The Dissertation Committee for George Marcus Turner certifies that this is the approved version of the following dissertation:

The Role of Cognitive Schemas in a

Web-Based Student Evaluation of Teaching System:

Usability Issues of Design and Implementation

Committee:

Marilla Svinicki, Supervisor

Barbara Dodd

William Koch

Daniel Robinson

Paul Raffeld

The Role of Cognitive Schemas in a Web-Based Student Evaluation of Teaching System: Usability Issues of Design and Implementation

by

George Marcus Turner, B.S., M.Ed.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2004

The Role of Cognitive Schemas in a

Web-Based Student Evaluation of Teaching System:

Usability Issues of Design and Implementation

Publication No.

George Marcus Turner, Ph.D.
The University of Texas at Austin, 2004

Supervisor: Marilla Svinicki

With the increase in the use of web-based student evaluations of teaching (SETs) by institutions of higher education, the importance of understanding the usability of such web-based systems has increased. Such understanding is needed to ensure that web-based SETs produce the best information possible. By using schema theory from cognitive psychology as a backdrop, this study seeks to identify the impact that the method of response and grouping of items displayed per page have on the usability of web-based SETs. Issues of user satisfaction, error rate, and time on task are examined. Participants were divided into one of six similarity groups based on the degree to which participant expectation and the design of the web-based SET system matched. Analyzing data from 791 university students at a large southern University, it was determined that the use of a radio-button response format with SET items grouped by area produces the best results from a usability perspective. The use of drop-down boxes and text boxes is discouraged as a response format. Limitations of this study and suggested directions for future research are discussed.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Institutions of higher education are constantly seeking ways of evaluating the quality of instruction given in their courses. One of the most common and widespread ways in which this information is gathered is through the use of student evaluations of teaching (SETs). The information gained from SETs is frequently used during the review of faculty to aid in determining tenure, promotion, teaching awards, and pay increases. During the 1970s, 1980s, and extending into the 1990s, there was extensive research looking at the reliability and validity of SETs (e.g., Kerin, Peterson, & Martin, 1975; Feldman, 1976; Feldman, 1979; Marsh, 1984; Basow & Spielberg, 1987; Feldman, 1988; Marsh & Dunkin, 1992; Marsh & Roche, 1993). Although few firm conclusions have been reached from the research, it is safe to assume that SETs are here and they are here to stay.

Throughout the wide body of research there is considerable discussion of both the limitations and benefits of administering SETs using paper-and-pencil during class time. The lack of time for students to provide written comments, the logistical complications of paper based administration, and the bias that can be introduced by faculty before SETs are given are commonly mentioned as limitations (Layne, DeCristoforo, & McGinty, 1999; Dommeyer, Baum, Chapman, & Hanna, 2002). Many involved with SETs have proposed that a computerized administration of SETs would help eliminate many of these limitations, but not without a cost. Preliminary evidence in the literature suggests that data collected from on-line SETs are not directly comparable with the information collected from paper and pencil administrations. Low response rates, changes in the way

students respond to items, and a perceived lack of anonymity for students are often cited as potential flaws caused by moving SETs to an on-line format (Layne, DeCristoforo, & McGinty, 1999; Dommeyer, Baum, Chapman, & Hanna, 2002).

However, as more courses are moved to computer & web-based delivery, there is a need to evaluate these courses in a new format. In addition, some faculty who teach in face to face class settings are hesitant about giving up valuable instruction time to fill out evaluations. Also, the cost of scanning and scoring course evaluations from paper and pencil administration is growing as more students are enrolling in college and class sizes are increasing.

Web based administration of SETs provides an alternative to paper and pencil administration that helps with all these issues. The physical classroom does not limit web-based administration of SETs. Therefore, it provides a means for web-based and distance education courses to be evaluated in the same manner as classroom based courses. Additionally, it allows the evaluations to be completed outside of the normal class time, allowing for more time to be spent on instruction and more time for students to compose constructive written comments. Finally, because the data are entered directly into a computer scorable format, it saves both time and money for the scoring of student evaluations. But, before the move to computerizing SETs can take place, a closer look at the impact that such a move could have on SETs is needed, especially from the stand point of how a web-based system should be developed and how students will approach and use such a system.

