2008

Instructions, response anchors, and scale length: How do variations affect student evaluations of teaching effectiveness?

Pamela G. Ing
University of Windsor

Instructions, Response Anchors, and Scale Length: How Do Variations Affect Student Evaluations of Teaching Effectiveness?

by

Pamela G. Ing

A Thesis
Submitted to Faculty of Graduate Studies
through Psychology
in Partial Fulfillment of the Requirements for
the Degree of Master of Arts
at the University of Windsor

Windsor, Ontario, Canada

2008

Your file Votre référence
ISBN: 978-0-494-47021-3
Our file Notre référence
ISBN: 978-0-494-47021-3

AUTHOR'S DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

This study investigated the effects of varying instructions, response anchor wording, and response scale length on students' evaluations of teaching effectiveness for liked and disliked courses. Four separate (one for each teaching effectiveness dimension) 2 (course) x 3 (instructions) x 2 (response anchor wording) x 2 (response scale length) analyses of variance revealed main effects for the course and length of response scale variables and an interaction effect between the length of response scale and course variables. No other main or interaction effects were observed. Implications resulting from this study and possible future directions are discussed.

DEDICATION

For my Mom and Daddy, who have continually inspired and supported me.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dennis Jackson, and my thesis committee members, Dr. Robert Orr and Dr. Kai Hildebrandt, for their advice and sincere interest in my research.

I need to thank my sister Cynthia for her unconditional support and encouragement.

I must also thank Lili, Mark, and Zakiya for their continuous support, encouragement, and love.

I would also like to thank three very good friends that have always supported my academic endeavours, Dawne, Mario, and Michelle.

TABLE OF CONTENTS

CHAPTER I

Introduction

Within psychology, there is a pervasive reliance upon questionnaire data to provide the critical information required to make decisions regarding individual outcomes. For instance, a variety of measures are employed by industrial-organizational psychologists in personnel selection (Kanning & Kuhne, 2006), while various scales are implemented by health psychologists for patient assessment and program design (Gotay et al., 2002). Perhaps the most familiar example can be witnessed through the use of student evaluations of teaching effectiveness for promotion and tenure decisions (Chen & Hoshower, 2003; Zabaleta, 2007).

Existing scales are sometimes revised, while new surveys are continually being developed across all subfields of psychology. A striking absence in the questionnaire development literature involves providing rationale for decision making throughout survey construction (Jackson, Gillaspy, & Purc-Stephenson, 2007; Jackson, Ing, & Arseneault, 2007). For example, researchers rarely explain how they select response scale anchors or the number of included response categories. Lack of conscientious decision making in scale construction is an immediate cause for concern when one considers the amount research conducted in psychology that is reliant upon survey data.

The intention of this research is to investigate critical components of questionnaire construction, with the purpose of investigating whether variations in certain scale properties independently or interactively influence individuals' ratings of others. More specifically, the effects of varying instructions, response scale anchors, and length of response scale on student evaluations of teaching effectiveness (SETEs) will be

examined. Due to the nature of this study, literature will be reviewed for: student evaluations of teaching effectiveness; scale construction; sources of response biases and errors; cognitive psychology applied to survey methodology; and recent studies involving instructions, instructor ratings, and response anchors.

CHAPTER II

Literature Review

Student Evaluations of Teaching Effectiveness

Student evaluations of teaching effectiveness (SETE) have gained widespread acceptance in universities around the world, including institutions in Canada, the United States, the United Kingdom, Australia, and several European countries (Chen & Hoshower, 2003; Moore & Kuol, 2005; Zabaleta, 2007). Their implementation has been intended to facilitate two functions: formative, by providing feedback to instructors for teaching style enhancement; and summative, by producing evaluations of teaching effectiveness, which serve as relevant criteria in administrative and personnel decisions (Chen & Hoshower, 2003; Crumbley, Henry, & Kratchman, 2001; Moore & Kuol, 2005; Sedlmeier, 2006; Zabaleta, 2007). SETE ratings often contribute to irreversible administrative decisions, while also possessing the influential power to enhance or defame an instructor's reputation (Crumbley et al., 2001). Despite the nearly universal employment of SETE instruments, there exist no standards by which teaching evaluation forms are to be constructed (Sedlmeier, 2006; Zabaleta, 2007).

Instructor ratings, derived from SETE, have proven useful for many functions, including: personnel, curriculum, and faculty resource allocation decisions (Crumbley et al. 2001); enriching student learning outcomes; improving instructor performance (Zabaleta, 2007); revealing the presence of and/or severity of teaching related concerns or problems; and identifying groups of students experiencing shared difficulties (Moore & Kuol, 2005). Student evaluations of instructor performance commonly take the form of self-administered paper and pencil questionnaires (Cashin, 1995; Jackson et al, 1999;

Marsh, 1982), while technological advances have led some institutions to administer these evaluations through a web survey format (Thorpe, 2002, as cited in Gamliel & Davidovitz, 2005). These SETE forms are generally filled out by university students towards the end of each academic course (Zabaleta, 2007; Koh & Tan, 1997; Marsh, 1982).

Previous research regarding sources of variability in SETE ratings has included characteristics of instructors, classes, students, and subject matter (Crumbley et al., 2001; Koh & Tan, 1997). Additionally, many studies have incorporated structural equation modeling techniques to investigate the validity of specific SETE instruments and the dimensionality (unidimensional versus multidimensional) of SETE ratings (Jackson, et al., 1999; Marsh, 1982, 1984, 1991). Recent studies have examined the appropriate use of SETE instruments (Emery, Kramer, & Tian, 2003) and the effects of variations of item wording and scale numbering (Sedlmeier, 2006).

Prior research has also provided evidence for certain factors that affect the variability of results in SETE ratings. Crumbley et al. (2001) presented several sources of variation in SETE results, including: instructor gender (Basnow & Silber, 1987); class size (Meredith, 1984; Toby, 1993); student gender (Sidanius & Crane, 1989); subject matter; course content; and student age, abilities, achievements, expectations, and classification (Perkins, Gueri, & Schleh, 1990).

Koh & Tan (1997) also acknowledged variables that affect SETE ratings. These factors include characteristics of instructors, classes, evaluations, and subjects. Instructor characteristics consisted of age, gender, and rank. Some researchers have indicated that better SETE ratings are received by more experienced and older instructors (Smith &

Kinney, 1992), while others have suggested that the relationship between SETE ratings and instructor age/experience is non-linear in nature (Langbein, 1994). Research involving instructor gender has resulted in conflicting findings, with both females and males receiving lower ratings than their counterparts in different studies (Fandt & Stevens, 1991; Langbein, 1994; Lueck, Endres, & Caplan, 1993). Instructor rank has been positively associated with SETE ratings, with higher ranking leading to higher ratings (Holtfreter, 1991). DeBerg & Wilson, 1990) and class size (Holtfreter, 1991; Toby, 1993). With regard to evaluation characteristics, greater differences have been found for SETE ratings between dates of administration than across different respondents (Cronin & Capie, 1986). Higher level courses (Holtfreter, 1991; Marsh, 1984; Langbein, 1994) and smaller class sizes (Holtfreter, 1991) are likely to receive more favourable ratings than lower level courses and larger sized classes respectively.

Because students' perceptions of instructional quality are usually viewed as multidimensional in nature (Burdsal & Bardo, 1986; Harrison et al., 2004; Jackson et al., 1999; Marsh, 1982, 1984, 1991; Marsh & Roche, 1997), results from SETE forms are intended to reflect various aspects of instructor performance. Several factor analytic studies have confirmed the multidimensionality of SETE instruments (Burdsal & Bardo, 1986; Cashin, 1995; Jackson et al., 1999; Marsh, 1991; 1984). Cashin (1995) presented six factors that are often identified in teaching evaluation questionnaires as: course organization and planning; clarity, communication skills; teacher student interaction, rapport; course difficulty, workload; grading and examinations; and student self-rated learning (Centra, 1993; Breskamp & Ory, 1994). Analyses by Burdsal & Bardo (1986) and Jackson et al. (1999) have identified a similar six-dimension structure for the Student

Perceptions of Teaching Effectiveness Scale (SPTE-II), including: rapport with students; course value; course organization and design; grading fairness; course difficulty; and workload. Analyses of Marsh's (1984, 1991) Students' Evaluations of Educational Quality (SEEQ) form identified the following nine dimensions of teaching effectiveness: learning/value; enthusiasm; organization; group interaction; individual rapport; breadth of coverage; exams/grades; assignments; and workload.

Some recent studies have examined the effects of scale properties on SETE ratings. It has been suggested that responses may be markedly influenced by seemingly harmless variations in questionnaire construction, such as item wording and response scale formats (Jackson et al., 2007b; Sedlmeier, 2006). According to Sedlmeier (2006), a respondent's level of certainty, regarding an item, may act as a moderating variable with the influence of response anchor labels, wherein greater certainty decreases the ability of anchors to affect responses. Emery et al. (2003) have recommended that SETE be constructed with an achievement orientation (i.e., the level of achievement gained through the course being evaluated) versus a satisfaction orientation (i.e., how satisfied students are with the instructor's performance). Furthermore, they have suggested that students be obliged to provide specific comments regarding less than satisfactory ratings, thus providing a means to investigate the validity of negative ratings.

Scale Construction

For questionnaires in general and SETE in particular, there is a lack of consensus regarding the optimal number of scale points and the most appropriate response scale anchors (Alwin, 1997; Jackson et al., 2007b; Sedlmeier, 2006). According to Weng (2004) the relationships of number of scale points with internal-consistency reliability

and test-retest reliability have been both inconsistent and inconclusive. Furthermore, recommendations for number of response categories have ranged from two to 11 scale points (Alwin, 1997; Weng, 2004). Higher numbers of scale points have been found to more efficiently facilitate the accurate measurement of participants' responses (Alwin, 1997; Weng, 2004). It has been suggested that a minimum of five scale points is required to accurately represent the direction, intensity, and neutrality region in the measurement of attitudes (Alwin, 1997; Weng, 2004). Moreover, it has been reported that if respondents are cognitively operating at a post-secondary education level of ability, consistent and reliable data may be acquired through the use of even (six points) or odd (seven points) numbered response categories (Weng, 2004).

The implementation of Likert-type scales has become so commonplace in psychological research, that it appears to have become the default method for researchers interested in measuring participant attitudes, attributes, and ratings of others (Dawis, 1987; Weng, 2004). The conventional order of response anchors for Likert-type scales presents the most favourable option first (left side) and the least favourable option last (right side) (Chan, 1991). In other words, the response anchors move from positive to negative wording, following a left to right direction. Thus, it is important to consider that presenting a scale with the response anchors reversed (i.e., negative anchor on the left side and positive anchor on the right side), changes not only the placement of the anchors, but also the sequence of respondent's information (Chan, 1991). An additional consideration with the use of Likert-type scales is whether to label each scale point or to label only the endpoints (anchors). Several studies indicate that reliability is not adversely affected by the selection of one alternative over the other (e.g. Weng, 2004).

Sources of Response Biases and Errors

There are certain elements that researchers should be aware of when implementing SETE forms. Difficulties may arise when using self-administered questionnaires, which could complicate the interpretation of survey results. Researchers should be conscious of the effects of such problems as: acquiescence bias, social desirability, middle position and "don't know" responses, and context effects caused by question order and response alternatives. Survey responses may be sensitive to the precise wording, format, and placement of questions asked (Kalton & Schuman, 1982).

Acquiescence bias is the tendency to endorse questionnaire items in an unreasonably affirmative manner, regardless of item content (Johanson & Osborn, 2004; Knowles & Nathan 1997; Schuman & Presser, 1981). It has been suggested that acquiescence bias is a response pattern caused by ambiguity in questionnaire items and that the magnitude of the bias is influenced by the items and the sampled subpopulation (Hurd, 1999). The problem that acquiescent responding poses is that endorsements lose meaning and researchers are left unable to interpret respondents' results (Ray, 1983). According to Knowles and Nathan (1997), most measurement specialists recommend controlling for acquiescence problems by creating balanced item sets which include equal numbers of trait indicators and trait contraindicators (items for which a "no" response identifies the presence of the trait in question). Questionnaires comprised of balanced item sets enable researchers to develop separate measures of the trait and of the acquiescent responding.

