those who completed all items on the form. To examine the course effect, pairs of instructors were randomly sampled for eight different courses from the Fall 2001 and Winter 2002 academic terms. Only eight courses satisfied the sampling criteria for inclusion, and all eight courses were utilized. Again, to ensure a balanced design, each instructor selected had at least 15 students who completed all items on the form, from which 14 students were randomly sampled.

Only instructors or courses that were rated at least two different times were utilized. For those teachers or courses that were rated more than two times with at least 15 students, two of the courses or instructors were randomly selected. If the instructor was rated in more than one section of a course, only one section was used, and it was randomly selected. If the course was taught in different sections by the same instructor, only one section for that instructor was used, and it was randomly selected.

Design. Each student responded to the same set of items, but a different set of students rated each teacher and each course. The design is s:c:t x i (students nested within course nested within teacher crossed with items). Figure 1 illustrates the main effects and interactions of the design.

The linear model for the design is

$$X_{sict} = \mu + \nu_{t} + \nu_{c:t} + \nu_{s:c:t} + \nu_{i} + \nu_{ti} + \nu_{ci:t} + \nu_{si:c:t}\,. \qquad (3)$$

The observed score variance, as described by Gillmore et al. (1978), is:

$$\sigma^2(X_t) = \sigma^2(t) + \frac{\sigma^2(c:t)}{n'_c} + \frac{\sigma^2(s:c:t)}{n'_c n'_s} + \frac{\sigma^2(ti)}{n'_i} + \frac{\sigma^2(ci:t)}{n'_c n'_i} + \frac{\sigma^2(si:c:t)}{n'_c n'_s n'_i}\,. (4)$$

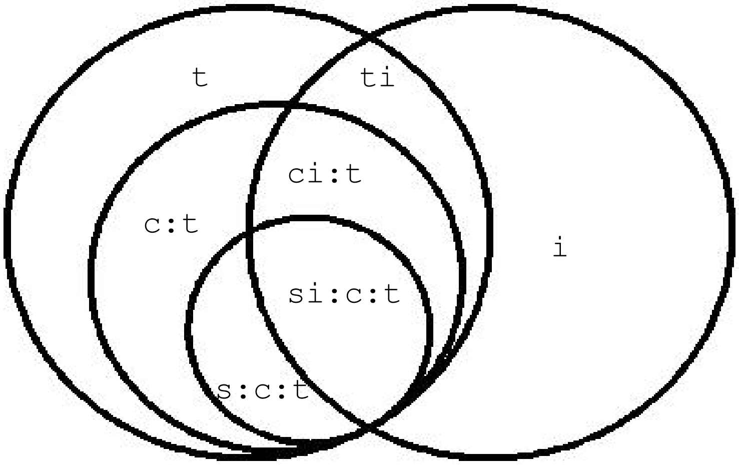


Figure 1. Venn diagram of generalizability study s:c:t x i

Following Gillmore et al. (1978), the variance due to the item main effect, $\sigma^2(i)$ is not included in Equation 4 since in student ratings the students all respond to the same items and therefore the item main effect would be a constant.

Table 3 shows the expected mean squares and estimation of variance components for the G study s:c:t x i; this design will determine teacher effect.

To determine course effect (c) and (t) will be interchanged in Table 3 and Equation 4. The generalizability coefficients that were used, as described by Gillmore et al. (1978), are

$$\rho^2(C,S,I) = \frac{\sigma^2(t)}{\sigma^2(X_t)}; \text{ and} \qquad (5)$$

$$\rho^2(T,S,I) = \frac{\sigma^2(c)}{\sigma^2(X_c)}. \qquad (6)$$

Equation 5 will be used to generalize over all courses, students, and items to determine teacher effect. Equation 6 will be used to generalize over all teachers, students, and items to determine the course effect.

$$\rho^2(C^*,S,I) = \frac{\sigma^2(t) + \frac{\sigma^2(c\ :\ t)}{n'_c}}{\sigma^2(X_t)}; \text{ and} \qquad (7)$$

$$\rho^2(T^*,S,I) = \frac{\sigma^2(c) + \frac{\sigma^2(t\ :\ c)}{n'_t}}{\sigma^2(X_t)}. \qquad (8)$$

The asterisks in Equations 7 and 8 indicate which facets will not be generalized over (Gillmore et al., 1978). Both equations are equal to $\rho^2(S, I)$ as described by Gillmore et al. (1976). These equations will be used to generalize over both students and items (Gillmore et al.). The generalizability coefficient $\rho^2(S, I)$ is favored for assessing the dependability of ratings for general instruction (Gillmore et al.).

Table 3

Summary of Random Effects ANOVA for Design s:c:t x i

Source (α)	*df*	E(*MS*)	$\sigma^2(\alpha)$
t	$n_t - 1$	$\sigma^2(e) +$ $n_i\sigma^2(s{:}c{:}t) +$ $n_s\sigma^2(ci{:}t)+$ $n_cn_s\sigma^2(c{:}t) +$ $n_cn_sn_i\sigma^2(t)$	$[MS(t)-MS(c{:}t)-$ $MS(ti) +$ $MS(ci{:}t)]/n_cn_sn_i$
i	$n_i - 1$	$\sigma^2(e) +$ $n_s\sigma^2(ci{:}t) +$ $n_cn_s\sigma^2(ti) +$ $n_tn_cn_s\sigma^2(i)$	$[MS(i)-MS(ti)]/n_tn_cn_s$
c:t	$n_t(n_c - 1)$	$\sigma^2(e) +$ $n_i\sigma^2(s{:}c{:}t) +$ $n_s\sigma^2(ci{:}t) +$ $n_cn_s\sigma^2(c{:}t)$	$[MS(c{:}t)-MS(ci{:}t)-$ $MS(s{:}c{:}t)+MS(e)]/n_in_s$
ti	$(n_t - 1)(n_i - 1)$	$\sigma^2(e) +$ $n_s\sigma^2(ci{:}t)+$ $n_cn_s\sigma^2(ti)$	$[MS(ti)-MS(ci{:}t)]/$ n_cn_s

s:c:t	$n_t n_c(n_s - 1)$	$\sigma^2(e) +$ $n_i\sigma^2(s{:}c{:}t)$	$[MS(s{:}c{:}t)-MS(e)]/n_i$
ci:t	$n_t(n_c - 1)$ $*(n_i - 1)$	$\sigma^2(e) +$ $n_s\sigma^2(ci{:}t)$	$[MS(ci{:}t)-MS(e)]/n_s$
si:c:t (e)	$n_t\ n_c(n_s - 1)$ $*(n_i - 1)$	$\sigma^2(e)$	$MS(e)$

Note. Portions of this table were adapted from Gillmore et al. (1978).

Results

Results for the G study analysis of variance involving students nested within courses nested within teachers are presented in Table 4.

As noted earlier, generalizability theory places emphasis on the magnitude of estimated variance components. For students nested within courses nested within teachers (s:c:t), the estimated variance component = 0.33. The student by item interaction (si:c:t) estimated variance components = 0.40. The teacher effect (t) estimated variance component = 0.09. The courses within teachers effect (c:t) estimated variance component = 0.04. The teacher by item interaction (ti) estimated variance component = 0.03. The course by item interaction (ci:t) estimated variance component = 0.03. These results, except for the courses within teachers effect (c:t), are consistent with the results of Gillmore et al. (1978).

Table 5 presents the D study generalizability coefficients for s:c:t x i. Equations 5 and 7 were used to estimate generalizability coefficients for several different courses and students within courses. Coefficients are presented for one, two, five, and 10 courses and for three levels of students: five students to represent a small section, 20 students to represent the mode of the data set, and 32 students to represent the mean of the data set for students in a course. The generalizability coefficients are not greatly influenced by more than 20 students, nor are they greatly influenced by the number of items.

The analysis shows that as a general measure of teaching effectiveness, $\rho^2(C, S, I)$, reliable results can be achieved using two items and 10 courses, regardless of the number of students in the section. Reliable results were obtained when 32 or more students in 10 or more sections completed responses for one item. Reliable results were obtained when 20 or more students in two or more sections completed responses for five items. Reliable results were also obtained when five or more students in five or more sections completed responses to five items. Using more than five items seems to make little difference. If one does not want to generalize over courses, $\rho^2(C^*, S, I)$, adequate reliability can be achieved using one item unless the class is small; then two items would be need.