When students engage in the task of completing a SET, they will rely in part on their previous experiences. The move to a web-based format will add a new dimension to the process which could change the way students respond to items on the SET. Research in the area of schema theory has investigated the ways in which prior experiences can guide people in the completion of a task (e.g. Hall, 2004; Fiore, Cuevas, & Oser, 2003; and Hicken, 1991). A schema can be thought of as a script or mental framework that represents what a person knows about a task and the steps needed to complete that task. The schema is based upon that person's previous experiences. The term *mental model* is often used to refer to a special type of schema that ties directly to the solving of a particular problem or performance of a given task (Park & Gittelman, 1995). From this perspective, students are likely to have developed schemata about the SET process over the course of their college careers. These schemata will guide the way they go about responding to questions. The move to a web-based system will require a modification to the student's existing schema. The greater the differences between the web-based system and the student's schema about SETs, the more challenging use of the web-based system will be for the students since it is violating their expectations and does not fit with what they are familiar with.

The idea that user expectations play a role in determining "ease-of-use" with a computer system, like a web-based survey system, has also been around for many years. However, the majority of the articles talking about this approach are aimed as "How To" guides rather than from the "How and Why" questioning of user actions. Additionally, much of the research related to cognitive theory in usability has been tied only to the area

of computer-assisted instruction, and has not examined within other applications, such as web-based SET systems (e.g., Chalmers, 2003).

Usability, as defined by the International Organization of Standards in ISO 9241.11 (1998), is the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." These three areas are commonly examined by measuring the frequency of errors by a user, the time taken to complete the task, and the user's overall satisfaction with the process. It is proposed that by viewing the usability of web-based SETs as a function of a student's schema about SETs a more complete model and set of recommendations for web-based SETs can be determined.

Despite the limited research and findings regarding web-based SET systems, many universities have already chosen to make the move to web-based SETs. However, with such limited research into web-based SETs, a wide variety of designs and implementations have been undertaken. For example, the University of Washington has selected to use a format that presents all items on a single page with students responding using radio buttons. California State Polytechnic University, Pomona and Edgecombe Community College have also started to administer course evaluations on-line, but instead of radio-buttons, they have chosen to use a drop-down list format for responses. Other universities have chosen to use text boxes in which students have to enter a number to respond to each item. So, with limited information about the quality of web-based SETs and even less research on how to design a web-based system, many universities are already moving into the world of web-based SETs.

The purpose of this dissertation was to examine the effects on the SET process that a move from paper-and-pencil to web-based format might have. More specifically, the study was looking to see how the number of items displayed per page and the manner in which students respond to questions influence the usability of a web-based SET system. Following the general description of usability provided by the ISO (1998), and as it is commonly measured, the time taken to complete the SET, the satisfaction with the web-based system, and the errors made will be used as measures usability for the web-based system. For purposes of this dissertation, the internal consistency of the SET and the discrepancy between the student's overall ratings and the "average" of ratings from other items will both be used as measures of error. By examining the impact different configurations have on usability, universities choosing to make the transition to administering SETs via the web will be able to approach the task of implementing such a system from a more informed position.

CHAPTER 2: REVIEW OF THE LITERATURE

Student Evaluations of Teaching (SETs)

In 1929, the University of Washington set forth guidelines for a system in which students would be allowed to evaluate the quality of the content and instruction in their courses (Wilson, 1999). Today, almost 75 years later, student evaluations of teaching (SETs) have become one of the most important, and influential, means for assessing instructional effectiveness of both courses and instructors (Marsh, 1987; d'Appolonia & Abrami, 1997; Shelvin, Banyard, Davies, & Griffiths, 2000; Millea & Grimes, 2002). Over the years, several factors related to SETs have received the attention of researchers from a variety of disciplines. At times it seems that research into this area is motivated, not by an interest in SETs themselves, but in answering the question "Why were my ratings so low?" (e.g., Greenwald, 1997). However, more often than not, the researchers become fascinated with the complexity of issues central to SETs and have produced a wide variety of research studies related to the use, interpretation, and administration of SETs in higher education.