Knowles and Nathan (1997) acknowledged two perspectives regarding acquiescent response sets. The first views acquiescence as a motivational issue, often as

an impression management problem (Couch & Keniston, 1960, 1961). In the second perspective, Cronbach (1942, 1950) describes acquiescent response sets as the uncritical acceptance of items, caused by cognitive processing problems. Knowles and Nathan (1997) found evidence of a general acquiescence trait, exhibiting a fairly clear pattern of correlations with personality variables, thus, indicating a cognitive style. Results suggested that acquiescent responders, yea-sayers, are rigid in their mental organization, cognitively simple, and intolerant of alternatives, while oppositional responders, nay-sayers, are forgetful, disordered, welcoming of dissent and more cognitively complex. Couch and Keniston (1960, 1961) concluded that impulsiveness, extraversion, emotionality, and a lack of control characterized yea-sayers, while rationality, caution, introversion, and a surplus of control described nay-sayers (as cited in Knowles & Condon, 1999).

Acquiescence bias may also appear when increased cognitive demands cause responders to use heuristic (shortcut) and peripheral route (non-direct) processes, creating a confirmation bias (Knowles & Nathan, 1997; Knowles & Condon, 1999). According to Krosnick (1991), acquiescent responding results from weak satisficing (selecting the first acceptable response option presented), whereas social desirability bias is linked to strong satisficing. Knowles and Condon (1999) supported a two-stage Spinozan belief process described by Gilbert (1991), in which an item is first comprehended and then reconsidered. It is important to be aware of and control for such response biases because they can inflate reliability and depress validity (Knowles & Condon, 1999).

When completing self-administered questionnaires, there is often a logical middle position that appeals to certain respondents more so than either end of the implicit

attitude dimension (Presser & Schuman, 1980). If a middle alternative is not offered in a response scale, concerns may arise regarding whether answers from respondents that would have preferred a middle position have contributed to some form of random or systematic error, through the use of forced-choice responses (Presser & Schuman, 1980). However, it has been reported that the quality of data is not affected by the presence of a distinct middle response category (Andrews, 1984, as cited in Si & Cullen, 1998). In other words, middle responses are generally derived equally from both extremes of a given scale.

When an explicit middle option is offered as part of the question being asked, the proportion of responses in that category becomes considerable, with the majority stemming from decreased extreme positions, versus declines in "don't know" responses (Bishop, 1987; Schuman & Presser, 1981). The presentation of a middle alternative may be especially attractive for respondents who are ambivalent to the other options available to them. With regard to whether a middle response is offered or not, there is generally no significant effect on the proportion of responses distributed in either end of the scale (Schuman & Presser, 1981; Presser & Schuman, 1980). However, it is possible that the absence or presence of a middle position may influence respondents' perceptions of the kind of information being elicited from the question (Presser & Schuman, 1980).

Social desirability bias occurs when respondents display a tendency to present themselves in a favourable light, as prescribed by social mores and norms, in order to gain the approval of others (King & Bruner, 2000). When respondents are completing self-administered questionnaires, they may edit their answers to create socially desirable responses (Sudman, Bradburn & Schwarz, 1996). Thomas and Kilmann (1975) proposed

that the use of ratings for assessing variables with evaluative overtones generates the contaminating effects of social desirability bias (as cited in King & Bruner, 2000). Thus it is possible that such a bias could influence responses, even when certain items do not directly refer to the individual respondent. Thus, items included in student evaluation forms which do not presumably reveal respondents' personal information, may unsuspectingly elicit the need to edit questionnaire responses. In other words, students may feel compelled to provide the most socially acceptable (as prescribed by norms, values, and morals) responses when they are completing course evaluations. Social desirability bias has been referred to as one of the most common and pervasive sources of bias affecting the validity of survey research findings in psychology (King & Bruner, 2000). This bias has the capability of decreasing the validity of research involving multi-item scales (King & Bruner, 2000). It may also moderate or suppress correlations of constructs of interest or cause spurious correlations between variables (King & Bruner, 2000).

Schuman and Presser (1981) address order effects on questionnaire type surveys. Due to the context-dependent nature of survey responses, order effects may result from the order in which questions are listed, as well as the order in which response alternatives are presented. Answers to items completed initially in a sequence of questionnaire items may influence the responses to subsequent questions. Ordering of response options may lead to primacy or recency effects where individuals systematically choose the first (primacy) or last (recency) alternatives available in response scales. When two or more items address topics that are closely related or facets of the same subject, context effects occur. General summary-type questions appear to be more easily influenced by order

effects than more specific questions. When considered in conjunction, two items may establish or highlight a norm that is not apparent when each item stands alone. In other words, when answers from specific questions remain salient in respondents' minds, they are likely to influence the responses given for overarching or summary items.

Schuman and Presser (1981) also discuss the issues surrounding "don't know" responses. It is reasonable to believe that for a given questionnaire item, a "don't know" response may be the most appropriate answer for some respondents. Filtered question forms (i.e., explicitly provide a "don't know" or "no opinion" response alternative. Implementing filtered questions can significantly increase the proportion of respondents selecting "don't know" options.

Cognitive Psychology Applied to Survey Methodology

Researchers have also applied cognitive psychology to the survey methodology literature. Sudman, Bradburn and Schwarz (1996) discuss the application of cognitive processes to survey methodology. They identify the main psychological components involved with answering survey items as: question interpretation; information retrieval; opinion generation or representation of relevant behaviour; response formation; and response editing. A respondent's cognitive depiction of an issue, as well as subsequent behaviour, may be influenced by the cognitive processes involved with answering a question.

The psychological sources of context effects were also discussed by Sudman et al. (1996). When context effects occur, the content of previous items may influence the interpretation of following items. Information used to respond to preceding items may become more readily accessible in memory and influence responses to subsequent

questions. The accessibility of cognitive procedures may also be enhanced by implementation in preceding items, thereby increasing the potential for use in subsequent questions. Influence of previous items may also extend to individuals' anchoring of response scales, as well as concerns regarding social desirability and self-presentation. It is likely that the introduction of a general norm in one item will promote the use of that norm in subsequent questions to which the norm is applicable. For example, if a job applicant is completing an integrity test and the social norm of "not stealing from one's employer" is mentioned in one of the test items, this norm is likely to remain salient for the individual, thus influencing responses that follow. Ambiguity also plays a role in context effects. Context effects are more likely to arise when ambiguity is present in questionnaire items. Questions creating subjective experiences, such as mood change in respondents are also likely to affect subsequent judgments and responses.

Sudman et al. (1996) cite the recurrent attribution of response order effects to the limitations of respondents' memories. Response choices are constrained by memory performance, leaving more accessible options more likely to be selected. Cognitive psychologists have examined the serial position effect under immediate and delayed recall conditions (Rundus, 1971). With immediate recall, items appearing at the end of a list are most likely to be recalled, followed by items presented at the beginning of a list, with middle options being least likely to be recalled (recency effect). Under delayed recall conditions, primacy effects (better recall for items appearing at the beginning of a list) are expected to emerge. Krosnick (1991) explains the appearance of primacy effects with satisficing, where respondents select the first acceptable option from among all presented responses. When respondents are satisficing, they answer items in a manner

that allows them to conserve time and energy, thus reducing cognitive strain, while still arriving at acceptable responses.

Primacy effects have consistently been identified in questionnaire research involving extensive lists of response alternatives (Payne, 1951; Becker, 1954; Krosnick & Alwin, 1987; Mueller, 1970; Ring, 1975, all as cited in Sudman et al., 1996). However, recency effects are most likely to occur when numerous or complicated response options are offered, causing strain on respondents' memory. A three-way interaction between item plausibility, presentation mode, and serial position predict that a plausible item, eliciting agreeing thoughts, with a visual (versus auditory) presentation will produce a primacy effect. In other words, if an item is believable, requires the respondent's confirmation, and is presented visually, recall will be best for first presented options. The size and frequency of response order effects along rating scales appear less pronounced than for those involving discrete response alternatives.

Survey questions often require respondents to search their memories in order to aid in making judgments for specific items. Williams and Hollan (1981, as cited in Sudman et al., 1996) designate remembering as a multi-stage reconstructive retrieval process. This process involves the use of information about items to generate descriptions, which may in turn retrieve additional information used to further refine descriptions until desired items are finally retrieved. They suggest that when respondents experience partial memory failure, they may commit errors of omission, in which they forget relevant information, or errors of commission, in which they report additional, yet inaccurate information. Further, respondents do not depend solely on their memories for specific events. Individuals integrate generic information regarding types of events into

their memories of specific events. They also summarize and classify events into groups and add logically consistent information to the memories they are able to retrieve.

Events will appear to be more frequent, likely or recent when relevant memories can be easily brought to mind. When such an availability heuristic (Tversky & Kahneman, 1973) is employed, it may lead to systematic error, appearing when frequent events are difficult to recall, or when rare events are easily remembered. According to Sudman et al. (1996) the key to understanding the retrieval process is to understand the way in which memory is organized. Evidence exists for a hierarchical organization and the use of coding schemes as guideposts in information/memory retrieval. Remembering is a complex process which may be influenced by habitual behaviour and the attitudes, emotion, and events occurring at the time of retrieval, including the context and wording of the question initiating the retrieval process.

Sudman et al. (1996) identified some implications for questionnaire design, in reference to the application of cognitive psychology to survey methodology. They note that survey questionnaires are channeled through the use of language. Possessing awareness of how individuals organize spoken and written language is beneficial in the comprehension of response effects. It is necessary for formal structure (syntax) and pragmatic factors (that may affect meaning) to be clear. Attitude questions require respondents to assign evaluative ratings. The perpetual context-dependency of human judgment becomes problematic when researchers wish to generalize sample specific survey results to the general population. Researchers designing questionnaires require an understanding of the communicative and cognitive processes fundamental to question proposals and responses, in order to recognize expected problems.

According to the relevant literature, the implementation of SETE forms is a widely accepted practice in university settings for obtaining instructor ratings (Chen & Hoshower, 2003; Moore & Kuol, 2005; Zabaleta, 2007). Self-administered pencil and paper questionnaires have traditionally been the most common method for measuring student ratings of teachers (Cashin, 1995; Marsh, 1982), although advances in technology are allowing for more universities to employ web-based SETEs (Thorpe, 2002, as cited in Gamliel & Davidovitz, 2005). Such survey questionnaires are vulnerable to various sources of biases and errors, which may contaminate the interpretation of results. Researchers should be aware of possible complications which may arise as a result of random or systematic error caused by response biases. In addition, an understanding of cognitive psychology enables researchers to be aware of the possible influences of cognitive processes on respondents' answers. Furthermore, with an absence of agreed upon guidelines for scale construction, it is important to consider the possibility of interaction effects between biases and other variables. For example, could scale length have an effect on undergraduates but not on graduate level students?

Recent Studies Involving Instructions, Instructor Ratings, and Response Anchors

Three recent studies have investigated the effects of varying instructions and response anchor wording on students' ratings of teaching effectiveness (Arseneault & Jackson, 2005; Ing & Jackson, 2006, 2007). Arseneault and Jackson (2005) employed instructional, course, and response anchor wording manipulations in their study. They examined whether asking for an opinion or an evaluation would result in significantly different responses, a possible suggestion from levels-of-processing theory. For their

study, 80 undergraduate psychology students completed two SETE questionnaires each, one for a liked course and one for a disliked course.

Items appearing on their SETE questionnaires represented four dimensions of teaching effectiveness, derived from the Student Perceptions of Teaching Effectiveness Scale, second edition, (SPTE-II) and presented on seven-point Likert type scales, with agreement (*Strongly Agree* or *Strongly Disagree*) and evaluative (*Very Good* or *Very Poor*) response anchors. Four (one for each teaching effectiveness dimension) separate 2 (course) x 2 (instructions) x 2 (response anchor wording) mixed randomized by repeated measures analyses of variance (ANOVAs) indicated that there was no main effect for instructions. In some conditions, though, there was a main effect for response anchor wording and some significant two-way and three-way interactions between the independent variables. Specifically, all of the significant interaction effects were due to more negative ratings by participants who had received *agreement* response anchors, especially when rating a *disliked* course. These results were contrary to conventional beliefs about acquiescence bias. This research question was extended in two follow-up studies by Ing and Jackson (2006, 2007).