Table 4

G Study Analysis of Variance Summary Table for s:c:t x i

Source (α)	*SS*	*df*	*MS*	Estimated Variance Component
t	706.42	13	54.34	0.09
c:t	204.34	14	14.59	0.04
s:c:t	1939.11	364	5.33	0.33
i	120.21	14	8.59	0.02
ti	280.31	182	1.54	0.03
ci:t	158.23	196	0.81	0.03
si:c:t (e)	2036.31	5096	0.40	0.40

Table 5

Estimated Generalizability Coefficients for Various Conditions for s:c:t x i

		$\rho^2(C, S, I)$			$\rho^2(C^*, S, I)$		
	n'_c	$n'_s = 5$	$n'_s = 20$	$n'_s = 32$	$n'_s = 5$	$n'_s = 20$	$n'_s = 32$
$n'_i = 1$	1	.27	.40	.42	.40	.57	.61
	2	.39	.52	.54	.48	.63	.66
	5	.55	.64	.65	.60	.69	.72
	10	.64	.69	.70	.66	.72	.73
$n'_i = 2$	1	.34	.48	.51	.49	.70	.74
	2	.49	.62	.67	.59	.75	.78
	5	.67	.74	.75	.71	.81	.82
	10	.74	.80	.80	.78	.83	.84
$n'_i = 5$	1	.40	.55	.58	.58	.80	.84
	2	.56	.70	.72	.69	.85	.88
	5	.74	.82	.84	.81	.90	.91
	10	.83	.88	.88	.86	.92	.92
$n'_i = 10$	1	.42	.58	.61	.62	.84	.88
	2	.59	.73	.75	.73	.89	.91
	5	.77	.85	.87	.84	.93	.94
	10	.86	.91	.91	.90	.95	.95

$n_i' = 15$	1	.43	.59	.62	.63	.85	.90
	2	.61	.74	.76	.74	.90	.93
	5	.78	.87	.87	.85	.94	.95
	10	.87	.92	.92	.91	.97	.97

Results for the G study analysis of variance involving students nested within teachers nested within courses are presented in Table 6.

The students nested within teachers nested within courses(s:t:c) estimated variance component = 0.28. The student by item interaction (si:t:c) estimated variance component = 0.38. The teachers nested within courses effect (t:c) estimated variance component = 0.09. The teacher by item interaction (ti:c) estimated variance component = 0.03. The item effect (i) estimated variance component = 0.02. The course effect (c) estimated variance component = 0.01. The course by item interaction (ci) estimated variance component = 0.01. These results are consistent with the results of the Gillmore et al.(1978) study.

Table 7 presents the D study generalizability coefficients for s:t:c x i. Equations 6 and 8 were used to estimate generalizability coefficients for several different instructors and students. Coefficients are presented for one, two, five, and 10 instructors and for three levels of students: five students to represent a small section, 20 students to represent the mode of the data set, and 32 students to represent the mean of the data set for students within a section of a course.

The analyses shows that as a general measure of course effectiveness, reliable results cannot be achieved regardless of the number of students, instructors, or items used. These findings are consistent with those of Gillmore et al. (1978).

If one does not want to generalize over instructors, $\rho^2(T^*,S,I)$, reliable results can be achieved using two items unless the section was small; then 10 or more items and 10 or more sections are necessary. It is noted that when using one or two items the generalizability coefficient actually decreases for 20 or more students, as more sections are added. When data for five or more items and for 20 or more students were analyzed, the generalizability coefficients did not increase much as sections were added. This trend is not seen in the results of the s:c:t x i D studies. According to Gillmore et al. (1978) the $\rho^2(C^*,S,I)$ and $\rho^2(T^*,S,I)$ should be equivalent estimates. This does not hold true for most of the results in this study, particularly for the results when utilizing 20 or more students.

Table 6

G Study Analysis of Variance Summary Table for s:t:c x i

				Estimated Variance
Source (α)	*SS*	*df*	*MS*	Component
c	204.86	7	29.26	0.01
t:c	185.04	8	23.13	0.09
s:t:c	965.84	208	4.64	0.28
i	66.60	14	4.76	0.02
ci	119.01	98	1.21	0.01
ti:c	89.81	112	0.80	0.03
si:t:c (e)	1098.45	2912	0.38	0.38

Table 7

Estimated Generalizability Coefficients for Various Conditions for s:t:c x i

		$\rho^2(T, S, I)$			$\rho^2(T^*, S, I)$		
	n'_t	$n'_s = 5$	$n'_s = 20$	$n'_s = 32$	$n'_s = 5$	$n'_s = 20$	$n'_s = 32$
$n'_i = 1$	1	.04	.06	.06	.37	.58	.62
	2	.07	.10	.11	.38	.57	.61
	5	.14	.20	.21	.40	.55	.58
	10	.22	.28	.29	.42	.54	.56
$n'_i = 2$	1	.05	.07	.07	.47	.70	.74
	2	.09	.13	.13	.48	.69	.73
	5	.18	.25	.26	.51	.69	.72
	10	.29	.36	.36	.54	.68	.70
$n'_i = 5$	1	.06	.08	.08	.56	.79	.84
	2	.10	.15	.15	.58	.80	.84
	5	.22	.29	.30	.62	.81	.84
	10	.35	.43	.44	.66	.81	.84
$n'_i = 10$	1	.06	.08	.09	.60	.83	.88
	2	.11	.15	.16	.62	.84	.88
	5	.24	.31	.32	.66	.85	.89
	10	.38	.46	.47	.71	.87	.89

$n_i' = 15$	1	.06	.08	.09	.61	.85	.89
	2	.11	.16	.16	.63	.86	.90
	5	.24	.31	.32	.68	.87	.90
	10	.39	.47	.48	.73	.89	.91

Discussion

In important decisions, especially regarding faculty promotion, raises, and tenure, the individuals using the data to make the decisions must know that the data is a reliable measure of the instructor's effectiveness, a question that is not adequately addressed by classical test theory. Previous research suggests that generalizability theory is a preferred method over classical test theory to assess reliability of evaluations. One of the major benefits of generalizability theory is the partitioning of variance to determine what the measure is actually measuring. In the present study, generalizability theory was applied to the Department of Psychology student evaluation forms to determine if they were measuring teaching effectiveness, course effectiveness, both, or something else.

Gillmore et al. (1978), in assessing student ratings of teaching effectiveness, proposed that it would be appropriate to generalize over all courses instructors might teach. They suggested $\rho^2(C, S, I)$ as the most appropriate index of dependability (Gillmore et al., 1978). Consistent with Gillmore, our results suggest data should be collected from as many courses as possible to assess the reliability of student ratings of teacher effectiveness. Basing a decision on less than five course using two items would probably be questionable. If the section has a low enrollment, more courses would be required to achieve adequate dependability. A total of five items appears to provide an increase in reliability, but beyond that little is gained by adding items.

Gillmore et al. (1978) also suggested that if an instructor were to teach only multiple sections of one specific course the generalizability coefficient of $\rho^2(C^*, S, I)$ would be an appropriate index of reliability. Our results suggest that in such cases, using two items and one course would achieve adequate dependability unless based on a small section.

Results regarding the dependability of courses by generalizing over teachers, $\rho^2(T, S, I)$, are consistent with findings from the study by Gillmore et al. The current study and the Gillmore et al. (1978) found that the variance component for the course main effect to be low. Reliable results cannot be obtained regardless of the number of items utilized, instructors, or number of students in each course.

If the same course were taught in multiple sections by different instructors, the generalizability coefficient of $\rho^2(T^*, S, I)$ would be appropriate. It was noted in the results that they were not consistent with the results of the s:c:t x i study. As more sections were added, particularly for larger sections, the coefficients got smaller for one and two items. For five or more items, the coefficients were nearly equal regardless of the number of sections added, for larger sections. The results for the coefficient $\rho^2(T*, S, I)$ were not consistent with Gillmore et al. (1978). This is possibly due to Gillmore reporting results for only one level of items. However, for the same number of items, the current results are not consistent with Gillmore et al. (1978), nor are they comparable to the results of the s:c:t x i study. The results of Gillmore et al. (1978) differed from the results of the current study because they used multiple departments, whereas the current study used only one department.

When considering courses the variance of the teacher nested with courses was nine times as large as the variance of the course effect. Comparing the two studies, the teacher effect is also nine times as large as the course effect. From the s:c:t x i we see that 69% of the estimated class variance component, $\sigma^2(t) + \sigma^2(c : t)$, is attributable to teacher effect. Similarly,from the s:t:c x i we see that 10% of the estimated class variance component, $\sigma^2(c) + \sigma^2(t : c)$, is attributable to the course effect. This suggests that the rating of the course is a function of the rating of the instructor. The small variance in course effect suggests that there is little reason to indicate that some courses are rated less favorably than others are.

A limitation of the study is that it focused only on a single department and therefore cannot be generalized to other departments. Considering courses, it was noted earlier that the small variance in the course effect suggests that the courses were not rated less favorably than others were. Previous research suggests that, in particular, courses in mathematics tend to be rated lower than courses in social sciences. Had this study included other departments, the results might suggest a similar difference. The sample size of the s:t:c x i G study also may have affected the results. It is nearly half the size of the other G study, which may make comparing the results of the two studies difficult.

The findings of the G studies were not completely consistent with those from previous research, particularly the Gillmore et al. (1978) study. The inconsistency is not surprising since the student evaluation forms used are different. But, overall, the findings were similar to those of previous research. As expected, the evaluation form used by the Department of Psychology tends to measure the students' rating of teaching effectiveness better than that of the students' ratings of courses.