Over the years SETs have been interpreted and used for a variety of purposes, including the evaluation of courses, instructors, and the combination of instructors teaching particular courses (Chang & Hocevar, 2000). In practice, however, the use of SETs tends to focus on the formative and summative evaluation of instructors (Spencer & Schmelkin, 2002; Hobson & Talbot, 2001). As described by Scriven (1967), SETs can be seen as formative in that they are designed to give feedback to instructors with the goal of improving the quality of teaching they provide. On the other side, SETs have evolved

into a summative evaluation used by administrators and personnel committees in the determination of tenure, promotion, and pay increases (Hobson & Talbot, 2001). This dual purpose guiding the use of SETs has led to a variety of research studies focused on the usefulness of SETs in higher education.

Although the majority of faculty prefer to believe that SETs lead to better teaching, several studies have discovered that SETs often have little impact on the quality of teaching with ratings of professors and departments often remaining constant over time. On one side of the debate over SETs, there are researchers who have provided evidence that given the proper environment and type of feedback SETs can lead to higher quality teaching (e.g., Stevens & Alaemoni, 1985; Marsh & Roche, 1993; and Piccinin, Cristi, & McCoy, 1999). However, these findings have to be interpreted with some caution since frequently the only measure of the quality of instruction used is the SET rating, and an increase in ratings might be reflective of issues unrelated to teaching itself, an issue discussed in more detail later. Also, it has been informally argued and discussed that there are ways to improve ratings on SETs that are unrelated to the quality of teaching, such as the timing of the administration of SETs during the semester in relation to exam feedback, announcing changes in the course such as extra credit opportunities immediately prior to giving SETs, and a variety of other somewhat questionable means.

On the other side, a wide body of research has developed questioning whether SETs do in fact lead to improved quality of teaching. One proposed reason for the lack of connection between SET ratings and improved quality of teaching is that students frequently do not know what constitutes “good” teaching and, as a result, have difficulty

providing feedback that can be usefully applied by faculty (Feldman, 1988). In some ways, this view can be characterized as the situation of an expert in a field being given advice by a novice. At times novices have good, creative ideas about how to improve the process because they are not constrained by a previous mindset, but more often they do not understand the far reaching results that their suggestions would create. Along the same lines, it has also been suggested that due to the underpinnings of SETs being based on the work of Feldman (1976), SETs are typically designed to evaluate a more didactic teaching style and, by their nature, result in low ratings for those who employ more interactive teaching methodologies that involve high levels of student interaction (McKeachie, 1997; d'Apollonia & Abrami, 1997; Koltich & Dean, 1999; Kember, Leung, & Kwan, 2002).

Additionally, as reported in the literature, faculty often feel that there is little incentive to improve the quality of teaching once SETs have reached an acceptable level for the administration (Kember, Leung, & Kwan, 2002). This ties directly into the use of SETs for summative evaluation by administrators and personnel committees in making tenure, promotion, and pay increase decisions. Numerous studies cite the use, and sometimes misuse, of SETs in making these decisions (McKeachie, 1997; Millea & Grimes, 2002). There is a growing sentiment on university campuses that, although student feedback is important, SETs are often overly weighted in tenure and promotion decisions (Obenchain, Abernathy, & Wiest, 2001). A large contribution to this attitude among faculty comes from the feeling that administrators often talk about the importance of quality teaching, but often do little to actually support and reward high quality

teaching (Boyer, 1990; Centra, 1993; Ramsden & Martin, 1996). In some cases, the perceived imbalance between the importance placed on poor SET ratings, without equivalent recognition to high SET ratings, leads to a resentment of the SET process.

A variety of factors unrelated to teaching, but which appear related to student ratings on SETS, are often cited as possible explanations of why SETs might not lead to increased ratings on future SETs. These factors can be categorized along the same lines that Chang & Hocevar (2000) presented as two of the evaluative objects to which SETs could be applied: instructor factors and course factors. Additionally, aspects related to the students themselves can influence the ratings they give on SETs.