Ing and Jackson (2006) extended the Arseneault and Jackson (2005) study by implementing some changes to the original design. To increase power, the seven-point Likert-type response scale was increased to a 10-point scale and the sample size was increased from 80 to 192 participants. Because items appearing on the SPTE-II are presented with semantic differential (item-specific) response anchors, the researchers decided to include the original scale items that parallel the items implemented in the previous study (presented with agreement and evaluative anchors). The original items

were included in order to investigate how the ratings elicited through all three response anchor wording variations would compare. To examine whether the significant ratings associated with the evaluative anchors were artifactual in nature, the *Very Good* response scale anchor was replaced with an *Excellent* anchor. The change in anchor wording was made because the researchers believed that *Excellent* was more semantically equivalent with *Strongly Agree*, than *Very Good.* To ensure that participants were aware of the instructions, the researcher read the instructions aloud, requesting either an opinion or an evaluation for one liked and one disliked class.

Due to an error in data collection, the intended methodological design was modified and took the form of four separate 2 (course) x 2 (instructions) x 3 (response anchor wording) mixed randomized between subjects ANOVAs (one for each dimension of teaching effectiveness being assessed). Consistent with the original study, no main effect for instructions was found. As expected there was a main effect for course across all conditions. There were no significant three-way interactions, although various two-way interactions appeared across the different conditions.

A significant main effect for response anchors appeared for only one dimension (course organization and design). The two-way interaction between response anchor wording and course revealed significant mean differences, only under the *disliked* course condition. For rapport with students, scale means from all three response anchor wording conditions differed significantly from one another. While in course organization and design, and course value, agreement response anchors produced significantly different means from the evaluation and semantic differential response anchors. The two-way interaction between instructions and response anchors, for the fairness of grading factor,

presented the most curious results. With *evaluation* instructions, the evaluation anchor scale mean differed significantly from both the agreement and semantic differential scale means. However, with *opinion* instructions, a significant difference appeared between the semantic differential scale mean and both the evaluative and agreement scale means.

Ing and Jackson (2007) continued with an additional study to implement the methodological design they had intended to use in their previous study. The researchers collected data from 246 undergraduate students at the University of Windsor. The questionnaire format was altered to present all of the items in one SETE form. The order for rating courses (liked and disliked) was counterbalanced and labeled directly on each questionnaire, to ensure that participants were assessing appropriate courses. Instructions were once again read aloud to participants prior to commencing and the three response anchor wording variations remained consistent with the previous study. A 2 (course) x 4 (dimensions of teaching effectiveness) x 2 (instructions) x 3 (response anchor wording) mixed randomized by repeated measures ANOVA was conducted on the collected data.

There was a main effect for response anchor wording, where the semantic differential anchors resulted in significantly different mean responses from those elicited through agreement and evaluative response anchors. Significant two-way interactions included SETE dimension by response anchor wording and course by SETE dimension. With respect to the former, the semantic differential response anchors elicited in the most positive ratings in three of the four SETE dimensions. While with the latter, instructor ratings were especially negative for one dimension (course value), for the *disliked* course condition. The only significant three-way interaction appeared as *SETE dimension* by *response anchor wording* by *course*. The semantic differential response anchors elicited

the most negative ratings in the fairness of grading dimension for *liked* classes, and also displayed a response pattern differing from the evaluative and agreement anchors, for *disliked* classes. Additionally, the agreement anchors displayed differing response patterns for *liked* versus *disliked* courses. The evaluative response anchors elicited a similar response pattern to that of the agreement worded anchors.

The appearance of a significant effect for response anchor wording suggests that not all scales are created equal (Ing & Jackson, 2007). This result highlights the importance of scale consideration when selecting measures for research, assessment, and evaluation in psychology. The question items included in the study were thought to be parallel across scale formats, with the only differences appearing in the anchors. The differing response patterns, elicited through the three response anchor wording variations, draw attention to the impact of changing just a few words on a questionnaire (Ing & Jackson, 2007).

Taken together, these three studies suggest that: 1) the semantic differential response anchors elicit significantly different response patterns than the other two types of anchors (i.e., agreement and evaluative); 2) the type of course being assessed (i.e., liked versus disliked) may act as a moderating variable for the response patterns elicited through the different response anchor wording; 3) seemingly harmless variations in scale properties (i.e., response anchors, scale length, instruction wording) can result in significant main and interaction effects on elicited response patterns; 4) SETE scale construction should result from careful analyses of the implications associated with each selection decision.

Purpose

At this point, there appear to be several questions regarding scale construction that need to be systematically investigated. Most recently, Jackson et al. (2007) initiated an effort to establish general guidelines for questionnaire construction. One of the questions raised by the researchers is whether the choice of response anchors increases or decreases in importance for longer versus shorter scales. The purpose of this study is to investigate the effect of varying three questionnaire components on students' ratings of instructor performance. More specifically, how do variations in instructions, response anchors, and number of scale points affect student evaluations of teaching effectiveness?

The following hypotheses are based upon the literature reviewed, with special influences from the research (described above) conducted by Arseneault & Jackson (2005) and Ing & Jackson (2007, 2006). First, it is hypothesized that shorter response scales will be more sensitive to variations in instructions and response anchor wording. Therefore, a three-way interaction effect is anticipated between *instructions*, *response anchors*, and *length of response scale*. It is expected that with *opinion* instructions, significantly different ratings will be elicited through use of *agreement* anchors and *evaluative* anchors, for *five-point* response scales. The mean differences between ratings elicited through *evaluative* instructions and *no instructions*, and measured on *nine-point* response scales are expected to be less pronounced. Ratings of instructor performance are also likely to be influenced by whether the course taken was *liked* or *disliked*. Therefore, it is also hypothesized that type of *course* (liked versus disliked) will act as a moderating variable for the anticipated interaction effects. Moderators are variables that enhance or reduce the strength of causal relationships between independent and dependent variables (Baron

& Kenny, 1986). Thus, it is expected that the three-way interaction effects, involving the independent variables (instructions, anchor wording, and scale length), will be *more pronounced* for the *disliked* course condition.

CHAPTER III

Method

Participants

Participants consisted of 431 students enrolled in undergraduate psychology courses at the University of Windsor. Students registered for participation in the study through the Psychology Department's online Participant Pool. Available appointments for this study were organized in groups of 16 students per timeslot. Participants were requested to have completed at least one full semester of courses at the University of Windsor or at another Canadian University, to ensure previous exposure to SETE forms. Responses from 26 participants were omitted from the analyses due to various problems with their data (described below in the results section). After removing the troublesome cases, the sample size consisted of 405 participants, with 79.3% female and 20.7% male students. The majority of the participants were in their second (30.1%), third (29.9%), or fourth (21.0%) year of study at the University of Windsor (first = 12.1% and other = 6.2%).

The largest percentages for participant ages were identified as 19 (24.2%), 20 (22.7%), and 21 (17.5), with the remaining percentages falling below 9% (17 = .7%, 18 = 7.2%, 22 = 8.1 %, 23 = 6.2%, 24 = 2.7%, 25 = 2%, and 26 or older = 8.6%). Although students were told that they required previous experiences with completing SETE forms, to participate in this study, 2.7% of participants indicated on their surveys that they had no such previous experience, while the remaining 97.2% had completed a minimum of one to five course evaluations (1-5 = 15.6%, 6-10 = 21.5%, 11-15 = 13.3%, 16-20 = 15.6%, 21-25 = 10.4%, 26-30 = 7.2%, 31-35 = 6.7%, 36-40 = 4.7%, and 40+ = 2.5%).

With respect to participants' programs of study, 58.3% were Social Science students, 15.1% were Science students, and the rest of the students were evenly distributed amongst other academic programs (Arts = 7.7%, Engineering = 1.7%, Human Kinetics = 5.9%, Nursing = 3.7%, Business = 4%, and Education = 3.7%).

Measures

The SETE form was comprised of two identical halves: one to assess a liked course; one to assess a disliked course. The following statement appeared at the beginning of each SETE form:

> Please answer all of the following questions, as honestly as possible, on the bubble sheet provided.

The *Liked Class* and *Disliked Class* sections were clearly marked within the questionnaire. The order of presentation for liked versus disliked course sections was counterbalanced across all experimental conditions, in order to control for any potential order effects.

The instructions variable was comprised of three levels, with the first being no instructions (described below). The second level of instructions requested students' opinions and was worded in the following manner:

> You are being asked to fill out a questionnaire about **two classes** that you have taken at the University of Windsor: one that you **liked** and one that you **disliked**. You are being asked to give your personal **opinion** about the class and professor characteristics for each course, so feel free to be as **subjective** as you wish.

For the third level of instructions, students were asked to give an evaluation:

You are being asked to fill out a questionnaire about **two classes** that you have taken at the University of Windsor: one that you **liked** and one that you **disliked.** You are being asked to give an **evaluation** of the class and professor characteristics for each course. Please be as **objective** as you can.

Questions for this study were derived from the Students' Perceptions of Teaching Effectiveness scale, second edition (SPTE II), a 58-item measure that is employed to collect data on instructor effectiveness, from students. The SPTE was developed at Wichita State University in 1975 by the Liberal Arts and Sciences Teaching Improvement Committee (Jackson et al., 1999). It consists of student and professor demographic questions and 39 items designed to provide an evaluation of the particular professor being rated. Each item is measured on a five-point semantic differential response scale (i.e., each item has a different set of opposite response choice anchors). According to Jackson et al. (1999), these items load on one of six factors: Rapport with Students, Course Value, Course Organization and Design, Fairness of Grading, Course Difficulty, and Workload.

As an extension of Ing and Jackson's (2007) study, this investigation included the four previously examined dimensions of teaching effectiveness: rapport with students; course value; course organization and design; and fairness of grading. The selected items were manipulated to represent agreement (opinion-eliciting) and evaluative (evaluation-eliciting) response scales. The two scales were anchored as *Strongly Agree – Strongly Disagree* (agreement) and *Excellent – Very Poor* (evaluative). As a result, the parallel items in each of these two conditions were not worded identically, but were thought to be

equivalent. The selected items also appeared with two different scale lengths: five points and nine points.

Validity checks were included throughout the questionnaire, to ensure that participants had read each question prior to responding. Since students were asked to record their responses on scantron/bubble sheets, the validity checks included questions in which respondents were asked to *'leave this line blank'* or to *'fill in any two bubbles'*. Participants were also asked to record whether they were assessing a *liked* or *disliked* course. The questionnaire was comprised of two sections: one for a liked class and one for a disliked class. The order of presentation for liked and disliked classes were counterbalanced in order to control for order effects.

The SETE form that was implemented in this study was comprised of two identical sections, each containing 22 items designed to assess the four dimensions of teaching effectiveness. The first factor, *Rapport with Students,* captures the participants' perceptions of the instructor's ability to establish a classroom rapport that is conducive to learning. A representative rapport item is "From my own experience, the instructor came across as a person as well as a teacher very well ...*Strongly Agree/Strongly Disagree*" or "From my own experience, the instructor's ability to come across as a person as well as a teacher was ...*Excellent/Very Poor*".

The second factor, *Course Value*, encapsulates students' assessments of the value of a course as reflected in the following facets: knowledge gained; expected retention; enthusiasm for attending class; recommendation of the course to others; and further interest in the subject matter, resulting from taking the course. A representative item from this factor is "As a result of this course, my knowledge level in this area has greatly

increased...*Strongly Agree/Strongly Disagree*" or "As a result of this course, the increase in my knowledge level in this area has been...*Excellent/Very Poor*".

The third factor, *Course Organization and Design*, encompasses instructor skills and competencies, such as: organization; preparation; clarity and suitability of presentation in conveying course concepts and objectives; and answering questions. An example of an item for this dimension is "The degree to which the material covered in this course was interrelated and consistent with subject area was excellent...*Strongly Agree/Strongly Disagree*" or "The degree to which the material covered in this course was interrelated and consistent with subject area was...*Excellent/Very Poor*".

The fourth factor, *Grading Fairness*, identifies participant evaluations of the quantity, clarity, and validity (appropriateness) of the methods employed in determining final grades, against student perceptions of personal performance in the classes assessed. A sample grading item is "The examination questions, or other evaluative methods used by the instructor, seemed to be very clear and fair ...*Strongly Agree/Strongly Disagree*" or "The clarity and fairness of the examination questions, or other evaluative methods used by the instructor, seemed to be ...*Excellent/Very Poor*". Although additional items were included in the SETE questionnaires (intended to measure two additional dimensions of teaching effectiveness), these items were excluded from the analyses, in order to maintain consistency with the previous studies conducted by Arseneault & Jackson (2005) and Ing & Jackson (2007, 2006) (described above).