Future research may want to attempt to determine what role repeated assessments of the same instructor by the same student may have. Is there a need to be concerned that the same group of students are rating the same instructor? This could easily happen in graduate courses as well as upper level undergraduate courses, which made up about 50% of the courses included in the s:c:t x i sample.

Study 3: Potential Bias by Absolute Expected Grade

Methodology

Sample. Data for the third study was based on student ratings obtained by the Department of Psychology at Eastern Michigan University for the Fall 2001 and Winter 2002 academic terms. The evaluation form used by the Department of Psychology contains 15 items, the first two of which are university-wide items of overall teacher effectiveness rating and overall course rating, as well as a question regarding absolute expected grade received.

Evaluations that did not contain complete response sets were eliminated. Given the size of the sample, several students may have rated the same teacher in different courses on multiple occasions, but anonymous ratings make these impossible to tease out. Greenwald and Gillmore (1997a) suggest that the effects of repeated measure would be negligible; however, they provided no data to support this assumption. There were 33 different instructors for the Fall 2001 term and 29 for the Winter 2002 term. There were 39 different courses taught in the Fall 2001 term and 37 in the Winter 2002 term.

Design. Analyses examined the mean responses for each instructor and each course within the psychology department, for each academic semester based on the student ratings of the instructor and the overall course ratings as completed by students. The mean overall teaching effectiveness rating collapsed over courses and students was computed for each instructor. The mean course rating collapsed over instructor and students was also computed. Mean data on the additional variables were obtained for the instructor regardless of the course taught and for the course regardless of which instructor taught it. Zero-order correlation, semi-partial correlation, and stepwise multiple regression were used to determine which variables made the largest contribution to the overall student ratings of teaching effectiveness and overall course ratings. The expected grade variable was entered first, and then the remaining variables were entered in a stepwise manner.

Results

The Department of Psychology includes thirteen items on its evaluation form, aside from the overall teaching and course ratings. The independent variables are (a) my instructor has an effective style of presentation [style], (b) my instructor seems well-prepared for class [prep], (c) my instructor stimulates interest in the course [inter], (d) my instructor displays enthusiasm when teaching [enthu], (e) my instructor is actively helpful when students have problems [help], (f) I understand what is expected of me in this course

[expt], (g) exams are fair [exam], (h) grades are assigned fairly and impartially [grd_1], (i) I would recommend this course to another student [rec_c], (j) I would recommend this instructor to another student [rec_i], (k) I learned a lot in this course [learn], (l) I looked forward to taking this course before it began [fwd], and (m) the grade I expect to receive in this course is (A, B, C, D) [grd_2].

For each regression analysis, expected grade by itself did not significantly predict the overall student ratings of teaching effectiveness. The bivariate correlations between expected grade and the overall rating was .216 for the Fall 2001 term and -.145 for the Winter 2002 term. Neither of these correlations is significant. However, expected grade did become a significant predictor once other variables entered the regression model for the Winter 2002 term. This occurred despite the fact that the part correlations show expected grade accounting for less than 1% of the variance in overall effectiveness ratings. Non-significant relationships leading to statistically significant prediction can largely be attributed to the miniscule error that remains in prediction once other variables were permitted to enter the model and suggest that any results associated with expected grade should be regarded with suspicion.

Table 8 shows summaries of the final regression models for the overall student ratings of teaching effectiveness for the individual academic terms. In addition, Table 9 shows summaries of regression coefficients and bivariate and semi-partial correlation coefficients.

Table 8

Model Summaries of Stepwise Regression Analysis for Predicting Overall Student Ratings of Teaching Effectiveness for the Fall 2001 (N=33) and Winter 2002 (N=29) Academic Terms

Step	R	R^2	R^2_{adj}	ΔR^2	F_{chg}	p	df_1	df_2
Fall 2001 Academic Term								
grd_2	.216	.047	.016	.047	1.514	ns	1	31
style	.949	.900	.893	.853	256.075	<.001	1	30
exam	.958	.918	.909	.018	6.183	<.02	1	29
learn	.965	.931	.922	.014	5.597	<.05	1	28
Winter 2002 Academic Term								
grd_2	.145	.021	-.015	.021	0.583	ns	1	27
style	.983	.966	.953	.945	716.257	<.001	1	26
expt	.991	.981	.979	.016	21.244	<.001	1	25
rec_i	.992	.985	.982	.004	5.599	<.05	1	24

Table 9

Summary of Regression, Bivariate, and Semi-partial Correlation Coefficients for Predicting Overall Student Ratings of Teaching Effectiveness for the Fall 2001 (N=33) and Winter (N=29) 2002 Academic Terms

	B	*SE B*	*β*	*t*	Bivariate *r*	Part *r*
Fall 2001 Academic Term						
grd_2	.079	.071	.063	1.124	.216	.056
style	.652	.083	.701	7.884***	.936	.390
exam	.177	.064	.163	2.760*	.533	.137
learn	.253	.107	.206	2.366*	.844	.117
Winter 2002 Academic Term						
grd_2	.157	.052	.091	3.003*	-.145	.075
style	.655	.103	.647	6.374***	.980	.160
expt	.192	.059	.141	3.235**	.758	.081
rec_i	.225	.095	.263	2.366*	.979	.059

$^{*}p<.05$, ** $p<.005$, $^{***}p<.001$

As with overall student ratings of teaching effectiveness, expected grade by itself did not significantly predict the overall course rating for any analysis. The bivariate correlations between expected grade and the overall rating was -.011 for the Fall 2001 term and -.088 for the Winter 2002 term. Neither of these correlations is significant. However, expected grade did become a significant predictor once other variables entered the regression model for the Winter 2002 term, just as it did for the overall student ratings of teaching effectiveness. This occurred despite the fact that the part correlations show expected grade accounting for less than 1% of the variance in overall course ratings. Again as with the overall student ratings of teaching effectiveness ratings, non-significant relationships leading to significant statistically significant prediction can largely be attributed to the miniscule error that remains in prediction once other variables were permitted to enter the model and suggest that any results associated with expected grade should be regarded with suspicion.

Table 10 shows summaries of the final regression models for the overall course ratings for the individual academic terms. In addition, Table 11 shows summaries of regression coefficients and bivariate and semi-partial correlation coefficients.

Table 10

Model Summaries of Stepwise Regression Analysis for Predicting Overall Course Ratings for the Fall 2001 (N=39), Winter 2002 Academic Terms (N=37) Academic Terms

Step	R	R^2	R^2_{adj}	ΔR^2	F_{chg}	p	df_1	df_2
Fall 2001 Academic Term								
grd_2	.011	.000	.000	.000	.004	ns	1	37
style	.903	.815	.805	.815	158.934	<.001	1	36
learn	.929	.862	.851	.047	11.946	<.001	1	35
rec_c	.937	.878	.863	.015	4.262	<.05	1	34
Winter 2002 Academic Term								
grd_2	.088	.008	.000	.008	.270	ns	1	35
rec_c	.940	.883	.876	.876	255.182	<.001	1	34
learn	.964	.929	.922	.045	20.912	<.001	1	33

Table 11

Summary of Regression, Bivariate, and Semi-partial Correlation Coefficients for Predicting Overall Course Rating for the Fall 2001 (N=39), Winter 2002 (N=37) Academic Terms

	B	*SE B*	*β*	*t*	Bivariate *r*	Part *r*
			Fall 2001 Academic Term			
grd_2	-.123	.098	-.084	-1.250	-.011	-.075
style	.290	.099	.397	2.922**	.903	.175
learn	.295	.136	.319	2.117*	.874	.131
rec_c	.244	.118	.272	2.064*	.879	.124
			Winter 2002 Academic Term			
grd_2	-.126	.060	-.100	-2.108*	-.088	-.098
rec_c	.515	.079	.585	6.499***	.930	.302
learn	.354	.077	.411	4.573***	.914	.213

*p<.05, **p<.01, ***p<.001

Discussion

Study 3 attempted to determine which variables used on the Department of Psychology student evaluation form would predict the overall student ratings of teaching effectiveness and overall course ratings. The study provided several regression analyses based on instructor means and course means, one for the Fall 2001 academic term and one for the Winter 2002 academic term. Chang (2000) performed a similar study but used data from only a single academic term. The current study used two separate terms and found that for both overall variables, the final regression models were not similar.

In particular, this study hypothesized that the absolute expected grade variable would account for a significant amount of the variance when predicting the overall student ratings of teaching effectiveness but would not when predicting the overall course rating. When utilizing only the expected grade variable it did not significantly predict the overall student ratings of teaching effectiveness variable; therefore, the hypothesis would have to be rejected. As hypothesized, the expected grade did not significantly predict overall course rating.

An interesting result emerged for both sets of regression when analyzing the data from the Winter 2002 academic term.

Even though expected grade by itself did not significantly predict either dependent variable, when additional variables entered the regression models, expected grade became a significant predictor in the model. It is highly likely that these results are due to collinearity in the data, and create a spurious role of expected grade in this data. Appendix D reports the bivariate correlations on which the regression analyses were based. All but two variables had significant correlations with the overall instructor rating and course ratings. Many of the other variables have higher and significant correlations with the dependent variables. This could explain why in general the regression models were not the same throughout the analyses.