A wide body of research has developed around the notion that aspects of instructors unrelated to their teaching can impact the ratings they receive on SETs. The fact that SET ratings might be unrelated to teaching quality has been a recurring theme in the literature for many years. Kerin, Peterson, & Martin (1975) showed that the enthusiasm instructors presented in class and the student's perceptions about the instructor's knowledge in the area played a major role in the determination of ratings students gave. Likewise, instructor charisma has been shown to relate directly to SET ratings, although there has been debate over whether charisma leads directly to better ratings, or if in fact a certain degree of charisma is necessary to provide good instruction (Shelvin, Banyard, Davies, & Griffiths, 2000). Additionally, a wide variety of other personal traits of instructors, such as warmth, supportiveness, personality, teaching style, and sense of humor have been shown to impact SET ratings (Langbein, 1994; Williams & Ceci, 1997; Clayson, 1999; Sojka, Gupta, & Deeter-Schmetz, 2002). Enthusiasm and

knowledge can easily be seen as potentially being related to the quality of instruction being provided. Even the warmth, supportiveness and charisma can be understood as qualities of good instructors.

Other factors shown to influence SETs raise some additional questions about the legitimacy of using SETs. In addition to a wide variety of personal traits being related to SET ratings by students, gender has also been shown to have an impact on ratings. Perhaps most interesting in relation to gender have been the findings that show students tend to rate same gender instructors higher than instructors of the opposite gender (Basow & Spielberg, 1987; Andersen & Miller, 1997). Also of interest is that fact that certain instructor aspects that would seem should be related to ratings, such as years of teaching experience, have been shown to not be related to SET ratings (Feldman, 1983). It is the impact that these non-teaching related aspects of instructors can have on student ratings that is the source of greatest concern for instructors, but they are not the only source of non-teaching related factors impacting the ratings.

Aspects of the course design have also been shown to influence the way students respond to SETs. Although a variety of issues, such as class size (Marsh, 1984; McKeachie, 1997) have received attention in the literature, course grades and difficulty have received the greatest attention. This research has varied from upwards of 70% of participants indicating that their expected grade in a course influenced their rating of an instructor to studies where participants claimed course grades had no influence on their ratings (Goldman, 1985). At this point, an extensive body of literature has evolved providing evidence that expected grades do in fact influence the ratings students give on

SETs, despite students not always being cognizant of the relationship (Feldman, 1976; Stumpf & Freedman, 1979; Cohen, 1981; Marsh, 1987; Marsh & Dunkin, 1992, Millea & Grimes, 2002). Likewise a relationship has been shown between the perceived course difficulty and grading leniency in a class and the ratings which students give on SETs (Greenwald & Gillmore, 1997; Millea & Grimes, 2002; Sojka, Gupta, & Deeter-Schmetz, 2002).

Finally, characteristics of the students themselves can have in impact on the SET ratings they provide. Aspects such as the student's overall GPA, hours spent studying for class, and interactions with the faculty member outside of class have all been shown to be related to the ratings given on SETs (Langbein, 1994). Perhaps more interestingly though is the impact that student perceptions and expectations can have on their ratings. Students just starting their college career often have expectations about grading, content, and difficulty that do not match the reality that they encounter in higher education (Krallman & Holcomb, 1997). Additionally, students tend to have preconceived ideas about the role they expect instructors to play in their education that frequently do not match the role instructors feel they serve (Andersen & Miller, 1997). This mismatch between the expectations students have coming into courses and the reality of the courses themselves creates a situation in which students might feel dissatisfied with instruction, not because the instruction is poor, but simply because it is not what they expected. This might also be a reflection of changes in the way students approach education, and higher education in particular, today versus students of the past.

The belief about the importance of SETs by students has also been examined as

an explanation of factors besides teaching quality that might impact ratings. This is an area that is only recently being examined in the literature, but it reveals some differences in the perceived importance of SETs by faculty and students. Whereas faculty see the impact that poor SETs can have on their careers through tenure, promotion, and pay decisions, students often feel that little attention is actually given to the feedback they provide on SETs (Sojka, Gupta, & Deeter-Schmetz, 2002; Spencer & Schmelkin, 2002). Since students seldom see changes that instructors might implement over time in response to SETs, they often feel that both administrators and faculty do not pay attention to the feedback they provide. As a result, some have argued that students choose not to provide quality feedback because they do not feel it will be taken into account (Sojka, Gupta, & Deeter-Schmetz, 2002). However, other research has shown that although students feel SETs are not taken seriously by administrators and faculty, SETs are still a good means, and in some cases the only means, of providing feedback to professors without fear of retaliation for the opinions they express (Spencer & Schmelkin, 2002). Also contributing to the student perceptions of the importance of SETs is the way in which they are typically administered.