The demographic questions included: whether the class that was rated was liked or disliked by the participant; participant's degree of certainty regarding the accuracy of responses; participant gender; grade received in the course that was assessed; participant

year in university; age of participant; participant ethnic orientation; ethnicity of the instructor for the rated course; whether the participant had been taught by the instructor more than once; instructor gender; level of course that was assessed; and whether the course was a degree requirement for the participant.

Procedure

Participants registered for available time slots through the Psychology Department's online Participant Pool. All of the appointments were conducted in a classroom, in the Psychology Department, at the University of Windsor, in groups of 16. Upon arrival, participants were greeted and asked to select a seat where papers had been preset. Participants experienced one of three possible instructions conditions: (1) *No instructions,* in which students were told that they could begin and asked to bring their sheets to the researcher once they were completed; (2) *Opinion*, in which the researcher read the opinion instructions (described above) aloud to students before informing them that they could begin and asked to bring their sheets to the researcher once they were completed; and (3) *Evaluation*, in which the researcher read the evaluation instructions (detailed above) aloud to students prior to informing them that they could begin and asked to bring their sheets to the researched once they were completed. After handing in their sheets to the researcher, students were thanked for their participation.

Methodology

An experimental design facilitated this investigation, with the inclusion of one within-subjects variable and three between-subjects variables (experimental manipulations described below). Each participant received one survey, comprised of two sections: one section containing questions regarding a class that the student *liked*, the

other section containing questions regarding a class that the student *disliked.* All of the participants received survey items intended to measure each of the four dimensions of teaching effectiveness. Half of the participants received agreement response scale anchors, labeled as *Strongly Agree* to *Strongly Disagree* (opinion-eliciting); while the other half of the participants received evaluative response scale anchors, worded *Excellent* to *Very Poor* (evaluation-eliciting). One third of the participants received instructions requesting their *subjective opinions* of instructor and class characteristics; one third of the participants received instructions requesting an *objective evaluation* of instructor and class characteristics; and one third of participants received *no instructions* at all.

CHAPTER IV

Results

Data Analysis

Student evaluations of teaching effectiveness (SETE) questionnaires were completed by 431 undergraduate students at the University of Windsor. Responses from nine participants were excluded from the analyses because these respondents had violated three or more of the six validity checks included in the survey. An additional 13 cases were excluded from the analyses when their scale values were identified at outliers; defined as having z scores with absolute values greater than 2.5 standard deviations away from the mean (Kirk, 1995). Three further cases were omitted for extreme missing values and evidence of not reading the questions prior to responding (i.e., responding to "gender" with an invalid option). Therefore, the final sample size for this study consisted of 405 participants, with the number of participants in each of the 12 experimental conditions ranging from 30 to 36.

The assumption of normality was assessed through the use of histograms and z scores and as mentioned above, values identified as outliers were omitted from the analyses. The homogeneity of variance assumption was assessed with the use of Levene's tests, where violations appeared in all but one condition (course organization and design for disliked course). Across most conditions, the ratio of the largest variance to the smallest variance was less than four (ranging from 2.55 to 3.6) with exceptions appearing in only three conditions: fairness of grading for disliked course, 4.30; course organization and design for liked course, 4.48; and rapport for liked course, 6.63. Because the cell sizes were roughly equal across conditions (30-36 per cell) and all larger

than n=20, and the data were normally distributed, the repeated measures designs (described below) were considered to be robust to the violation of the homogeneity of variance assumption (Stevens, 2002). The assumption of the equivalence of variance-covariance matrices was violated across all four repeated measures analyses of variance. However, for each analysis, the determinants of the cell matrices were compared and the largest value was always less than four times greater than the smallest value. Since the group sizes were roughly equal with $n > 20$, the analyses were viewed as robust to the violation of this assumption (Stevens, 2002).

The data were analyzed through the use of four (one for each teaching effectiveness dimension) separate *2* (course) x *3* (instructions) x *2* (response anchor wording) x *2* (length of response scale) mixed-randomized by repeated measures analyses of variance. The *within-subjects* variable was *course*, with the levels of (1) liked and (2) disliked. The *between-subjects* variables included: *instructions*, evenly divided between (1) opinion, (2) evaluation, and (3) no instructions; *response anchor wording*, alternating between (1) agreement (opinion-eliciting) and (2) evaluative (evaluation-eliciting); and *length of response scale*, separated into (1) five-point and (2) nine-point scales. Responses measured with five-point scales were converted to values on a nine-point scale, in order to facilitate comparisons between the two levels of this variable. The item response values from the five-point scale (i.e., 0 - 4) were recoded into their numerical equivalents on a nine-point scale (i.e., 0 - 8). The statistical analyses for this study were conducted through the Statistical Package for Social Sciences (SPSS). All analyses were conducted with an alpha level of .05.

Means were computed for each of the four teaching effectiveness dimension scales. The scale means, standard deviations, and reliability estimates are presented in Table 1. One item was omitted from the *fairness of grading* scale, due to low correlations with the other scale items. Removal of this item increased the internal consistency reliability between the remaining scale items. The internal consistency reliability estimates for the four teaching effectiveness dimensions were good (greater than .70) for items measured on the nine-point response scale and for items collapsed across the five-point and nine-point response scales (Kaplan & Succuzzo, 2005). With the five-point response scale, some of the Cronbach's α values were less than .70, with the lowest estimate at .63. A complete list of items included in each teaching effectiveness dimension scale can be found in Appendix C.

Significant Between-Subjects Effects

A main effect for *instructions* appeared for only one dimension of teaching effectiveness, *rapport with students* ($F_{[2,393]}=3.72$, $p < .05$, partial $\omega^2=.00$), with a very small effect size (Cohen, 1977). A Bonferroni post-hoc test revealed that the mean difference (.03) between the "no instructions" and "opinion" conditions was significant at an alpha level of .05 ($p <.05$).

Across all four dimensions of teaching effectiveness, there was a significant effect for *length of response scale*: Rapport with students ($F_{[1,393]}=134.35$, $p < .05$, partial $\omega^2=.14$); Course value ($F_{[1,393]}=103.29$, $p < .05$, partial $\omega^2=.11$); Course organization and design ($F_{[1,393]}=158.43$, $p < .05$, partial $\omega^2=.16$); and Fairness of grading ($F_{[1,393]}=93.74$, $p < .05$, partial $\omega^2=.13$).

Table 1

Means, Standard Deviations, and Reliability Estimates by Dimension

	Teaching Effectiveness Dimension							
	Rapport with Students		***Course Value***		***Course Organization and Design***		***Fairness of Grading***	
Five-Point Scale	*Liked Course*	*Disliked Course*	*Liked Course*	*Disliked Course*	*Liked Course*	*Disliked Course*	*Liked Course*	*Disliked Course*
N	203	203	203	203	203	203	203	203
M	2.38	5.55	2.75	6.33	2.34	5.01	2.90	5.40
SD	.68	1.23	.85	1.22	.57	1.09	.90	1.42
α	.68	.77	.65	.72	.66	.71	.67	.63
Nine-Point Scale								
N	202	202	202	202	202	202	202	202
M	1.05	5.08	1.29	6.01	1.01	4.40	1.64	5.02
SD	.91	1.59	1.00	1.58	.89	1.45	1.14	1.70
α	.83	.84	.76	.79	.88	.79	.74	.74
Combined Scale								
N	405	405	405	405	405	405	405	405
M	1.72	5.31	2.02	6.17	1.67	4.71	2.33	5.22
SD	1.04	1.44	1.18	1.42	1.00	1.32	1.27	1.39
α	.78	.82	.82	.77	.90	.77	.78	.78
Corrected Item-Total Correlation	.48-.68	.46-.67	.62-.69	.49-.64	.67-.75	.44-.59	.58-.64	.48-.56

Combined Scale refers to values collapsed across five-point and nine-point scale conditions.

In order to achieve more accurate estimates of effect size, partial ω^2 values were calculated for each effect (SPSS provides partial η^2 values). Partial ω^2 provides a considerably less biased estimate of effect size than partial η^2 (Howell, 2008). The effect sizes for *length of response scale* were strong across all four teaching effectiveness dimensions, with values ranging from .11 to .16 (Cohen, 1977). Scale means for each dimension of teaching effectiveness by length of response scale are presented in Table 2. The effect for length of response scale is observable through relatively less favourable ratings provided by participants when they completed course evaluations on five-point response scales. This effect suggests that when respondents are presented with longer response scales (i.e., nine-point versus five-point scales), they are likely to provide more positive ratings for both liked and disliked courses.

Table 2

Scale Means for Teaching Effectiveness Dimensions by Length of Response Scale

		Teaching Effectiveness Dimension			
		Rapport with Students	*Course Value*	*Course Organization and Design*	*Fairness of Grading*
Length of Response Scale	*Five-point Scale*	3.96	4.54	3.67	4.20
	Nine-point Scale	3.07	3.65	2.71	3.34

Significant Within-Subjects Effects

There was a significant within-subjects effect for *course*, across all four dimensions of teaching effectiveness: Rapport with students, ($F_{[1,393]}$=1810.46, $p < .05$, partial ω^2=.69); Course value ($F_{[1,393]}$=2689.57, $p < .05$, partial ω^2=.77); Course

organization and design ($F_{[1,393]}$=1819.92, $p < .05$, partial ω^2=.69); and Fairness of grading ($F_{[1,393]}$=1335.25, $p < .05$, partial ω^2=.62). The effects sizes were large for each dimension of teaching effectiveness, with all values being greater than .14 (Cohen, 1977). Because participants were instructed to assess one liked and one disliked course, and the resulting significant differences are in the expected directions (more positive ratings for a liked course and more negative ratings for a disliked course), this effect simply validates the assumption that students provided ratings for both liked and disliked courses. The scale means for liked and disliked courses, for each dimension of teaching effectiveness, are presented in Table 3.

Table 3

Scale Means for Teaching Effectiveness Dimensions by Course

		Dimension of Teaching Effectiveness			
		Rapport with Students	*Course Value*	*Course Organization and Design*	*Fairness of Grading*
Course	*Liked*	1.71	2.02	1.67	2.32
	Disliked	5.32	6.18	4.71	5.22

Across all four teaching effectiveness dimensions, there was a significant interaction between *length of response scale* and *course*: Rapport with students ($F_{[1,393]}$=26.22, $p < .05$, partial ω^2=.03); Course value ($F_{[1,393]}$=51.88, $p < .05$, partial ω^2=.06); Course organization and design ($F_{[1,393]}$=26.41, $p < .05$, partial ω^2=.04); and Fairness of grading ($F_{[1,393]}$=22.93, $p < .05$, partial ω^2=.03). The partial ω^2 values, ranging from .03 to .06, indicate small effect sizes for this interaction (Cohen, 1977).

This interaction effect was interpreted through the use of separate one-way between-subjects analyses of variance (for *liked* and *disliked* courses), instead of employing simple main effects analyses, because the homogeneity of variance assumption had been violated. The analyses revealed that for both liked and disliked courses, the five-point and nine-point response scales elicited significantly different ratings. In the liked course condition, the nine-point scale elicited significantly more favourable ratings than the five-point scale, with mean differences ranging from 1.24 to 1.46. For the disliked course condition, the five-point scale elicited significantly less positive ratings than the nine-point scale, with mean differences from .31 to .59. Thus, the mean differences between ratings measured on five-point scales versus nine-point scales are considerably higher for the *liked* course condition than for the *disliked* course condition. The scale means for each condition are presented below in Table 4.

Table 4

Scale Means for Teaching Dimensions by Length of Response Scale and Course

		Teaching Effectiveness Dimension			
		Rapport with Students	*Course Value*	*Course Organization and Design*	*Fairness of Grading*
Liked Course	*Five-point Response Scale*	2.38	2.75	2.33	2.94
	Nine-point Response Scale	1.05	1.29	1.01	1.70
Disliked Course	*Five-point Response Scale*	5.55	6.33	5.00	5.46
	Nine-point Response Scale	5.09	6.02	4.41	4.97

Rapport with Students

Scale means and standard deviations for each experimental condition, along with the analysis of variance source table for this dimension, are presented in Appendix D. As described above, there were main effects for instructions ($F_{[2,393]}$=3.72, $p < .05$, partial ω^2=.00), length of response scale ($F_{[1,393]}$=134.35, $p < .05$, partial ω^2=.14), and course ($F_{[1,393]}$=1810.46, $p < .05$, partial ω^2=.69), and an interaction effect between course and length of response scale ($F_{[1,393]}$=26.22, $p < .05$, partial ω^2=.03). There was no main effect for response anchor wording. There were no other interaction effects present between the remaining variables included in the analysis.