Even if absolute expected grade truly were a significant predictor in any of the models, the portion of variance accounted for was very small, less than 5% in all cases. It is probably safe to say that the effect of the absolute grade has little or no influence on either type of ratings.

The primary objective of this study was to determine if expected grade significantly contributed to the overall student ratings of teaching effectiveness rating. These findings are not consistent with previous research. However, previous research has shown that the relative expected grade may bias ratings more than an absolute expected grade. The

Department of Psychology evaluation form does not include a question regarding the relative expected grade. The form also does not include items regarding workload, which Greenwald and Gillmore (1997a, 1997b) suggest are an important indication of possible grading leniency. Had the present study found a significant effect from the expected grade, it would not necessarily have indicated grading leniency but possibly have indicated that the students worked hard to earn a higher grade or that the instructor actually is effective. The results of the study suggest that grading leniency is not of concern within the Department of Psychology.

Future research would want to address the issue of absolute versus relative expected grade. A question regarding the workload would be useful. As noted in the discussion of Study 2, future research may also want to look at other departments. It would be beneficial to determine the relationship of departmental grading standards to overall ratings. Research in the future should attempt to determine the stability of regression equations over semesters. Would it be better to develop a single regression equation for academic years, or does adjustment of the equations need to be performed each semester, if bias were found? If the regression equations do need adjustment each term, what does that tell the researcher about how students rate instructors and courses?

Conclusion

Results of Study 1 show that departments that used a common set of questions for their evaluation forms all achieve high levels of reliability. The difference between the departments does not suggest one set of questions was more reliable than another.

Results of Study 2 suggest that the evaluation form used by the Department of Psychology can be generalized across any psychology instructor, regardless of the psychology course taught, and result in reliable ratings. However, the same evaluation form, based on this study, cannot be used to achieve reliable ratings of any psychology course regardless of which instructor taught it. The results suggest that for reliable results of student ratings of teaching effectiveness, an evaluation would need five items and ratings from at least five courses.

Study 3 indicates that the expected grade has minimal influence on student ratings of teaching effectiveness and course ratings obtained in the Department of Psychology. However, more research is needed to confirm this conclusion.

Taking these results together, it would be difficult to say that any rating form is obviously better than another is.

The results of the studies on the Department of Psychology evaluation suggest that only five items are needed. All but one department used six or more items on their evaluation forms, aside from the overall items used by the entire university.

When important decisions about promotion, raises, and tenure are being made, the study results suggest that the use of one or two items may not be sufficient. It would likely be beneficial for the institution to utilize five items across the entire university. Revision or addition of certain items to assist in detecting bias could be added and used university-wide. If student evaluations of instructors are to be a fair assessment of teaching effectiveness, more research on the current evaluation system would be beneficial.

References

Aubrecht, J. D. (1979). *Are student ratings of teacher effectiveness valid?* (IDEA Paper No. 2), Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development (ERIC Document Reproduction Service No. ED 202 410)

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work.* San Francisco: Josey-Bass.

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice, Winter*, 27-34.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited.* (IDEA Paper No. 32), Manhattan KS: Kansas State University, Center for Faculty Evaluation and Development. (ERIC Document Reproduction Service No. ED 402 338)

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improve faculty performance and promotion/tenure decisions* (pp. 25-44). Bolton, MA: Anker Publishing Company, Inc.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Josey-Bass.

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

Chang, T. (2000, April). *An application of regression models with student ratings in determining course effectiveness.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 455 311)

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. (1986, April). *An updated and expanded meta-analysis of multisection student rating validity studies.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 270 471)

Crick, J. E., & Brennan, R. L. (1983, August). *Manual for GENOVA: A generalized analysis of variance system.* (ACT Technical Bulletin No. 43). Iowa City, IA: ACT.

Frances, S. J., & Gruber, M. B. (1981, August). *Student evaluations of psychology instructors.* Paper presented at the Annual Convention of the American Psychological Association, Los Angeles, CA. (ERIC Document Reproduction Service No. ED 210 572)

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82*, 341-351.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1976). *The generalizability of student ratings: General theory and application to the University of Washington instructional assessment system* (Educational Assessment Center Project: 506). Seattle, WA: University of Washington Educational Assessment Center. (ERIC Document Reproduction Service No. ED 146 192)

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings: Estimation of the teacher and course components. *Journal of Educational Measurement, 15*, 1-13.

Gilmore, D. C., Swerdlik, M. E., & Beehr, T. A. (1980). Effects of class size and college major on student ratings of psychology courses. *Teaching of Psychology,* 7, 210-214.

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217.

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743-751.

Haskell, R. W. (1997, February 12). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives, 5* Article 6. Retrieved October 18, 2002, from http://olam.ed.asu.edu/epaa/v5n6.html

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology, 63*, 130-133.

Huang, C., Go, S., Druva-Roush, C., & Moore, J. E. (1995, April). *A generalizability theory approach to examining teaching evaluation instruments completed by students.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 394 984)

Hudson, J. C. (1989). Expected grades correlate with evaluation of teaching. *Journalism Educator, Summer,* 38-44.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13*, 171-183.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754.

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

Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. Brown (Eds.) *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 527-551). San Diego: Harcourt Brace & Co.

McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology, 82*, 189-200.

Miller, R. I. (1974). *Developing programs for faculty evaluation.* San Francisco: Josey-Bass.

Neale, J. M., & Liebert, R. M. (1986). *Science and behavior: An introduction to methods of research* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ory, J. C. (2000). Teaching evaluations: Past, present, and future. *New Directions for Teaching and Learning, 83,* 13-18.

Remmers, H. (1930). To what extent do grades influence student ratings? *Journal of Educational Research, 21*, 314-316.

Remmers, H. (1934). Reliability and halo effect of high school and college students' judgments of their teachers. *Journal of Applied Psychology, 18*, 619-630.

Ronco, S. L. (1999, June). *Deconstructing the student assessment of instruction instrument: Some psychometric issues.* Paper presented at the Annual Forum of the Association for Institutional Research, Seattle, WA. (ERIC Document Reproduction Service No. ED 433 763)

Seldin, P. (1998). How colleges evaluate faculty. *AAHE Bulletin, 41* (7), 3-7.

Seldin. P. (1999). Current practices—good and bad—nationally. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 1-24). Bolton, MA: Anker Publishing Company, Inc.

Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25*, 269-291.

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71*, 293-302.

Sun, A., & Valiga, M. J. (1997, March). *Assessing reliability of student ratings of advisor: A comparison of univariate and multivariate generalizability approaches.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 411 262)

Sun, A., Valiga, M. J., Gao, X., & ACT (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. *The Journal of Experimental Education, 64*, 367-379.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. Brown (Eds.) *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 95-124). San Diego: Harcourt Brace & Co.

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

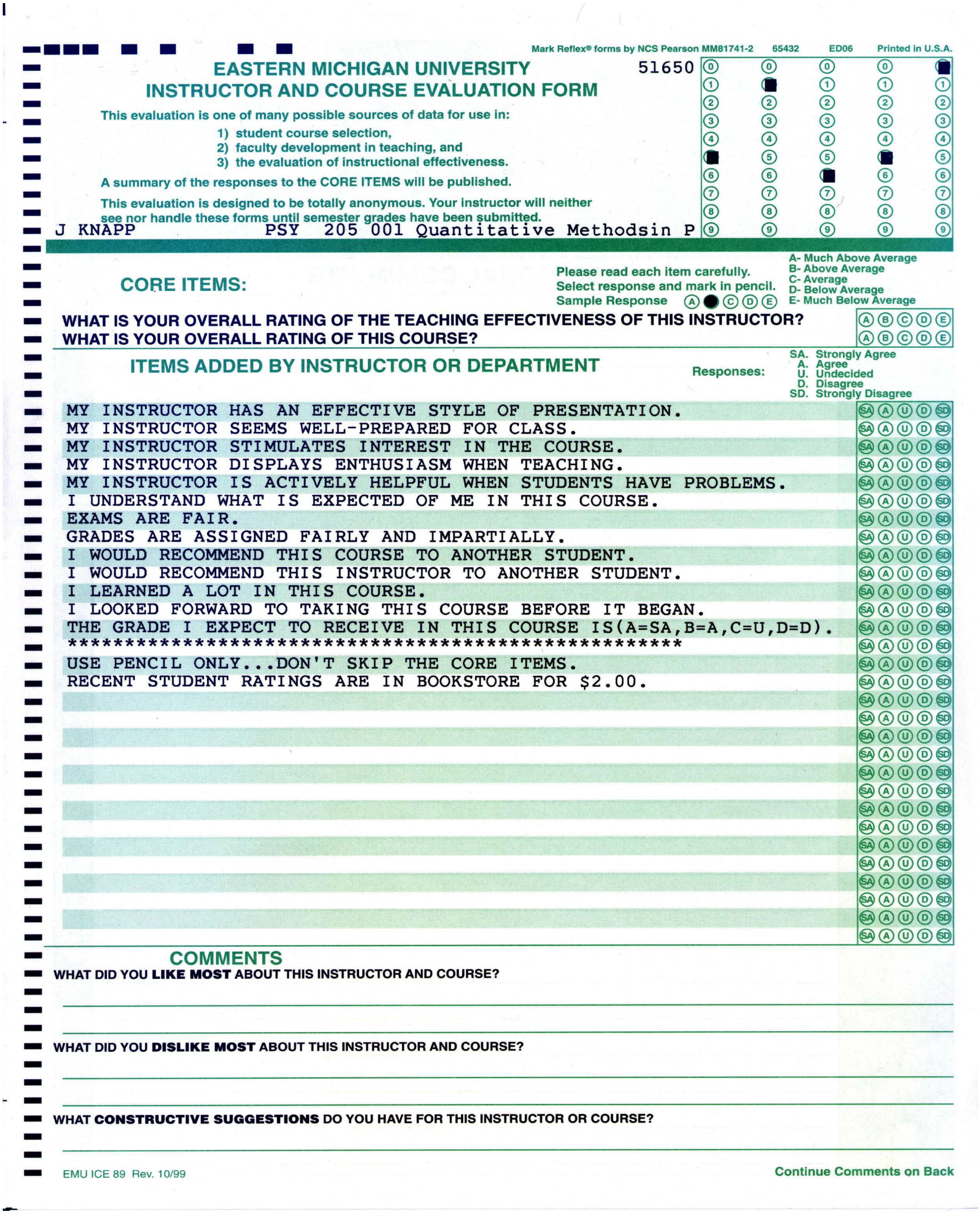
Williams, W. M., & Ceci, S. J. (1997). How'm I doing? *Change, 29* (5), 12-23.