Research has revealed both positive and negative impacts caused by different approaches to implementing SETs. One of the earliest findings related to the means in which SETs are conducted is the student's desire for anonymity (Feldman, 1979; Braskamp & Ory, 1994). Although this confidentiality aspect has become an integral part of most SETs, other aspects of the SET administration can often leave students with the wrong impression about their importance. Although faculty might proclaim the

importance of SETs, they are often administered during the last few minutes of a class session towards the end of a semester when students are focused on upcoming final projects and exams rather than reflection on the preceding weeks and months of the class (Chonko, Tanner, & Davis, 2002). Given the limited amount of time provided to students to complete SETs, it is little wonder that they feel faculty do not place much importance on them and the opportunity for written comments is frequently underutilized by students. Even when faculty believe they give ample time for students to respond, students often feel they are working against an invisible clock in a situation that frequently resembles that of course examinations more than it does a chance for reflection on the educational experiences the semester has provided.

Likewise, aspects of the SET administration process as it is commonly conducted have received some criticism in the literature that bears some mention. The way in which SETs are administered varies widely between campuses. On some campuses, faculty and teaching assistants are not allowed to have any contact with the SET forms, requiring a student in the course or some outside person to pick up the forms, administer them, and then return them to a location for scoring. However, other campuses take the other extreme and allow faculty to both distribute and collect their own SETs. And then there are the campuses that fall somewhere between the two, allowing faculty to take the SETs into the classroom, but requiring a student in the course to distribute, collect, and return the SETs to the administration or other office responsible for the scoring of the SETs. All of these methods of administration have received some criticism in the literature for being subject to tampering and abuse (e.g., Steinbart, 1989; Stanfel, 1991; Stanfel, 1998;

Layne, DeCristoforo, & McGinty, 1999; Dommeyer, Baum, Chapman, & Hanna, 2002). In many instances it would be possible for a faculty member, disgruntled student, or a staff member, to tamper with the SETs prior to their scoring. This could range from the instructor unduly influencing the SETs by making an announcement in class prior to their administration to the extreme of someone falsifying SETs to create either a favorable or unfavorable result (Stanfel, 1998). All of these procedural concerns have prompted some to suggest that an alternative means of administering SETs is needed.

As technological advances have been making increasing strides over the past 10 years, including the transition of the Internet from an arcane technology used by a few major research institutions into a worldwide system integrated into most aspects of daily life, many have provided suggestions about making use of technology to change the mode in which SETs are administered. One of the earliest proponents of moving to an electronic delivery system for SETs was Steinbart (1989), and his feelings and approach have been echoed in more recent research by Layne, DeCristoforo, & McGinty (1999) and Dommeyer, Baum, Chapman, & Hanna (2002). Although the exact nature of the electronic delivery system has changed over the years, the underlying principles have remained the same. In fact, some institutions (e.g., University of Washington, Georgia Institute of Technology, and Harvard Medical School) are in the process of making, or have already made, the transition to web-based administration of SETs to try and grapple with the changes caused by budget cuts and distance learning. Technology proponents and administrators, with little examination that the change in mode of administration might produce, have pushed the majority of these moves.