Course Value

Means and standard deviations by condition, and the analysis of variance source table for *course value*, are displayed in Appendix E. As was reported above, there were main effects for course ($F_{[1,393]}$=2689.57, $p < .05$, partial ω^2=.77) and length of response scale ($F_{[1,393]}$=103.29, $p < .05$, partial ω^2=.11), along with an interaction effect between course and length of response scale ($F_{[1,393]}$=51.88, $p < .05$, partial ω^2=.06). There were no main effects for instructions or response anchor wording. No interactions effects were present between the other variables included in the analysis.

Course Organization and Design

The means and standard deviations for each experimental condition, along with the analysis of variance source table for this dimension, can be found Appendix F. The main effects for length of response scale ($F_{[1,393]}$=158.43, $p < .05$, partial ω^2=.16) and course ($F_{[1,393]}$=1819.92, $p < .05$, partial ω^2=.69) and the interaction between length of response scale and course ($F_{[1,393]}$=26.41, $p < .05$, partial ω^2=.04) were described above.

There were no main effects for instructions or response anchor wording. There were no additional interactions appearing between the remaining variables included in the analysis.

Fairness of Grading

Means and standard deviations by condition, and the analysis of variance source table for *fairness of grading* can be found in Appendix G. As described above, there were main effects for length of response scale ($F_{[1,393]}$=93.74, $p < .05$, partial ω^2=.13) and course ($F_{[1,393]}$=1335.25, $p < .05$, partial ω^2=.62), with an interaction effect between length of response scale and course ($F_{[1,393]}$=22.93, $p < .05$, partial ω^2=.03). There were no main effects for instructions or response anchor wording. No other interactions were present between the other variables included in the analysis.

CHAPTER V

Discussion

As was expected, there was a main effect for *course*, in each of the separate analyses, with large effect sizes across all four teaching effectiveness dimensions. The presence of this effect provides confirmation that participants were in fact observing the instructions to rate both a *liked* and a *disliked* course. The differences in scale means appeared in the expected directions, with the liked course scale receiving more positive ratings and the disliked course scale eliciting more negative ratings.

Across all dimensions, there was a significant main effect for *length of response scale*. This effect is observable through the mean differences between ratings measured on five-point response scales versus those measured on nine-point scales, where more positive ratings resulted from the implementation of the nine-point response scales. This effect implies that more favourable evaluations may be elicited through the use of longer response scales. Such an effect could be influential in SETE questionnaire design and should be examined in more detail to determine its broader implications.

For instance, if longer response scales elicit more positive ratings than shorter scales, it is easy to imagine course instructors wanting to be evaluated on SETE forms with the longest response scale possible. Alternatively, university administrators incorporating SETE ratings into tenure and promotion decisions may prefer to implement shorter response scales, to control for the possibility of artificially inflated ratings being included as relevant decision making criteria. Outside of academic settings, this effect could be relevant for the design and implementation of any questionnaires which measure

respondents' ratings of others. In any case, further investigations into the effects of length of response scale on instructor ratings are necessary.

An interaction effect between *length of response scale* and *course* appeared across all four analyses. This effect is observable through the significantly different means for each condition (presented above in Table 4). For the *liked* course condition, the ratings measured on a nine-point response scale were significantly more favourable than those measured on a five-point scale. In the *disliked* course condition, the five-point scale elicited significantly less favourable ratings than the nine-point response scale. For the *liked* course condition, the mean differences between the ratings measured on the five-point scale versus the nine-point scale were considerably greater than the mean differences found between the ratings from five-point versus nine-point response scales, for the *disliked* course condition.

This interaction highlights the differential influences of *length of response scale* for *liked* versus *disliked* courses. It is unclear as to why these two variables interact and result in significant mean differences across conditions. The larger mean differences for the *liked* course condition suggest that respondents receiving longer response scales are more likely to endorse the most positive response options, whereas respondents receiving shorter response scales are more likely to select response options located further away from the extreme positive anchor. The smaller mean differences for the *disliked* course condition may be attributable to a lack of distinction between disliked courses. In other words, when respondents are rating disliked courses, they tend to endorse response options that fall somewhere between the middle point, and the extreme negative anchor, regardless of response scale length. These differential approaches to completing course

evaluations may be elicited through the presentation of shorter versus longer response scales, thus resulting in significantly different mean differences across course conditions. A thorough investigation of this interaction effect is required, in order to determine its cause and practical implications.

A main effect for *instructions* appeared for only one teaching effectiveness dimension, *rapport with students*. A bonferroni post-hoc test revealed a significant difference between the "no instructions" and "opinion" conditions ($p < .05$). This meant that collapsing across other variables (i.e., response anchors, length of response scale, and course), participants receiving the "opinion" instructions provided significantly different ratings from those in the "no instructions" condition. According to the means for the instructions conditions, "opinion" instructions elicit the most positive ratings with "evaluation" instructions in the middle and "no instructions" resulting in the most negative ratings. These differences, however, were very small, with an effect size of $\omega^2=.00$. This effect for instructions is curious in that it was not present for the other three dimensions of teaching effectiveness and the effect size was equal to zero. The main effects for *course* and *length of response scale* were present across all four teaching effectiveness dimensions, with effect sizes ranging from .11 to .77. Given the effect size of zero, the absence of a significant effect for instructions for the other three dimensions, the presence of stronger main effects appearing across all four dimensions, and the lack of correction for type I error, it seems possible that this effect is spurious in nature. In other words, this effect would not be hypothesized in a replication of this study.

It was hypothesized that the five-point response scale would be more sensitive to variations in instructions and response anchor wording, thus resulting in a three-way

interaction between instructions, response anchor wording, and length of response scale. Results from the analyses did not support this hypothesis. In fact, there were no main or interaction effects involving response anchor wording. Also, as previously mentioned, the only significant effect involving the instructions variable was a main effect, observed for only one teaching effectiveness dimension. Furthermore, the only significant interaction effect, across all of the analyses, was a two-way interaction between length of response scale and course (present for all dimensions).

It was also hypothesized that the course variable (liked versus disliked) would act as a moderator for the anticipated interaction between the instructions, response anchor wording, and length of response scale variables. In the absence of the anticipated significant three-way interaction, no further analyses were warranted to investigate the hypothesized moderator variable. Therefore, no evidence was found to support this hypothesis.

The complete absence of support for the two hypotheses included in this study is surprising, when one considers the significant effects that were present in previous studies involving instructor ratings and variations in instructions and response anchor wording (Arseneault & Jackson, 2005; Ing & Jackson, 2006, 2007). Results from this study suggest that variations in wording of instructions and response anchors are not influential on respondents' ratings of teaching effectiveness. With respect to variations in instruction wording, the absence of effect for instructions (in all but one condition) provides evidence that students are not influenced by requests for opinions (subjectivity) or evaluations (objectivity), which suggests that they employ the same degree of objectivity and subjectivity into their instructor ratings, regardless of instructions.

The absence of a main effect for instructions has remained consistent across four separate studies. Interactions involving instruction wording has been less consistent, appearing in only two of the four studies. At this point, it may be necessary to reconsider the potential effects of varying instructions on SETE ratings. In other words, the limited and inconsistent interaction effects appearing randomly for different teaching effectiveness dimensions and the complete absence of any main effects for instructions suggest that instruction variations do not significantly influence students' ratings of instructor performance. This suggestion may also be true for other evaluative situations, in which respondents are providing ratings of other individuals.

With regard to response anchor wording, the absence of significant effects, across all conditions and dimensions was most unexpected, given the presence of both main and interaction effects in previous studies (Arseneault & Jackson, 2005; Ing & Jackson, 2006, 2007). However, across the four studies involving response anchor wording variations, it appears that as the sample sizes (power) increase, the effects for response anchor wording (i.e., agreement versus evaluative anchors) decrease. For instance, while Ing and Jackson (2007) found a main effect for response anchor wording, the significant differences separated the response pattern elicited through semantic differential anchors from those elicited through agreement and evaluative anchors, which resulted in similar response patterns. Furthermore, results from this study (along with results from Ing and Jackson (2007) provide evidence that students are not influenced by requests for opinions (subjectivity) or evaluations (objectivity), which suggests that they employ similar degrees of objectivity and subjectivity into their instructor ratings, regardless of the instructions that they receive.

The objectives of the recent studies involving response anchor variations have been to identify and examine the effects of such variations, including the potential replication of the two-way and three-way interactions that were present in Arseneault and Jackson's (2005) study. However, results from this study, combined with the results from Ing & Jackson's (2007) study, provide evidence that agreement and evaluative anchors can be used interchangeably without influencing responses. This implied equivalency between the two response anchor variations is contradictory to the effects exhibited in the first two studies (Arseneault & Jackson, 2005; Ing & Jackson, 2006). It is important to consider that the two most recent studies have had increased power over the first two studies, through the use of longer response scales, substantially larger sample sizes, and improved statistical analysis design. At this point, it seems as though the previous interaction effects associated with response anchor wording may have been attributable to some unexamined factors, rather than the specific response anchor wording. Researchers interested in pursuing this line of inquiry, may benefit from investigating other (unexamined) potential sources of interactions involving response anchor wording.

While the hypotheses of this study were not supported, there is valuable information to be gained from this research. The practical implications resulting from this study are three-fold. First, administrators of SETE instruments do not need to be particularly concerned about the type of instructions that students receive prior to completing instructor/course evaluations. The type of instructions provided to students does not appear to significantly influence students' ratings of instructor performance.

Second, SETE instrument creators and administrators do not need to be highly concerned with the selection of agreement response anchors versus evaluative response

anchors. The analyses from this study and from Ing & Jackson (2007) did not result in any significant differences between the ratings elicited through agreement versus evaluative worded response anchors. These results suggest that variations in response anchor wording (agreement versus evaluative) are not responsible for significant differences in course/instructor ratings.

Third, SETE creators and administrators should be aware of the influence that length of response scale has on students' ratings of teaching effectiveness. The effect of length of response scale has been documented in the scale construction literature (see *Scale Construction* literature review above). Longer scales are better equipped to measure the true nuances of respondent opinions (i.e., direction, intensity, neutrality), in comparison with shorter scales, which are more restrictive. Better reliability has been associated with longer rather than shorter response scales. Results from this study suggest that longer response scales also elicit more favourable ratings than shorter response scales.

Although the hypotheses associated with this study were not supported, there remain aspects of response scale construction that are worthy of investigation. First, an extension of this study could include a reversal of the numerical values associated with the anchors, or a reversal of the placement of the response anchors. The response scales implemented in this study, as well as the studies conducted by Arseneault and Jackson (2005) and Ing and Jackson (2007; 2006), were arranged from left to right, with the most positive option appearing first and the most negative appearing last. However, the numerical values associated with the scales appeared in ascending order from left to right, resulting in the most positive option represented by zero and the most negative option

represented by 8 (or some other value higher than 0). This representation of positive options with lower numerical values and negative options with higher numerical values may cause some form of cognitive dissonance for respondents and subconsciously affect their ratings.

A second extension of this study could include replacing the *Excellent* anchor on the evaluative scale with the *Very Good* anchor that was implemented by Arseneault and Jackson (2005). This would enable the researcher to determine the equivalency between the two anchor wording variations, and to investigate whether the significant effects from the Arseneault and Jackson (2005) study were associated with the response anchor wording. The significant differences in instructor ratings, elicited through variations in response anchor wording became less pronounced (Ing & Jackson, 2007) and absent (current study) when the *Very Good* anchor was replaced with the *Excellent* anchor. This could be the result of some sort of semantic association with the word *Excellent* versus *Very Good* or the equivalency of those two anchors with the agreement anchor of *Strongly Agree*.