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.

Appendices

APPENDIX A

Sample Eastern Michigan University Instructor and Course Evaluation Form



Mark Reflex® forms by NCS Pearson MM81741-2 65432 ED06 Printed in U.S.A.

EASTERN MICHIGAN UNIVERSITY
INSTRUCTOR AND COURSE EVALUATION FORM

51650

⓪	⓪	⓪	⓪	●
①	●	①	①	①
②	②	②	②	②
③	③	③	③	③
④	④	④	④	④
●	⑤	⑤	●	⑤
⑥	⑥	●	⑥	⑥
⑦	⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨	⑨

This evaluation is one of many possible sources of data for use in:

1) student course selection,
2) faculty development in teaching, and
3) the evaluation of instructional effectiveness.

A summary of the responses to the CORE ITEMS will be published.

This evaluation is designed to be totally anonymous. Your instructor will neither see nor handle these forms until semester grades have been submitted.

J KNAPP PSY 205 001 Quantitative Methodsin P

CORE ITEMS:

Please read each item carefully.
Select response and mark in pencil.
Sample Response Ⓐ ● Ⓒ Ⓓ Ⓔ

A- Much Above Average
B- Above Average
C- Average
D- Below Average
E- Much Below Average

WHAT IS YOUR OVERALL RATING OF THE TEACHING EFFECTIVENESS OF THIS INSTRUCTOR?	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ
WHAT IS YOUR OVERALL RATING OF THIS COURSE?	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ

ITEMS ADDED BY INSTRUCTOR OR DEPARTMENT

Responses:
SA. Strongly Agree
A. Agree
U. Undecided
D. Disagree
SD. Strongly Disagree

MY INSTRUCTOR HAS AN EFFECTIVE STYLE OF PRESENTATION.	(SA) (A) (U) (D) (SD)
MY INSTRUCTOR SEEMS WELL-PREPARED FOR CLASS.	(SA) (A) (U) (D) (SD)
MY INSTRUCTOR STIMULATES INTEREST IN THE COURSE.	(SA) (A) (U) (D) (SD)
MY INSTRUCTOR DISPLAYS ENTHUSIASM WHEN TEACHING.	(SA) (A) (U) (D) (SD)
MY INSTRUCTOR IS ACTIVELY HELPFUL WHEN STUDENTS HAVE PROBLEMS.	(SA) (A) (U) (D) (SD)
I UNDERSTAND WHAT IS EXPECTED OF ME IN THIS COURSE.	(SA) (A) (U) (D) (SD)
EXAMS ARE FAIR.	(SA) (A) (U) (D) (SD)
GRADES ARE ASSIGNED FAIRLY AND IMPARTIALLY.	(SA) (A) (U) (D) (SD)
I WOULD RECOMMEND THIS COURSE TO ANOTHER STUDENT.	(SA) (A) (U) (D) (SD)
I WOULD RECOMMEND THIS INSTRUCTOR TO ANOTHER STUDENT.	(SA) (A) (U) (D) (SD)
I LEARNED A LOT IN THIS COURSE.	(SA) (A) (U) (D) (SD)
I LOOKED FORWARD TO TAKING THIS COURSE BEFORE IT BEGAN.	(SA) (A) (U) (D) (SD)
THE GRADE I EXPECT TO RECEIVE IN THIS COURSE IS(A=SA,B=A,C=U,D=D).	(SA) (A) (U) (D) (SD)
**	(SA) (A) (U) (D) (SD)
USE PENCIL ONLY...DON'T SKIP THE CORE ITEMS.	(SA) (A) (U) (D) (SD)
RECENT STUDENT RATINGS ARE IN BOOKSTORE FOR $2.00.	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)
	(SA) (A) (U) (D) (SD)

COMMENTS

WHAT DID YOU **LIKE MOST** ABOUT THIS INSTRUCTOR AND COURSE?

WHAT DID YOU **DISLIKE MOST** ABOUT THIS INSTRUCTOR AND COURSE?

WHAT **CONSTRUCTIVE SUGGESTIONS** DO YOU HAVE FOR THIS INSTRUCTOR OR COURSE?

EMU ICE 89 Rev. 10/99

Continue Comments on Back

APPENDIX B

Items Available for use on Eastern Michigan University Evaluation Forms

University Wide Items

What is your overall rating of the teaching effectiveness of this instructor?

What is your overall rating of this course?