Computerized Attitude Measurement

A slowly increasing body of research examining the impact that changing modes of administration can have on SETs is beginning to take shape. However, before examining the research on this aspect of SETs in more detail, it is important to examine the research on the more general issue of changing the mode of administration for attitudinal measures. A review of the literature on cognitive and aptitude measures is not included since the results of such studies have remained rather constant over the years with no differences being discovered between computer and paper-and-pencil administered measures for non-speeded tests, though some differences do appear with speeded tests (Mead & Drasgow, 1993; Rosenfeld, Booth-Kewley, & Edwards, 1993; King & Miles, 1995; Stanton, 1998). A possible explanation for the difference in effect between speeded and non-speeded tests is that a respondent's proficiency with computers could play a larger role when a test is timed. Individuals who are proficient with computers would be able to focus their concentration on figuring out the correct answer to a problem, whereas the less proficient respondents would be forced to divide their attention between determining the correct answer and using the computer to make their response. Under a non-timed condition the extra time spent on remembering how to use the system would not be an issue, so no differences would be found. But, on a speeded test, the extra time necessary to remember how to use the computer to make the response could hinder the participant's performance. This issue, however, is related to cognitive tests so it will be left to other researchers to more fully address these questions. Research on mode differences for attitudinal measures has found a similar pattern, but the debate

and issues involved with attitudinal measures warrant further discussion because of their unique relation to SETs.

As far back as 1969, researchers have been investigating the impact that computerized data collection for attitudinal measures might have on responses. When issues such as social desirability were first examined, it appeared that computerized administration would lead to a decrease in socially desirable responses (Evans & Miller, 1969). In general, this trend continued throughout the 1970s and into the 1990s. A substantial body of research looking at the impact computerized data collection would have on social desirability all pointed to decreases in socially desirable responses (Scissons, 1976; Slack & Slack, 1977; Hart & Goldstein, 1985; Keisler & Sproull, 1986; Davis & Cowles, 1989; Koch, Dodd, & Fitzpatrick, 1990; King & Miles, 1995). However, beginning in the late 1970s and gaining momentum into the 1990s were findings that the decrease in socially desirable responses was reversed as the general population became more familiar with computers. During part of this time, the mode of administration appeared to create no difference with regard to social desirability (Rezmovic, 1977; Booth-Kewley, Edwards, & Rosenfeld, 1992; Dwight & Feigelson, 2000). Others actually noticed an increase in socially desirable responses (White, Clements, & Fowler, 1985; Martin & Nagao, 1989; Lautenshlager & Flaherty, 1990).

The primary reasons given for the changes in socially desirable responses all related to the issue of anonymity, which is central to SETs as well. As previously noted, a student's perceived level of anonymity does appear to have an impact on how they evaluate a course or an instructor. With regards to the computerized measurement of

attitudes, the original belief was that participants would be more honest when responding to the computer since it did not know who the person was, a finding supported over the years (Evan & Miller, 1969; Scissons, 1976; Hart & Goldstein, 1985; King & Miles, 1995). Another explanation for the decrease could be that participants are less likely to respond in a socially desirable way because they feel that the computer somehow "knows" if they are lying or misrepresenting themselves. But, on the other side of the issue is the "Big Brother" hypothesis which actually causes participants to increase their socially desirable responses because they feel they are being monitored (Rosenfeld, Booth-Kewley, Edwards, & Thomas, 1996; Martin & Nagao, 1989). Under the "Big Brother" hypothesis, the sense of anonymity being caused by the computer is decreased. Dwight and Feigelson (2000) noted that the trend in socially desirable responses has changed over the years. In the 1960s participants felt a greater sense of anonymity, but as people became more familiar with computers and the monitoring capabilities associated with them, we have achieved a point where there appear to be no differences in terms of socially desirable responses between computer and paper-and-pencil administered surveys. However, it has also been noted that as the sense of anonymity decreases, as would be the case for students whose enrollment in a course must be verified before completing an SET, the tendency to respond in socially desirable ways increases. This finding is in line with that of Kantor (1991), Stanton (1998), and Lautenschlager and Flaherty (1990) who found that as anonymity decreased, socially desirable responses increased. Given the evaluative nature of SETs, the impact that perceived anonymity might have on responses is important to keep in mind.

In addition to the comparison in terms of social desirability, research has also examined the equivalence of surveys given under different modes of administration. Again, similar to the research on social desirability, the findings have been mixed. In many cases there appear to be no significant differences created by changes in the mode of administration (Wilson, Genco, & Yager, 1985; Koch, Dodd, & Fitzpatrick, 1990; Rosenfeld, Booth-Kewley, & Edwards, 1993; Potosky & Bobko, 1997). Others, however, have noticed substantial changes in responses by participants (Katz & Dalby, 1981; White, Clements, & Fowler, 1985). Sproull (1986) noted that participants appeared to choose more extreme response options, but that it did not create any significant differences in the mean scores. However, others have noted mean differences, but provided little evidence to explain why those differences might exist (King & Miles, 1995; Potosky & Bobko, 1997).