A third extension of this study could employ SETE forms in which numerical values are not included on the response scale (e.g., University of Windsor's course evaluation instrument). Removing the numerical references may affect the elicited ratings, if respondents are in fact relying on numerical representations to support their ratings. Relying solely on worded response anchors may result in more accurate ratings, if the inclusion of verbal anchors (words) and numerical anchors (numbers) are creating conflicting prompts/guides for respondents.

Finally, several extensions of this study could be implemented though the use of web-based SETE instruments. For instance, it would be interesting to compare the instructor ratings from web-based SETE instruments with traditional paper-based versions. As technology advances, more universities are certain to implement their instructor and course evaluations through the use of internet-based questionnaires. This transfer of instrument medium opens the door for research based upon the effects of technology on how students rate teaching effectiveness in particular, and how individuals rate others' performance in general.

REFERENCES

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales. Sociological Methods & Research, 25(3), 318-340.

Aubrecht, J. D., Hanna, G. S., & Hoyt, D. P. (1986). A comparison of high school student ratings of teaching effectiveness with teacher self-ratings: Factor analytic and multitrait-multimethod analyses. *Educational and Psychological Measurement*, 46, 223-231.

Arseneault, J. M. & Jackson, D. L. (2005). *Opinion versus evaluation: Do instructions and response choice anchors influence students' ratings?* Unpublished honours thesis, University of Windsor, Windsor, Ontario, Canada.

Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, 51, 220-232.

Bardo, J. W., Yeager, S. J., & Burdsal, C. A. (1985). Examination of response formats without anchoring items. *Perceptual and Motor Skills*, 61, 287-297.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variables distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Burdsal, C. A., & Bardo, J. W. (1986). Measuring student's perceptions of teaching; Dimensions of evaluation. *Educational and Psychological Measurement*, 46, 63-79.

Cadwell, J., & Jenkins, J. (1985). The effects of the semantic similarity of items on student ratings of instructors. *Journal of Educational Psychology*, 77, 383-393.

Cashin, W. E. (1995). Student ratings of teaching: The research revisited. *Center for Faculty Evaluation and Development. Idea Paper*, 32.

Chan, J. C. (1991). Response-order effects in likert-type scales. *Educational and Psychological Measurement,* 1991, 51, 531-540.

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education*, 9(4), 197-207.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481-489.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. Quality Assurance in Education, 11(1), 37-46.

Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement*, 16(1), 43-48.

Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluation: mode can matter. *Assessment & Evaluation in Higher Education*, 30(6), 581-592.

Gotay, C. C., Blaine, D., Haynes, S. N., Holup, J., & Pagano, I. S. (2002). Assessment of quality of life in a multicultural cancer patient population. *Psychological Assessment*, 14(4), 439-450.

Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, 45(3), 311-323.

Harrison, P. D., Ryan, J. M., & Moore, P. S. (1996). College students' self-insight and common implicit theories in ratings of teaching effectiveness. *Journal of Educational Psychology*, 88(4), 775-782.

Howell, D. C. (2008). *Fundamental Statistics for the Behavioural Sciences* (6th ed.). Belmont, CA: Thomson Wadsworth.

Hurd, M. C. (1999). Anchoring and acquiescence bias in measuring assets in household surveys. *Journal of Risk and Uncertainty, 19*, 111-136.

Ing, P. G., & Jackson, D. L. (2006). *Evaluation, opinion, and semantic differential: A reevaluation of the influence of instructions and response scale choices on students' ratings.* Unpublished honours thesis, University of Windsor, Windsor, Ontario, Canada.

Ing, P. G., & Jackson, D. L. (2007, April). *Varying instructions and response anchors for students' evaluations of teaching.* Paper presented at the meeting of the Society for Applied Multivariate Research, Fort Worth, TX.

Jackson, D. L., Ing, P. G., & Arseneault, J. M. (2007). *Opinion versus Evaluation: Do Instructions and Response Scale Anchors Influence Students' Ratings?* Manuscript submitted for publication.

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2007). *Reporting Practices in Confirmatory Factor Analysis Research: An Overview and Some Recommendations*. Manuscript submitted for publication.

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*, 580-596.

Jenkins, J. (1987). Implicit theories and semantic similarities: Reply to Marsh and Groves (1987). *Journal of Educational Psychology*, 79, 490-493.

Johanson, G. A., & Osborn, C. J. (2004). Acquiescence as differential person functioning. *Assessment & Evaluation in Higher Education*, 29(5), 535-548.

Kalton, G., & Schuman, H. (1982). The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society*, 145(1), 42-73.

Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology*, 15(3), 241-261.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological Testing: Principles, applications, and issues* (6th ed.). Belmont, CA: Wadsworth Thomson Learning.

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. Psychology & Marketing, 17(2), 79-103.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioural sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Knowles, E., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2), 379-386.

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

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2), 379-386.

Knowles, E. S., & Nathan, K. T., (1997). Acquiescent responding in self-reports: Cognitive style or social concern. *Journal of Research in Personality*, 31, 293-301.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A monte carlo approach. *Journal of Applied Psychology*, 60(1), 10-13.

Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, 74(2), 264-279.

Marsh, H. W. (1984). Students' evaluations of university teaching dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707-754.

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83(2), 285-296.

Marsh, H. W., & Groves, M. A. (1987). Students' evaluations of teaching effectiveness and implicit theories: A critique of Cadwell and Jenkins. *Journal of Educational Psychology*, 79, 483-489.

Marsh, H. W., Overall, J. V., & Kesler, S. P. (1979). Validity of student evaluation of Instructional effectiveness: A comparison of faculty self-evaluation and evaluations by their students. *Journal of Educational Psychology*, 71, 149-160.

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

Moore, S. & Kuol, N. (2005). Students evaluating teachers: exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education*, 10(1), 57-73.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill, Inc.

Ory, J. C., & Poggio, J. P. (1981). Response-mode variation effects on affective measures. *Educational and Psychological Measurement*, 41, 625-634.

Presser, S. & Schuman, H. (1980). The measurement of a middle position in attitude surveys. *Public Opinion Quarterly*, 70-85

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology*, 121, 81-96.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the likert and thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, 59(2), 211-233.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *The Annual Review of Sociology*, 29, 65-88.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item Reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, 41, 1101-1114.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York, NY: Academic Press Inc.

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

Si, S. X., & Cullen, J. B. (1998). Response categories and potential cultural bias: Effects of an explicit middle point in cross-cultural surveys. *The International Journal of Organizational Analysis*, 6(3), 218-230.

Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, 72(3), 438-443.

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

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 207-232.

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, 28(6), 587-607.

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educational and Psychological Measurement, 64(6), 956-972.

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research*, 15, 261-267.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education,* 12(1), 55-76.

Appendix A

The following questions were presented with either *five* or *nine* scale points, with endpoints (anchors) labeled as *Strongly Agree* and *Strongly Disagree*, in counterbalanced order (i.e., *Liked Class* appeared first for half of the cases and second for the other half).

Questions

(Liked/Disliked) Class

1. With respect to your progress in the course, the instructor was concerned and actively helpful.

2. In terms of what I gained (learned) from the course, the grade that I obtained was an excellent reflection.

3. By raising challenging questions or problems for discussion, the instructor stimulated students to think for themselves in nearly every class.

4. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were judged to be conducive to learning.

5. Based on your experience, the instructor's attitude toward students as individuals was respectful.

6. As reflected in the classroom and in the presentation of course material, the instructor was very enthusiastic.

7. Judging only on the basis of your own experience, the instructor's knowledge of the subject material of the course appeared to be excellent.

8. As a result of this course, my knowledge level in this area has greatly increased.

9. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor freely permitted comments.

10. The number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were sufficient to reflect achievement.

11. Leave this line blank.

12. I usually went to classes with eager anticipation.

13. The degree to which the material covered in this course was interrelated and consistent with the subject area was excellent.

14. As a result of this course, my interest in pursuing additional knowledge in this area was stimulated.

15. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was very useful and worthwhile.

16. The examination questions, or other evaluative methods used by the instructor, seemed to be very clear and fair.

17. Fill in any two bubbles on this line.

18. The method of assigning grades was clearly understood and consistent.

19. In conveying the concepts of this course in a clear, meaningful, and appropriate way, the instructor's ability was very evident.

20. The instructor's classroom presentation was well prepared at all times.

21. The ability of the instructor in handling questions and answering them to the student's satisfaction was quite satisfactory.

22. From my own experience, the instructor came across as a person as well as a teacher very well.

23. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was most appropriate.

24. The general objectives of the course were clearly understood.

25. With respect to the goals of the course, the amount of material presented was appropriate.

26. Considering the level of the course, class composition, prerequisites, etc., the level of the material presented was appropriate.

27. Considering other courses of similar credit and level, the workload for this course was appropriate.

28. As an aid to learning, the number and difficulty of assignments were appropriate.

29. Considering the nature of the course and subject material, the rate of coverage of the material was appropriate.

30. With respect to my ability and prior preparation, the level of difficulty of this course was appropriate.

31. Overall, the instructor was effective in teaching this course.

Demographics – Class:
32. The class you are assessing is one that you:
0) liked 1) disliked

33. How certain are you that you have accurately assessed this class?
0) 81-100% 1) 61-80% 2) 41-60% 3) 21-40% 4) 0-20%

34. The gender of the instructor for this course is:
0) Female B) Male

35. What grade did you receive in this course?
0) A 1) B 2) C 3) D 4) F

36. What level was this course?
0) 100-level 1) 200-level 2) 300-level 3) 400-level
4) other

(Disliked/Liked) Class

37. With respect to your progress in the course, the instructor was concerned and actively helpful.

38. In terms of what I gained (learned) from the course, the grade that I received was an excellent reflection.

39. By raising challenging questions or problems for discussion, the instructor stimulated students to think for themselves in nearly every class.

40. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were judged to be conducive to learning.

41. Based on your experience, the instructor's attitude toward students as individuals was respectful.

42. As reflected in the classroom and in the presentation of course material, the instructor was very enthusiastic.

43. Judging only on the basis of your own experience, the instructor's knowledge of the subject material of the course appeared to be excellent.

44. As a result of this course, my knowledge level in this area has greatly increased.

45. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor freely permitted comments.

46. The number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were sufficient to reflect achievement.

47. Leave this line blank.

48. I usually went to classes with eager anticipation.

49. The degree to which the material covered in this course was interrelated and consistent with the subject area was excellent.

50. As a result of this course, my interest in pursuing additional knowledge in this area was stimulated.

51. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was very useful and worthwhile.

52. The examination questions, or other evaluative methods used by the instructor, seemed to be very clear and fair.

53. Fill in any two bubbles on this line.

54. The method of assigning grades was clearly understood and consistent.

55. In conveying the concepts of this course in a clear, meaningful, and appropriate way, the instructor's ability was very evident.

56. The instructor's classroom presentation was well prepared at all times.

57. The ability of the instructor in handling questions and answering them to the student's satisfaction was quite satisfactory.

58. From my own experience, the instructor came across as a person as well as a teacher very well.

59. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was most appropriate.

60. The general objectives of the course were clearly understood.

61. With respect to the goals of the course, the amount of material presented was appropriate.

62. Considering the level of the course, class composition, prerequisites, etc., the level of the material presented was appropriate.

63. Considering other courses of similar credit and level, the workload for this course was appropriate.

64. As an aid to learning, the number and difficulty of assignments were appropriate.

65. Considering the nature of the course and subject material, the rate of coverage of the material was appropriate.

66. With respect to my ability and prior preparation, the level of difficulty of this course was appropriate.

67. Overall, the instructor was effective in teaching this course.

Demographics – Class:

68. The class you are assessing is one that you:
0) liked 1) disliked

69. How certain are you that you have accurately assessed this class?
0) 81-100% 1) 61-80% 2) 41-60% 3) 21-40% 4) 0-20%

70. The gender of the instructor for this course is:
0) Female B) Male

71. What grade did you receive in this course?
0) A 1) B 2) C 3) D 4) F

72. What level was this course?
0) 100-level 1) 200-level 2) 300-level 3) 400-level
4) other

Demographics – Student:

73. Your gender is:
0 Female 1) Male

74. Your age is:
0) 17 1) 18 2) 19 3) 20 4) 21 5) 22
6) 23 7) 24 8) 25 9) 26 or older

75. Your year in university is:
0) First 1) Second 2) Third 3) Fourth 4) Other

76. The approximate number of course evaluation forms that you have previously completed is:
0) none 1) 1-5 2) 6-10 3) 11-15 4) 16-20 5) 21-25
6) 26-30 7) 31-35 8) 36-40 9) 40 +

77. Your program of study is part of which faculty?

0) Arts	1) Social Science	2) Engineering	3) Human Kinetics
4) Nursing	5) Business	6) Business	7) Education

Appendix B

The following questions were presented with either *five* or *nine* scale points, with endpoints (anchors) labeled as *Excellent* and *Very Poor*, in counterbalanced order (i.e., *Liked Class* appeared first for half of the cases and second for the other half).