Additional Items

001 I understand easily what my instructor is saying.

002 My instructor displays a clear understanding of course topics

003 My instructor is able to simplify difficult materials.

004 My instructor explains experiments and/or assignments clearly.

005 Difficult topics are structured in easily understood ways.

006 My instructor has an effective style of presentation.

007 My instructor seems well prepared for class.

008 My instructor talks at a pace suitable for maximum comprehension.

009 My instructor speaks audibly and clearly.

010 My instructor draws and explains diagrams effectively.

011 My instructor writes legibly on the blackboard.

012 My instructor has no distracting peculiarities.

013 My instructor makes learning easy and interesting.

014 My instructor holds the attention of the class.

015 My instructor senses when students are bored.

016 My instructor stimulates interest in the course.

017 My instructor displays enthusiasm when teaching.

018 The course supplies me with an effective range of challenges.

019 In this course, many methods are used to involve me in learning.

020 My instructor makes me feel involved with this course.

021 In this course, I always felt challenged and motivated to learn.

022 My instructor motivates me to do further independent study.

023 This course motivates me to take additional related courses.

024 This course has been intellectually fulfilling to me.

025 My instructor has stimulated my thinking.

026 My instructor has provided many challenging new viewpoints.

027 My instructor teaches one to value the viewpoint of others.

028 This course caused me to reconsider many of my former attitudes.

029 In this course, I have learned to value new viewpoints.

030 This course fosters respect for new viewpoints.

031 This course stretched and broadened my views greatly.

032 This course has effectively challenged me to think.

033 The class meetings helped me to see other points of view.

034 The course develops the creative ability of students.

035 My instructor encourages student creativity.

036 My instructor emphasizes relationships between and among topics.

037 My instructor helps me apply theory to solve problems.

038 My instructor emphasizes conceptual understanding of material.

039 My instructor effectively blends facts with theory.

040 My instructor clarifies topics with developments in other fields.

041 My instructor makes good use of examples and illustrations.

042 Relationships among course topics are clearly explained.

043 This course builds understanding of concepts and principles.

044 My instructor is actively helpful when students have problems.

045 My instructor recognizes when some students fail to comprehend.

046 Everything possible is provided to help me learn.

047 My instructor explanations and comments are always helpful.

048 My instructor evaluates often and provides help where needed.

049 My instructor appears to grasp quickly what a student is saying.

050 My instructor is careful and precise when answering questions.

051 My instructor is readily available for consultation.

052 My instructor regularly checks and rewards progress in learning.

053 My instructor suggests specific ways I can improve.

054 My instructor recognizes and rewards success in this course.

055 My instructor can gauge what I know and what I should do next.

056 Exams are used to help me find my strengths and weaknesses.

057 My instructor returns papers quickly enough to benefit me.

058 This course shows sensitivity to individual interests/abilities.

059 My instructor adjusts to fit individual abilities and interests.

060 The flexibility of this course helps all kinds of students learn.

061 My instructor tailors this course to help many kinds of students.

062 The design of this course lets me learn at my own pace.

063 Students proceed at their own pace in this course.

064 I was able to keep up with the workload in this course.

065 My background is sufficient to enable me to use course material.

066 A teacher/student partnership in learning is encouraged.

067 Each student is encouraged to contribute to class learning.

068 I am free to express and explain my own views in class.

069 When I have a question or comment I know it will be respected.

070 I feel free to ask questions in class.

071 I feel that I am an important member of this class.

072 Mutual respect is a concept practiced in this course.

073 My instructor respects divergent viewpoints.

074 My instructor respects constructive criticism.

075 I feel free to challenge my instructor’s ideas in class.

076 My instructor relates to me as an individual.

077 My instructor deals fairly and impartially with me.

078 My instructor readily maintains rapport with this class.

079 This instructor encourages divergent thinking.

080 The climate of this class is conducive to learning.

081 This course has clearly stated objectives.

082 The objectives of this course were clearly explained to me.

083 The stated goals of this course are consistently pursued.

084 I understand what is expected of me in this course.

085 The course objectives allow me to know when I am making progress.

086 I was able to set and achieve some of my own goals.

087 I had an opportunity to help determine course objectives.

088 Lecture information is highly relevant to course objectives.

089 The course content is consistent with my prior expectations.

090 This course material is pertinent to my professional training.

091 This course contributes significantly to my professional growth.

092 I can apply information/skills learned in this course.

093 This course will be of practical benefit to me as a student.

094 My technical skills were improved as a result of this course.

095 This course directly contributes to my vocational preparation.

096 This course is a valid requirement for my major.

097 The relationship of this course to my education is apparent.

098 The practical application of subject matter is apparent.

099 This course gives me an excellent background for further study.

100 This course is up-to-date with developments in the field.

101 This course includes adequate information on career opportunity.

102 This course includes a sufficient number of practical exercises.

103 The content of this course is relevant to my educational goals.

104 The amount of material covered was reasonable.

105 My instructor develops classroom discussion skillfully.

106 There is sufficient time in class for questions and discussions.

107 My instructor allows student discussion to proceed uninterrupted.

108 My instructor encourages students to debate conflicting views.

109 My instructor does not monopolize classroom discussion.

110 One real strength of this course is the classroom discussion.

111 Challenging questions are raised for discussion.

112 This course provides an opportunity to learn from other students.

113 Exams accurately assess what I have learned in this course.

114 Exams are fair.

115 Exams are free from ambiguity.

116 Exams cover a reasonable amount of the material.

117 Exams stress important points of the lectures/text.

118 Exams in this course have instructional value.

119 Exams are creative and require original thought.

120 I know how I stand relative to others in the class on exams.

121 Exams are reasonable in length and difficulty.

122 Exams are coordinated with major course objectives.

123 My final grade will accurately reflect my overall performance.

124 Grades are an accurate assessment of my knowledge.

125 Grades are assigned fairly and impartially.

126 The grading system was clearly explained.

127 The contract grading method is used appropriately in this course.

128 My instructor has a realistic definition of good performance.

129 The assigned readings significantly contributed to this course.

130 The assigned reading is well integrated into this course.

131 Length and difficulty of assigned readings are reasonable.

132 Assigned readings are interesting and hold my attention.

133 Assignments are of definite instructional value.

134 Assignments are related to goals of this course.

135 Complexity and length of course assignments are reasonable.

136 Directions for course assignments are clear and specific.

137 The number of course assignments is reasonable.

138 Class projects are related to course goals and objectives.

139 The course’s programmed learning materials are effective.

140 The group work contributes significantly to this course.

141 Student presentations significantly contribute to this course.

142 Student presentations in class are interesting/stimulating.

143 I am generally pleased with the text(s) required for this course.

144 I find the course emphasis on individual projects stimulating.

145 My instructor is not overly demanding of my time.

146 This course has made excellent use of TV.

147 The televised portions of class are a great help to learning.

148 TV reception was of good quality.

149 Audio reception (TV, recorder, etc.) was of good quality.

150 The use of television made the course very interesting.

151 Media (films, TV, etc.) used in this course are well chosen.

152 Media (film, TV, etc.) are an asset to this course.

153 Films in this course contributed significantly to my learning.

154 This course has made excellent use of films.

155 Films in class were well-integrated with course topics.

156 Team teaching is effectively used in this course.

157 Instruction is well-coordinated among the team teachers.

158 Team teaching provided insights as a single instructor could not.

159 The team teaching approach adequately meets my needs/interests.

160 Course topics are dealt with in sufficient depth.

161 Teaching methods used in this course are appropriate to course purposes.

162 The format of this course is appropriate to course purposes.

163 The teaching strategy used in this course is appropriate.

164 This course is accurately described in the catalog.

165 Lecture information is adequately supplemented by other work

166 Class lectures contain information not covered in the textbook.

167 Bibliographies for this course are current and extensive.

168 Mimeographed handouts are valuable supplements to this course.

169 The guest speakers contributed significantly to this course.

170 The speakers who address us communicated effectively.

171 An appropriate number of outside lectures is used.

172 Lab procedures are clearly explained to me.

173 My instructor thoroughly understands lab experiments/equipment

174 Assistance is always available throughout lab sessions.

175 The lab sessions are well-organized.

176 The content of the lab is a worthwhile part of this course.

177 Lab assignments are reasonable in length and complexity.

178 Lab assignments have instructional value.

179 The lab in this course has adequate facilities.

180 The lab assignments are promptly returned to me.

181 The class mixture of Fr., So., Jr., Sr., or Grad is appropriate.

182 The size of this class is appropriate to course objectives.

183 The facilities for this course are excellent.

184 I have easy access to equipment/tools required in this course.

185 I had sufficient opportunity to use lab/practice room facilities.

186 The lab/practice room is well equipped.

187 I highly recommend this course.

188 I would enjoy taking another course from this instructor.

189 I like the way the instructor conducts this course.

190 Frequent attendance in this class is essential to good learning.

191 I am satisfied with my accomplishments in this course.

192 These items let me appraise this course fully and fairly.

193 The services in the math student service center are helpful.

194 I frequently attend the math service center.

195 The grade I expect to receive in this course is (A=SA, B=A, C=U, D=D).

196 My instructor motivates me to do my best work.

197 My instructor explains difficult material clearly.

198 Course assignments are interesting and stimulating.

199 Overall, this instructor is among the best teachers I have known.

200 Overall, this course is among the best I have ever taken.

201 I would recommend this course to another student.

202 I would recommend this instructor to another student.

203 I learned a lot in the course.

204 I looked forward to taking this course before it began.

205 My instructor presents the course in a well-organized manner.

206 My instructor presents material clearly.

207 My instructor is helpful when I have questions.

208 The goals of the course are clearly stated and consistently pursued.

209 For this course, assignments are reasonable.

210 The instructor offers alternatives when criticizing my work.

211 The instructor uses beneficial class critiques in teaching.

212 The instructor uses beneficial individual critiques in teaching.

213 I understand the course objectives.

214 I can determine my standing in the class prior to final grades.

215 The instructor suggests investigation of other artists' work.

216 The instructor is reasonably accessible outside the classroom.

217 The instructor emphasizes various approaches to problem-solving

218 The instructor can clarify information on assignments.

219 The instructor meets class regularly.

220 My interest in this subject has increased as a result of this course.

221 This course has increased my critical thinking skills.

222 My instructor respects students from diverse cultural backgrounds.

223 My instructor respects students regardless of sex, age, or race.

APPENDIX C

Items Used by each Department

Q	A	B	C	E	E	F	H	C	M	P	P	P	S	C	A	M	T	L	S	A	H	N	S	B	I
#	R	I	H	C	N	N	I	M	A	H	O	S	O	T	C	A	E	E	P	S	E	U	W	T	N
	T	O	M	O	G	L	S	P	T	Y	L	Y	C	A	C	R	A	A	E	C	C	R	K	E	D
001																							X		
002						X				X				X	X	X	X		X				X	X	X
003										X															
004																							X		
005																									
006		X				X	X				X	X			X				X	X			X	X	X
007		X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X				X	X	
008																									
009																X									
010																									
011																									
012																									
013																									
014																									
015																									
016						X	X				X	X			X				X					X	
017				X				X	X	X	X	X	X	X		X									
018																									
019																		X				X	X		X

Q #	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MAR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
020																									
021			X																			X			
022																									
023																									
024																									
025					X						X						X			X					
026																									
027																									
028																									
029																									
030																									
031																									
032							X													X					
033																									
034																									
035																									
036							X																		
037																									
038																									
039																X									
040																									

Q#	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MAR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
041			X		X	X				X					X	X			X					X	
042																									
043																									X
044			X			X	X	X				X			X									X	X
045																									
046																									
047																									
048																				X					
049																									
050																									
051										X	X		X	X											
052																									
053																		X							
054																									
055																									
056																									
057		X																					X		
058																									
059																									
060																									
061																									

Q #	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MCR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
062																									
063																									
064																									
065																									
066																									
067																		X							
068																									
069		X																							
070																									
071																									
072																		X							
073											X			X											
074																									
075																									
076																									
077																									
078																	X								
079																									
080																				X					
081													X					X		X					
082								X																	