Perhaps one of the best examinations comparing data collection via paper-and-pencil versus the computer, more specifically the internet, was conducted by Stanton (1998). With a focus on the evaluation of day-to-day interactions with a supervisor, strong support for the equivalence of the two modes of administration was found. No differences in factor structures or inter-scale correlations were found. Additionally, the internet based survey resulted in less missing data than the paper-and-pencil format, but it was also noted that there appeared to be increased measurement error for the data collected via the internet. A possible explanation for these findings could be that respondents felt more compelled to respond to each item when administered via the internet, even when the item might not have applied, and were therefore introducing more

error into their responses. Likewise, similar to the findings of Sproull (1986), an increase in the variability of responses was noted for the internet group. Stanton (1998) also argued that participant motivation was an issue for internet data collection with only those participants with a strong view on the issue and sufficient free time taking part in the study, which was supported in part by the written comments received from those in the internet group. Additionally, it was noted that internet data collection does not allow for an understanding of the participant's state of mind while completing the survey. It was noted that "respondents might be sleepy, angry, bored, intoxicated, or otherwise in an unsuitable frame of mind to provide honest, accurate responses" (Stanton, 1998). This could potentially create an unusual situation in regards to SETs since when given in class participants are more likely to be in a suitable frame of mind, whereas when completing the SETs outside of class, the chances of a student being in an unsuitable state of mind might be higher.

Finally, it is important to note that overall one common theme has emerged from the computerized measurement of attitudes. Participants continually indicate a preference for the computerized format over paper-and-pencil.. This finding has been consistent across the years (e.g., Erdman, Klein, Greist, 1983; Skinner & Allen, 1983; Rozensky, Honor, Rasinski, Tovian, Herz, 1986; Matheson & Zanna, 1988; Koch, Dodd, & Fitzpatrick, 1990). Part of this preference for computerized surveys might be because participants view them as being shorter and quicker to complete. Likewise, and of special importance in regards to SETS, computerized attitude measures might lead to higher levels of self-awareness (Matheson & Zanna, 1988), been viewed as more useful (Davis

& Cowles, 1989) and seen as more relevant (Rozensky, et al, 1986). If students are able to better reflect on their feelings before responding, feel the SETs are more relevant, and view them as more useful, the quality of data collected by SETs might be improved by switching to a computerized format.

Computerized SETs

As alluded to in the introduction, the primary focus here is on the impact that changing the mode of administration for SETs has on the way students respond. Although more and more universities are making the move to web-based administration of SETs, the research in this area is still in its infancy.

Steinbart (1989) took one of the first looks at the effects moving to a computer based administration could have on SETs. In this particular examination of the effect mode might have on SETs, a computerized, but non-web-based SET system was used. Participants were randomly assigned to complete either the electronic or paper based SET during a regularly scheduled class time. This helped ensure that all students that were present took part in completing a SET, but does not address the issues over response rates that would occur if the electronic evaluation were to take place outside of the normally scheduled class time. This is important in that the web-based system, as it is commonly presented, uses the "out of class" aspect as a major reason for implementing such a system. Surprisingly, participants in this study tended to rate the professor lower when using the electronic SET versus the paper-and-pencil format (Steinbart, 1989). This finding might be explained by the lack of familiarity and unease about technology which was present in society at the time of the study, especially considering that more recent

studies of non-paper-and-pencil formatted SETs have resulted in either equivalent, or slightly higher, average ratings for professors. Still, this study deserves a special note in that it was one of the very first studies to examine the impact changing the mode of administration might have on SETs.