Questions

(Liked/Disliked) Class

1. The instructor's level of concern and active help with your progress in the course was:

2. In terms of what I gained (learned) from the course, the grade that I obtained provided a reflection that was:

3. By raising challenging questions or problems for discussion, the instructor's ability to simulate students to think for themselves in nearly every class was:

4. With regard to conduciveness of learning, the instructor's voice level, rate of speaking, appearance, mannerisms, and personal characteristics were:

5. Based on your own experience, the instructor's ability to display a respectful attitude towards students as individuals was:

6. As reflected in the classroom and in the presentation of course material, the instructor's enthusiasm was:

7. Judging only on the basis of your experience, the instructor's knowledge of the subject material of the course appeared to be:

8. As a result of this course, the increase in my knowledge level in this area has been:

9. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor's ability to freely permit comments was:

10. With regard to sufficiently reflecting achievement, the number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were:

11. Leave this line blank.

12. The degree to which I usually went to classes with eager anticipation was:

13. The degree to which the material covered in this course was interrelated and consistent with the subject area was:

14. The degree to which this course stimulated my interest in pursuing additional knowledge in this area was:

15. In one way or another (whether in relationship to my major, other courses, or just life in general) the degree to which this course has been very useful and worthwhile is:

16. The clarity and fairness of the examination questions, or other evaluative methods used by the instructor, seemed to be:

17. Fill in any two bubbles on this line.

18. The clarity of understanding and consistency of the method of assigning grades were:

19. In conveying the concepts of this course in a clear, meaningful, and appropriate way, evidence of the instructor's ability was:

20. The instructor's level of preparedness at all times was:

21. The ability of the instructor to handle questions and answer them to the student's satisfaction in a satisfactory manner was:

22. From my own experience, the instructor's ability to come across as a person as well as a teacher was:

23. Considering the nature of the course in terms of subject and class size, the appropriateness of the method of presentation of material (i.e., lecture, discussion groups, etc) was:

24. The degree to which the general objectives of the course were clearly understood was:

25. With respect to the goals of the course, the appropriateness of the amount of material presented was:

26. Considering the level of the course, class composition, prerequisites, etc., the appropriateness of the level of the material presented was:

27. Considering other courses of similar credit and level, the appropriateness of the workload for this course was:

28. As an aid to learning, the appropriateness of the number and difficulty of assignments was:

29. Considering the nature of the course and subject material, the appropriateness of the rate of coverage of the material was:

30. With respect to my ability and prior presentation, the appropriateness of the level of difficulty for this course was:

31. Overall, the instructor's effectiveness in teaching this course was:

Demographics – Class:
32. The class you are assessing is one that you:
0) liked 1) disliked

33. How certain are you that you have accurately assessed this class?
0) 81-100% 1) 61-80% 2) 41-60% 3) 21-40% 4) 0-20%

34. The gender of the instructor for this course is:
0) Female B) Male

35. What grade did you receive in this course?
0) A 1) B 2) C 3) D 4) F

36. What level was this course?
0) 100-level 1) 200-level 2) 300-level 3) 400-level
4) other

(Disliked/Liked) Class

37. The instructor's level of concern and active help with your progress in the course was:

38. In terms of what I gained (learned) from the course, the grade that I obtained provided a reflection that was:

39. By raising challenging questions or problems for discussion, the instructor's ability to simulate students to think for themselves in nearly every class was:

40. With regard to conduciveness of learning, the instructor's voice level, rate of speaking, appearance, mannerisms, and personal characteristics were:

41. Based on your own experience, the instructor's ability to display a respectful attitude towards students as individuals was:

42. As reflected in the classroom and in the presentation of course material, the instructor's enthusiasm was:

43. Judging only on the basis of your experience, the instructor's knowledge of the subject material of the course appeared to be:

44. As a result of this course, the increase in my knowledge level in this area has been:

45. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor's ability to freely permit comments was:

46. With regard to sufficiently reflecting achievement, the number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were:

47. Leave this line blank.

48. The degree to which I usually went to classes with eager anticipation was:

49. The degree to which the material covered in this course was interrelated and consistent with the subject area was:

50. The degree to which this course stimulated my interest in pursuing additional knowledge in this area was:

51. In one way or another (whether in relationship to my major, other courses, or just life in general) the degree to which this course has been very useful and worthwhile is:

52. The clarity and fairness of the examination questions, or other evaluative methods used by the instructor, seemed to be:

53. Fill in any two bubbles on this line.

54. The clarity of understanding and consistency of the method of assigning grades were:

55. In conveying the concepts of this course in a clear, meaningful, and appropriate way, evidence of the instructor's ability was:

56. The instructor's level of preparedness at all times was:

57. The ability of the instructor to handle questions and answer them to the student's satisfaction in a satisfactory manner was:

58. From my own experience, the instructor's ability to come across as a person as well as a teacher was:

59. Considering the nature of the course in terms of subject and class size, the appropriateness of the method of presentation of material (i.e., lecture, discussion groups, etc) was:

60. The degree to which the general objectives of the course were clearly understood was:

61. With respect to the goals of the course, the appropriateness of the amount of material presented was:

62. Considering the level of the course, class composition, prerequisites, etc., the appropriateness of the level of the material presented was:

63. Considering other courses of similar credit and level, the appropriateness of the workload for this course was:

64. As an aid to learning, the appropriateness of the number and difficulty of assignments was:
65. Considering the nature of the course and subject material, the appropriateness of the rate of coverage of the material was:

66. With respect to my ability and prior presentation, the appropriateness of the level of difficulty for this course was:

67. Overall, the instructor's effectiveness in teaching this course was:

Demographics – Class:
68. The class you are assessing is one that you:
0) liked 1) disliked

69. How certain are you that you have accurately assessed this class?
0) 81-100% 1) 61-80% 2) 41-60% 3) 21-40% 4) 0-20%

70. The gender of the instructor for this course is:
0) Female B) Male

71. What grade did you receive in this course?
0) A 1) B 2) C 3) D 4) F

72. What level was this course?
0) 100-level 1) 200-level 2) 300-level 3) 400-level
4) other

Demographics – Student:
73. Your gender is:
0) Female 1) Male

74. Your age is:
0) 17 1) 18 2) 19 3) 20 4) 21 5) 22
6) 23 7) 24 8) 25 9) 26 or older

75. Your year in university is:
0) First 1) Second 2) Third 3) Fourth 4) Other

76. The approximate number of course evaluation forms that you have previously completed is:

0) none	1) 1-5	2) 6-10	3) 11-15	4) 16-20	5) 21-25
6) 26-30	7) 31-35	8) 36-40	9) 40 +		

77. Your program of study is part of which faculty?

0) Arts	1) Social Science	2) Engineering	3) Human Kinetics
4) Nursing	5) Business	6) Business	7) Education

Appendix C

Rapport with Students

Questions with agreement anchors, worded as *Strongly Agree* and *Strongly Disagree*

1. & 37. With respect to your progress in the course, the instructor was concerned and actively helpful.

3. & 39. By raising challenging questions or problems for discussion, the instructor stimulated students to think for themselves in nearly every class.

4. & 40. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were judged to be conducive to learning.

5. &41. Based on your experience, the instructor's attitude toward students as individuals was respectful.

6. & 42. As reflected in the classroom and in the presentation of course material, the instructor was very enthusiastic.

9. & 45. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor freely permitted comments.

22. & 58. From my own experience, the instructor came across as a person as well as a teacher very well.

Questions with evaluative anchors, worded as *Excellent* and *Very Poor*

1. & 37. The instructor's level of concern and active help with your progress in the course was:

3. & 39. By raising challenging questions or problems for discussion, the instructor's ability to simulate students to think for themselves in nearly every class was:

4. & 40. With regard to conduciveness of learning, the instructor's voice level, rate of speaking, appearance, mannerisms, and personal characteristics were:

5. & 41. Based on your own experience, the instructor's ability to display a respectful attitude towards students as individuals was:

6. & 42. As reflected in the classroom and in the presentation of course material, the instructor's enthusiasm was:

9. & 45. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor's ability to freely permit comments was:

22. & 58. From my own experience, the instructor's ability to come across as a person as well as a teacher was:

Course Value

Questions with agreement anchors, worded as *Strongly Agree* and *Strongly Disagree*

8. & 44. As a result of this course, my knowledge level in this area has greatly increased.

12. & 48. I usually went to classes with eager anticipation.

14. & 50. As a result of this course, my interest in pursuing additional knowledge in this area was stimulated.

15. & 51. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was very useful and worthwhile.

Questions with evaluative anchors, worded as *Excellent* and *Very Poor*

8. & 44. As a result of this course, the increase in my knowledge level in this area has been:

12. & 48. The degree to which I usually went to classes with eager anticipation was:

14. & 50. The degree to which this course stimulated my interest in pursuing additional knowledge in this area was:

15. & 51. In one way or another (whether in relationship to my major, other courses, or just life in general) the degree to which this course has been very useful and worthwhile is:

Course Organization and Design

Questions with agreement anchors, worded as *Strongly Agree* and *Strongly Disagree*

7. & 43. Judging only on the basis of your own experience, the instructor's knowledge of the subject material of the course appeared to be excellent.

13. & 49. The degree to which the material covered in this course was interrelated and consistent with the subject area was excellent.

19. & 55. In conveying the concepts of this course in a clear, meaningful, and appropriate way, the instructor's ability was very evident.

20. & 56. The instructor's classroom presentation was well prepared at all times.

21. & 57. The ability of the instructor in handling questions and answering them to the student's satisfaction was quite satisfactory.

23. & 59. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was most appropriate.

24. & 60. The general objectives of the course were clearly understood.

Questions with evaluative anchors, worded as *Excellent* and *Very Poor*

7. & 43. Judging only on the basis of your experience, the instructor's knowledge of the subject material of the course appeared to be:

13. & 49. The degree to which the material covered in this course was interrelated and consistent with the subject area was:

19. & 55. In conveying the concepts of this course in a clear, meaningful, and appropriate way, evidence of the instructor's ability was:

20. & 56. The instructor's level of preparedness at all times was:

21. & 57. The ability of the instructor to handle questions and answer them to the student's satisfaction in a satisfactory manner was:

23. & 59. Considering the nature of the course in terms of subject and class size, the appropriateness of the method of presentation of material (i.e., lecture, discussion groups, etc) was:

24. & 60. The degree to which the general objectives of the course were clearly understood was:

Fairness of Grading

Questions with agreement anchors, worded as *Strongly Agree* and *Strongly Disagree*

*2. & 38. In terms of what I gained (learned) from the course, the grade that I obtained was an excellent reflection.

10. & 46. The number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were sufficient to reflect achievement.

16. & 52. The examination questions, or other evaluative methods used by the instructor, seemed to be very clear and fair.

18. & 54. The method of assigning grades was clearly understood and consistent.

Questions with evaluative anchors, worded as *Excellent* and *Very Poor*

*2. & 38. In terms of what I gained (learned) from the course, the grade that I obtained provided a reflection that was:

10. & 46. With regard to sufficiently reflecting achievement, the number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were:

16. & 52. The clarity and fairness of the examination questions, or other evaluative methods used by the instructor, seemed to be:

18. & 54. The clarity of understanding and consistency of the method of assigning grades were:

*Items omitted from analyses due to low correlations with other scale items and increased scale item-total correlations subsequent to removal.