Q	A	B	C	E	E	F	H	C	M	P	P	P	S	C	A	M	T	L	S	A	H	N	S	B	I
#	R	I	H	C	N	N	I	M	A	H	O	S	O	T	C	A	E	E	P	S	E	U	W	T	N
	T	O	M	O	G	L	S	P	T	Y	L	Y	C	A	C	R	A	A	E	C	C	R	K	E	D
083																									
084		X				X			X		X	X			X	X			X			X	X	X	
085																									
086																									
087																									
088																									
089																									
090																									
091																									
092																	X								
093																									
094																									
095																									
096																									
097																									
098																									
099																									
100		X																X							
101																									
102																									
103																									X

Q #	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MAR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
104																							X		
105											X														
106																									
107																									
108																									
109																									
110																									
111																									
112																									
113																									
114											X	X													
115																									
116																									
117		X																							
118																									
119																									
120																									
121																									
122																									
123																									
124																							X		

Q	A	B	C	E	E	F	H	C	M	P	P	P	S	C	A	M	T	L	S	A	H	N	S	B	I
#	R	I	H	C	N	N	I	M	A	H	O	S	O	T	C	A	E	E	P	S	E	U	W	T	N
	T	O	M	O	G	L	S	P	T	Y	L	Y	C	A	C	R	A	A	E	C	C	R	K	E	D
125			X	X		X		X	X		X	X			X		X	X	X					X	
126										X	X		X			X									
127																									
128																									
129																									
130																									
131																									
132																									
133																	X								
134																X							X		
135																									
136													X												
137																									
138																									
139																									
140																									
141																									
142																									
143											X														
144																									
145																									

Q	A	B	C	E	E	F	H	C	M	P	P	P	S	C	A	M	T	L	S	A	H	N	S	B	I
#	R	I	H	C	N	N	I	M	A	H	O	S	O	T	C	A	E	E	P	S	E	U	W	T	N
	T	O	M	O	G	L	S	P	T	Y	L	Y	C	A	C	R	A	A	E	C	C	R	K	E	D
146																									
147																									
148																									
149																									
150																									
151																									
152																									
153																									
154																									
155																									
156																									
157																									
158																									
159																									
160																									
161																									
162																									
163																					X				
164																									
165																									
166																									

Q #	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MAR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
167																									
168																									
169																									
170																									
171																									
172																									
173																									
174																									
175																									
176																									
177																									
178																									
179																									
180																									
181																									
182																									
183																									
184																									
185																									
186																									
187											X														

Q #	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MAR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
188											X														
189																									
190																									
191																									
192																									
193									X																
194									X																
195												X													
196															X				X					X	
197																									
198																									
199																									
200																									
201									X			X													
202						X						X													
203			X	X				X				X			X	X			X					X	
204												X													
205		X		X														X			X	X			
206				X																					
207				X																	X	X			
208		X		X	X																				

Q #	ART	BIO	CHM	ECO	ENG	FNL	HIS	CMP	MAT	PHY	POL	PSY	SOC	CTA	ACC	MAR	TEA	LEA	SPE	ASC	HEC	NUR	SWK	BTE	IND
209				X																					
210	X																								
211	X																	X							
212	X																								
213	X																								
214	X																								
215	X																								
216	X	X			X				X																
217	X																								
218	X																								
219	X								X					X											
220																									
221																									
222														X			X						X		
223					X												X						X		

List of table abbreviations:

Q#   Question Number (refer to Appendix B)

ART – Art Department

BIO – Biology Department

CHM – Chemistry Department

ECO – Economics Department

ENG - English Language and Literature Department

FNL - Foreign Languages Department

HIS - History and Philosophy Department

CMP - Computer Science Department

MAT - Mathematics Department

PHY - Physics and Astronomy Department

POL - Political Science Department

PSY - Psychology Department

SOC - Sociology, Anthropology and Criminology Department

CTA - Communications and Theatre Arts Department

ACC - Accounting and Finance Department

MAR - Marketing Department

TEA - Teacher Education Department

LEA - Leadership and Counseling Department

SPE - Special Education Department

ASC - Associated Health Department

HEC - Health, Environmental and Consumers Resources Department

NUR - Nursing Department

SWK - Social Work Department

BTE - Business and Technology Education Department

IND - Industrial Technology Department

APPENDIX D

CORRELATION MATRIXIES FOR STUDY 3

Overall Student Ratings of Teaching Effectiveness for Fall 2001 and Winter 2002

	DV	STYLE	PREP	INTER	ENTHU	HELP	EXPT	EXAM	GRD_1	REC_C	REC_I	LEARN	FWD	GRD_2
DV	-	.98*	.85*	.92*	.83*	.87*	.76*	.66*	.78*	.84*	.98*	.86*	.18	-.15
STYLE	.94*	-	.84*	.94*	.86*	.84*	.71*	.56‡	.74*	.80*	.97*	.84*	.19	-.22
PREP	.56*	.60*	-	.76*	.68*	.86*	.81*	.62*	.80*	.63*	.86*	.74*	-.01	-.34†
INTER	.88*	.89*	.53‡	-	.89*	.76*	.64*	.54‡	.66*	.86*	.91*	.89*	.32†	-.13
ENTHU	.74*	.72*	.42‡	.79*	-	.76*	.48‡	.40†	.54‡	.69*	.82*	.79*	.35†	-.06
HELP	.61*	.60*	.64*	.62*	.52‡	-	.72*	.67*	.77*	.69*	.89*	.75*	.18	-.14
EXPT	.41‡	.46‡	.56*	.41‡	.15	.56*	-	.75*	.88*	.70*	.75*	.64*	-.01	-.41†
EXAM	.53‡	.39†	.15	.37†	.21	.23	.46‡	-	.73*	.73*	.68*	.65*	.15	.15
GRD_1	.65*	.57*	.53‡	.62*	.40†	.51‡	.71*	.80*	-	.68*	.82*	.63*	.01	-.39†
REC_C	.75*	.72*	.49‡	.83*	.63*	.47‡	.48‡	.58*	.79*	-	.82*	.85*	.47‡	.05
REC_I	.92*	.95*	.63*	.87*	.70*	.70*	.62*	.49‡	.69*	.76*	-	.86*	.15	-.15
LEARN	.84*	.81*	.53‡	.87*	.63*	.55*	.43‡	.35†	.57*	.71*	.80*	-	.34†	-.04
FWD	.26	.14	.12	.41‡	.13	.11	.12	.26	.37†	.58*	.12	.44‡	-	.26
GRD_2	.22	.07	-.14	.15	.24	.05	-.26	.41†	.21	.32†	.06	.20	.28	-

Note. DV = Overall student ratings of teaching effectiveness. Correlations above the diagonal are for Winter 2002 and below are for Fall 2001.
†*p*<.05, ‡*p*<.01, **p*<.001

Overall Student Ratings of Courses for Fall 2001 and Winter 2002

	DV	STYLE	PREP	INTER	ENTHU	HELP	EXPT	EXAM	GRD_1	REC_C	REC_I	LEARN	FWD	GRD_2
DV	-	.89*	.80*	.89*	.74*	.76*	.78*	.77*	.78*	.93*	.92*	.91*	.27	-.09
STYLE	.90*	-	.90*	.91*	.87*	.88*	.79*	.67*	.80*	.86*	.97*	.86*	.16	-.24
PREP	.63*	.72*	-	.77*	.81*	.88*	.84*	.68*	.86*	.77*	.90*	.80*	.03	-.32†
INTER	.88*	.90*	.62*	-	.83*	.73*	.73*	.59*	.72*	.85*	.90*	.90*	.29†	-.12
ENTHU	.75*	.84*	.61*	.87*	-	.84*	.61*	.49‡	.68*	.74*	.84*	.78*	.26	-.17
HELP	.65*	.69*	.66*	.71*	.74*	-	.81*	.69*	.86*	.79*	.89*	.73*	.15	-.22
EXPT	.60*	.65*	.62*	.63*	.57*	.73*	-	.79*	.93*	.77*	.84*	.71*	-.02	-.30†
EXAM	.31†	.31†	.24	.29†	.26	.41‡	.63*	-	.77*	.75*	.76*	.72*	-.06	.01
GRD_1	.67*	.68*	.68*	.65*	.53*	.70*	.76*	.72*	-	.77*	.87*	.69*	.04	-.38†
REC_C	.88*	.85*	.71*	.88*	.74*	.64*	.63*	.27†	.65*	-	.90*	.85*	.33†	.05
REC_I	.88*	.95*	.72*	.90*	.85*	.76*	.75*	.41‡	.74*	.86*	-	.87*	.10	-.15
LEARN	.87*	.85*	.65*	.87*	.73*	.62*	.61*	.35†	.67*	.86*	.85*	-	.24	-.04
FWD	.39‡	.26	.25	.42‡	.24	.28†	.24	.21	.26	.48‡	.22	.49‡	-	.09
GRD_2	-.01	-.03	-.16	.14	.09	.13	-.10	.15	-.01	.07	.03	.21	.33†	-

Note. DV = Overall student ratings of courses. Correlations above the diagonal are for Winter 2002 and below are for Fall 2001.
†*p*<.05, ‡*p*<.01, **p*<.001

APPENDIX E

Glossary of Terms

Absolute Error Variance – one of two types of error variance within generalizability theory, also called Δ-type error, used when the researcher is interested in whether a person can perform at a pre-specified level or the researcher is interested in rank ordering and differences in average scores.

Absolute Expected Grades – the grade a student expects to receive within a course in absolute terms (i.e., A, B, C or 4.0, 3.5, 2.0, etc…).

Cafeteria-Style System – a system in which a list of available items is given for an instructor or committee could select from to generate the form used to rate the instructor and course.

Classical Test Theory (CTT) – theory that an observed score for any person obtained through some measurement can be decomposed into the true score and a random error component.

Coefficient Alpha – a measure of internal consistency which is equivalent to having conducted all possible split-half internal consistency analysis (also call Cronbach’s Alpha).

Contaminant – a variable that could potentially affect the relationship between independent and dependent variables (also called confounding variable).

Convergent Validity – an indication of validity that a measurement measures the construct of interest based on other measures of the same construct.

Decision (D) Study – within generalizability theory is used to emphasize the estimation, use and interpretation of variance components.

Discriminant Validity – An indication of validity that a measurement is not measuring some other construct than the one desired.

Expected Grade – the grade a student expects to receive within a course.

Facet – used within generalizability theory, refers to a set of similar conditions of measurement.

Generalizability (G) Study – refers to the initial study of a measurement procedure within generalizability theory. A G study is used to obtain estimates of variance components for the universe of admissible observations.

Generalizability Theory – a random sampling theory used to examine the dependability of a measurement (also called G theory).

GENOVA – GENeralized Analysis Of VAriance, computer program used to estimate variance components and calculated generalizability coefficients.

IDEA – a widely used student ratings form developed by Kansas State University.

Internal Consistency – a measure of reliability when only one administration of a measurement was performed to see if items in the measure are consistent with each other.

Inter-rater Agreement – the extent to which raters agree on the score of an item within a measure.

Population – used within generalizability theory, refers to the objects of measurement.

Relative Error Variance – one of two type of error variance within generalizability theory, also called δ-type error, used when the researcher wants to make decisions about individual differences between persons.

Relative Expected Grades – the grade a student expects to receive within a course in relative terms to grades received in other courses they have taken.

Reliability – the degree to which a measure would produce the same results from one occasion to another.

Spearman-Brown Prophecy Formula – a statistical formula used to determine the number of items that would be needed to achieve different levels of reliability.

Structural Modeling – mathematical method for explicitly testing a theoretical model.

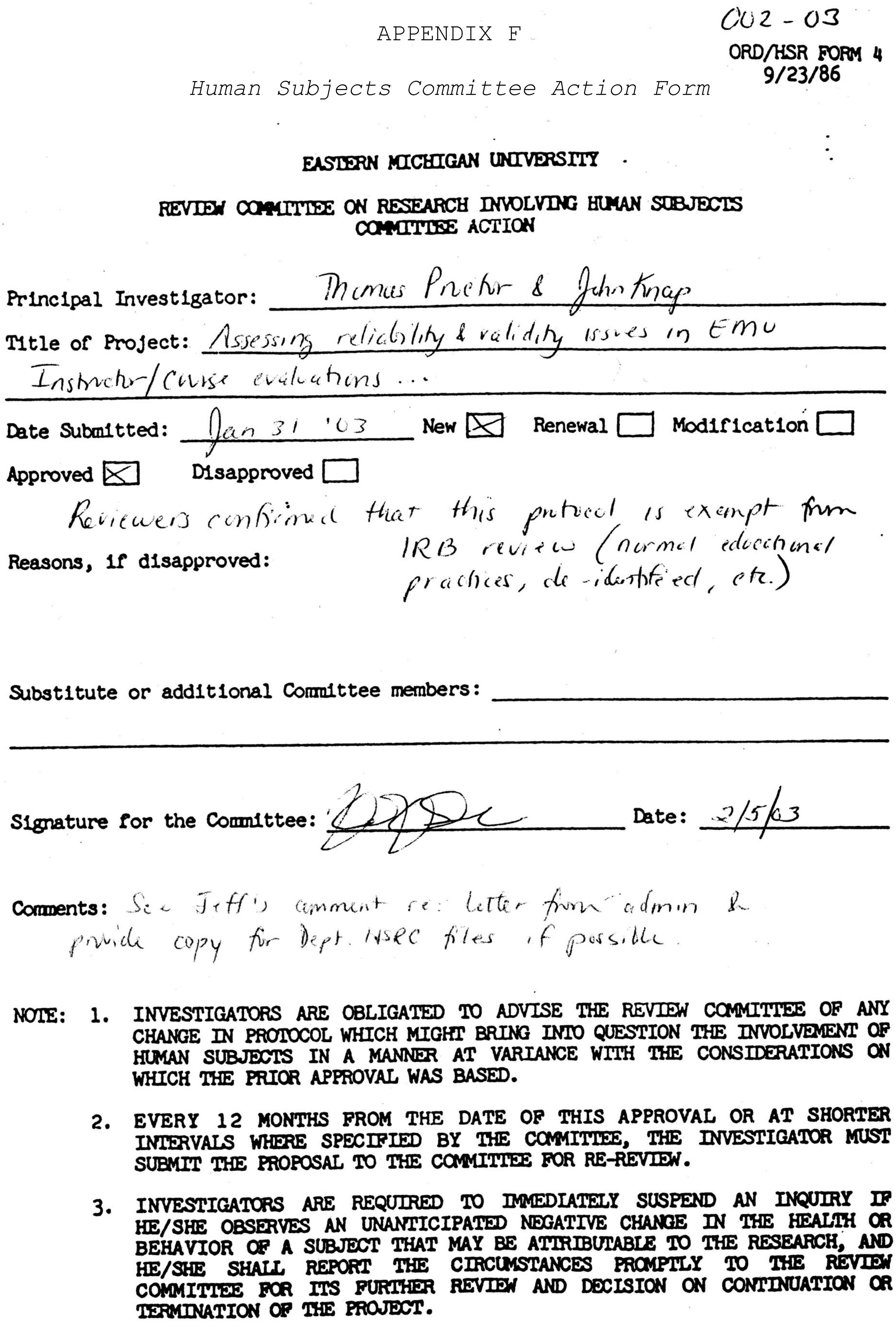
Student Ratings of Teacher Effectiveness – any systematic method of collecting ratings by students of a teacher or course.

Universe – used within generalizability theory, refers to all admissible conditions of measurement.

Validity – whether what is being measured is what the researcher really wants to measure.

APPENDIX F

Human Subjects Committee Action Form



002 - 03

ORD/HSR FORM 4
9/23/86

EASTERN MICHIGAN UNIVERSITY

REVIEW COMMITTEE ON RESEARCH INVOLVING HUMAN SUBJECTS
COMMITTEE ACTION

Principal Investigator: Thomas Prehr & John Knap

Title of Project: Assessing reliability & validity issues in EMU Instructor/Course evaluations ...

Date Submitted: Jan 31 '03 New ☒ Renewal □ Modification □

Approved ☒ Disapproved □

Reviewers confirmed that this protocol is exempt from

Reasons, if disapproved: IRB review (normal educational practices, de-identified, etc.)

Substitute or additional Committee members: ______

Signature for the Committee: ______ Date: 2/5/03

Comments: See Jeff's comment re: letter from admin & provide copy for Dept. HSRC files if possible.

NOTE: 1. INVESTIGATORS ARE OBLIGATED TO ADVISE THE REVIEW COMMITTEE OF ANY CHANGE IN PROTOCOL WHICH MIGHT BRING INTO QUESTION THE INVOLVEMENT OF HUMAN SUBJECTS IN A MANNER AT VARIANCE WITH THE CONSIDERATIONS ON WHICH THE PRIOR APPROVAL WAS BASED.

2. EVERY 12 MONTHS FROM THE DATE OF THIS APPROVAL OR AT SHORTER INTERVALS WHERE SPECIFIED BY THE COMMITTEE, THE INVESTIGATOR MUST SUBMIT THE PROPOSAL TO THE COMMITTEE FOR RE-REVIEW.

3. INVESTIGATORS ARE REQUIRED TO IMMEDIATELY SUSPEND AN INQUIRY IF HE/SHE OBSERVES AN UNANTICIPATED NEGATIVE CHANGE IN THE HEALTH OR BEHAVIOR OF A SUBJECT THAT MAY BE ATTRIBUTABLE TO THE RESEARCH, AND HE/SHE SHALL REPORT THE CIRCUMSTANCES PROMPTLY TO THE REVIEW COMMITTEE FOR ITS FURTHER REVIEW AND DECISION ON CONTINUATION OR TERMINATION OF THE PROJECT.