Appendix D

Means and Standard Deviations by Condition for Rapport with Students

Course	Instructions	Response Scale Anchors	Response Scale Length	M	SD	n
Liked	None	Agreement	Five point	2.67	1.01	34
			Nine point	1.14	.76	32
	None	Evaluative	Five point	2.45	.62	36
			Nine point	.91	.81	35
	Opinion	Agreement	Five point	2.44	.71	34
			Nine point	.92	.98	34
	Opinion	Evaluative	Five point	2.22	.63	35
			Nine point	1.18	1.06	35
	Evaluation	Agreement	Five point	2.13	.36	31
			Nine point	1.09	1.05	30
	Evaluation	Evaluative	Five point	2.28	.57	34
			Nine point	1.09	.79	35
Disliked	None	Agreement	Five point	5.62	1.03	34
			Nine point	5.58	1.53	32
	None	Evaluative	Five point	5.73	1.19	36
			Nine point	4.88	1.56	35
	Opinion	Agreement	Five point	5.23	1.25	34
			Nine point	4.83	1.86	34
	Opinion	Evaluative	Five point	5.39	1.39	35
			Nine point	4.73	1.51	35
	Evaluation	Agreement	Five point	5.60	1.35	31
			Nine point	5.39	1.78	30
	Evaluation	Evaluative	Five point	5.66	1.25	34
			Nine point	5.18	1.16	35

Analysis of Variance Source Table for Rapport with Students

Source	**df**	**F**	***p***	**Partial ω^2**
		Between Subjects		
Instructions (I)	2	3.72	.03	.00
Response Anchors (R)	1	.89	.35	.00
Length (L)	1	134.35	.00	.14
I x R	2	1.47	.23	.00
I x L	2	1.12	.33	.00
R x L	1	1.79	.18	.00
I x R x L	2	.63	.53	.00
S within-group error	393	(MS=1.20)		
		Within Subjects		
Course (C) – liked vs. disliked	1	1810.46	.00	.69
C x I	2	2.45	.09	.00
C x R	1	.17	.68	.00
C x L	1	26.22	.00	.03
C x I x R	2	.07	.93	.00
C x I x L	2	.43	.65	.00
C x R x L	1	2.71	.10	.00
C x I x R x L	2	.40	.67	.00
C x S within-group error	393	(MS=1.45)		

Appendix E

Means and Standard Deviations by Condition for Course Value

Course	Instructions	Response Scale Anchors	Response Scale Length	M	SD	n
Liked	None	Agreement	Five point	2.87	.84	34
			Nine point	1.12	.97	32
	None	Evaluative	Five point	2.90	1.00	36
			Nine point	1.23	.83	35
	Opinion	Agreement	Five point	2.65	.95	34
			Nine point	1.31	1.18	34
	Opinion	Evaluative	Five point	2.85	.93	35
			Nine point	1.18	.87	35
	Evaluation	Agreement	Five point	2.61	.68	31
			Nine point	1.30	1.24	30
	Evaluation	Evaluative	Five point	2.56	.65	34
			Nine point	1.59	.87	35
Disliked	None	Agreement	Five point	6.18	1.10	34
			Nine point	6.59	1.28	32
	None	Evaluative	Five point	6.51	1.26	36
			Nine point	5.95	1.35	35
	Opinion	Agreement	Five point	6.39	1.17	34
			Nine point	5.92	1.76	34
	Opinion	Evaluative	Five point	6.02	1.54	35
			Nine point	5.66	1.76	35
	Evaluation	Agreement	Five point	6.53	1.19	31
			Nine point	5.87	1.88	30
	Evaluation	Evaluative	Five point	6.29	1.04	34
			Nine point	6.19	1.32	35

Analysis of Variance Source Table for Course Value

Source	**df**	**F**	***p***	**Partial ω^2**
		Between Subjects		
Instructions (I)	2	1.25	.29	.00
Response Anchors (R)	1	.10	.76	.00
Length (L)	1	103.29	.00	.21
I x R	2	.48	.62	.00
I x L	2	.66	.52	.00
R x L	1	.15	.70	.00
I x R x L	2	2.28	.10	.00
S within-group error	393	(MS=1.54)		
		Within Subjects		
Course (C) – liked vs. disliked	1	2689.57	.00	.77
C x I	2	1.19	.31	.00
C x R	1	1.92	.17	.00
C x L	1	51.88	.00	.06
C x I x R	2	.26	.77	.00
C x I x L	2	2.54	.08	.00
C x R x L	1	.17	.68	.00
C x I x R x L	2	2.10	.12	.00
C x S within-group error	393	(MS=1.30)		

Appendix F

Means and Standard Deviations by Condition for Course Organization and Design

Course	Instructions	Response Scale Anchors	Response Scale Length	M	SD	n
Liked	None	Agreement	Five point	2.47	.77	34
			Nine point	1.13	.87	32
	None	Evaluative	Five point	2.42	.60	36
			Nine point	.98	.87	35
	Opinion	Agreement	Five point	2.36	.56	34
			Nine point	.98	1.00	34
	Opinion	Evaluative	Five point	2.19	.55	35
			Nine point	.97	1.04	35
	Evaluation	Agreement	Five point	2.20	.46	31
			Nine point	.96	.81	30
	Evaluation	Evaluative	Five point	2.29	.54	34
			Nine point	1.05	.77	35
Disliked	None	Agreement	Five point	5.17	1.12	34
			Nine point	4.77	1.45	32
	None	Evaluative	Five point	5.23	.99	36
			Nine point	3.98	1.37	35
	Opinion	Agreement	Five point	4.74	1.03	34
			Nine point	4.29	1.57	34
	Opinion	Evaluative	Five point	4.83	1.08	35
			Nine point	4.26	1.23	35
	Evaluation	Agreement	Five point	4.86	1.23	31
			Nine point	4.46	1.58	30
	Evaluation	Evaluative	Five point	5.14	1.14	34
			Nine point	4.74	1.41	35

Analysis of Variance Source Table for Course Organization and Design

Source	**df**	**F**	***p***	**Partial ω^2**
		Between Subjects		
Instructions (I)	2	2.10	.12	.00
Response Anchors (R)	1	.10	.77	.00
Length (L)	1	158.43	.00	.16
I x R	2	2.49	.09	.00
I x L	2	1.18	.31	.00
R x L	1	1.52	.22	.00
I x R x L	2	.93	.40	.00
S within-group error	393	(MS=1.17)		
		Within Subjects		
Course (C) – liked vs. disliked	1	1819.92	.00	.69
C x I	2	1.23	.29	.00
C x R	1	.01	.94	.00
C x L	1	26.41	.00	.04
C x I x R	2	.96	.38	.00
C x I x L	2	.37	.69	.00
C x R x L	1	1.44	.23	.00
C x I x R x L	2	.59	.56	.00
C x S within-group error	393	(MS=1.02)		

Appendix G

Means and Standard Deviations for Fairness of Grading

Course	Instructions	Response Scale Anchors	Response Scale Length	M	SD	n
Liked	None	Agreement	Five point	3.05	.96	34
			Nine point	1.66	.96	32
	None	Evaluative	Five point	3.03	.95	36
			Nine point	1.59	1.12	35
	Opinion	Agreement	Five point	2.87	1.04	34
			Nine point	1.55	1.44	34
	Opinion	Evaluative	Five point	2.87	.81	35
			Nine point	1.75	1.08	35
	Evaluation	Agreement	Five point	2.59	.83	31
			Nine point	1.63	1.39	30
	Evaluation	Evaluative	Five point	2.87	.91	34
			Nine point	1.70	.84	35
Disliked	None	Agreement	Five point	5.27	1.53	34
			Nine point	5.33	1.65	32
	None	Evaluative	Five point	5.82	1.25	36
			Nine point	4.79	1.60	35
	Opinion	Agreement	Five point	4.88	1.54	34
			Nine point	4.70	1.94	34
	Opinion	Evaluative	Five point	5.59	1.20	35
			Nine point	5.39	1.62	35
	Evaluation	Agreement	Five point	5.40	1.74	31
			Nine point	4.74	1.70	30
	Evaluation	Evaluative	Five point	5.25	1.22	34
			Nine point	5.24	1.55	35

Analysis of Variance Source Table for Fairness of Grading

Source	**df**	**F**	***p***	**Partial ω^2**
		Between Subjects		
Instructions (I)	2	.53	.59	.00
Response Anchors (R)	1	3.42	.07	.00
Length (L)	1	93.74	.00	.13
I x R	2	1.23	.29	00
I x L	2	.92	.40	.00
R x L	1	.194	.66	.00
I x R x L	2	.76	.47	.00
S within-group error	393	(MS=1.71)		
		Within Subjects		
Course (C) – liked vs. disliked	1	1335.25	.00	.62
C x I	2	.19	.82	.00
C x R	1	1.06	.31	.00
C x L	1	22.93	.00	.03
C x I x R	2	.65	.53	.00
C x I x L	2	.07	.93	.00
C x R x L	1	.01	.93	.00
C x I x R x L	2	1.48	.23	.00
C x S within-group error	393	(MS=1.38)		

VITA AUCTORIS

NAME: Pamela Grace Ing PLACE OF BIRTH: Windsor, Ontario

EDUCATION

Psychology Department Fellow 2008 - 2009
Quantitative Psychology Doctoral Program
The Ohio State University
Columbus, Ohio

Master of Arts Degree in Applied Social Psychology 2006 - 2008
with Special Distinction
University of Windsor
Windsor, Ontario

Bachelor of Arts (Hons) in Psychology with Thesis 2003 - 2006
with Distinction
University of Windsor
Windsor, Ontario

Honours Psychology 1999 - 2000
University of Windsor
Windsor, Ontario

Walkerville Collegiate Institute June 1999
Ontario Scholar

ACADEMIC SCHOLARSHIPS AND HONOURS

Graduate School Fellowship 2008 – 2009
The Ohio State University

Graduate Tuition Scholarship 2006 - 2008
University of Windsor

Masters Level Scholarship 2006 - 2007
Social Sciences and Humanities Research Council Canada

Dean's Honour Roll 2005 - 2006
Faculty of Arts and Social Sciences
University of Windsor

Undergraduate Tuition Scholarship Winter 2006
University of Windsor

Entrance Scholarship 1999 - 2000
University of Windsor
PROFESSIONAL EXPERIENCES

Paper Presenter April 2008
Free For All: A Practical Introduction to the Mx Graphical User Interface
Kansas City, Missouri
Society for Applied Multivariate Research
Southwestern Psychological Association Annual Convention

Program Chair 2007 - 2008
Annual Meeting for the Society for Applied Multivariate Research
Southwestern Psychological Association Annual Conference
Kansas City, Missouri (April 2008)

Practicum Placement 2007 – 2008
Execution of survey development, web implementation, analyses, and instrument assessment
Office of Institutional Analysis
University of Windsor

Article Reviewer 2007 - 2008
Applied Multivariate Research
Journal of the Society for Applied Multivariate Research

Graduate Student Representative 2006 - 2008
Appointments Committee
Department of Psychology
University of Windsor

Second Author June 2007
Opinion versus Evaluation: Do Instructions and Response Scale Anchors Influence Students' Ratings?
First Author: Dennis L. Jackson
Manuscript submitted for publication

Poster Presenter June 2007
Varying Instructions and Response Anchors for Students' Evaluations of Teaching Effectiveness
2nd Author: Dennis L. Jackson
Ottawa, Canada
Canadian Psychological Association Annual Convention

Collaborator (with Community Psychology Graduate Class) June 2007
Learning through Praxis: The Benefits of Incorporating Social Action in Teaching
Symposium: Teaching of Psychology
Ottawa, Canada
Canadian Psychological Association Annual Convention

Paper Presenter April 2007
Varying Instructions and Response Anchors for Students' Evaluations of Teaching
2nd Author: Dennis L. Jackson
Fort Worth, Texas
Society for Applied Multivariate Research
Southwestern Psychological Association Annual Convention

Workshop Presenter March 2007
Balance: Support for Graduate Student Life
Presented to psychology graduate students at the University of Windsor

Workshop Presenter February 2007
Fundamentals of Statistics
Presented to psychology honours thesis students at the University of Windsor

Copy Editor 2005 - 2006
Applied Multivariate Research Journal

Student Affiliate 2006 – Present

- Psychometric Society
- Society for Applied Multivariate Research
- Southwestern Psychological Association
- Canadian Psychological Association

TEACHING EXPERIENCE

Graduate Assistant/Lab Instructor 2007 - 2008
Graduate Statistics for Psychological Research
Computer Laboratory

Graduate Assistant/Lab Instructor Spring 2007
Introductory Statistics for Social Sciences

Graduate Assistant 2006 - 2007
Advanced Multivariate Statistics (graduate level)
Graduate Statistics for Psychological Research
Community Psychology
Computer Laboratory