However, despite the early beginning of research into the area, it took almost 10 years before a similar look was taken using a web-based SET system. Layne, DeCristoforo, and McGinty (1999) examined the effects on both ratings and student perceptions that a move to a web-based SET system might create. In perhaps one of the best examples of research in this area, no significant difference in average ratings for instructors was discovered based on mode of administration. In contrast to the Steinbert (1989) study, the participants of this study came from a variety of courses across several disciplines. Perhaps more importantly, all had active computer accounts with the university and were experienced with using the campus wide network for other activities (Layne, DeCristoforo, & McGinty, 1999). Likewise, rather than forcing students in the web-based format to complete the SET during the last few minutes of class time, this group was given until the end of the semester to complete the SET. As some might expect, the response rate of all students was higher for the paper-and-pencil based administration (60.6%), in which students completed the SET during the last few minutes of class, in comparison to the electronic format group (47.8%), which completed the SET outside of class. This difference in response rates between the two groups might be cause for concern in that certain groups of students might have been disenfranchised from the SET experience. No data were collected that would allow for analysis of differences in

responders versus non-responders between the two conditions. However, there was an attempt to determine the likelihood of completing the electronic SET, and it was discovered that:

- students with higher GPAs were more likely to complete the SET than those with lower GPAs,
- seniors were the least likely to complete the SET, whereas sophomores were the most likely,
- students majoring in science were more likely to complete the SET than those in other academic areas, and
- participants in the computerized SET condition were more likely to provide written comments on the SET.

Interestingly, feelings of anonymity by students in the study by Layne, DeCristoforo, & McGinty (1999) were mixed with some sensing a higher level of anonymity while others felt a lower degree of anonymity in the computerized format. The feeling of a lack of anonymity might have resulted from the fact that they had to "log in" to the campus system in order to complete the evaluation. However, despite this perceived lower degree of anonymity, the majority of participants in the computerized format favored it over the paper-and-pencil method. One of their main concerns was that the majority of students would not take the time to complete the electronic SET outside of class, a finding supported by the results showing less than half of those in the computerized SET group actually completed the SET (Layne, DeCristoforo, & McGinty, 1999).

More recently, research by Dommeyer, Baum, Chapman, & Hanna (2002) has examined the impact that paper versus web-based administration can have on SETs. The main focus of this article is that on-line administration of SETs can address a large number of concerns about the short-comings present with paper-and-pencil administered SETs, namely less chance for faculty to unduly influence their SETs, more time for students to think about responses to items on the SETs, and an opportunity for students who miss class on a given day to still have the opportunity to complete an SET. Unlike the previously mentioned studies, this study focused on the preferences of faculty for different modes of administration of SETs and the rational for those preferences. The majority of faculty surveyed in the Dommeyer, Baum, Chapman, & Hanna (2002) study preferred the paper mode of administration because of the reasons listed below:

- higher response rate to paper based administrations,
- paper-based was deemed more convenient for students,
- gives more weight to students who regularly attended class,
- more accurate responses from paper based administrations, and
- students more likely to feel their responses are anonymous.

Although evidence can be found to support the first two assumptions given, and the third assumption can be assumed given the nature of paper-based versus computerized administrations, little evidence exists for the fourth assumption (paper based being more accurate) and evidence on the last assumption is, at best, mixed.

Overview of Schema Theory and Mental Models

Over the past century, the notion of schemata has become firmly entrenched in the field of applied cognitive psychology. The schema concept has been traced back to the work of Kant, but has been modified and refined in more recent years (Dahlin, 2001). In early work by Bartlett (1932), schemata were described in terms of an individual's interaction with the environment and over time schema theory has come to refer more to mental structures or models that guide an individual's actions within a given environment or on a particular task. Works by Bransford (1979, 1985) and Eysenck (1993) further expand on the notion of schemata as mental models, and this approach represents the common way in which researchers in human factors and usability current use the term schema.

Mental models provide a dynamic model that helps guide an individual in solving problems and/or completing tasks. Mental models are more general schemata because the model more explicitly defines the relationships between objects in the model and how to perform specific tasks by making connections between new information and existing schema (Price & Driscoll, 1997; Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Winn and Snyder, 1996). Although mental models are often thought of as being constructed from mental images, they have sometimes been referred to as task schemata because they are often viewed as the knowledge acquired about how to perform a given task and they provide a context to help in the interpretation and understanding of new information (Kraiger, Ford, & Salas, 1993; Cantor & Mischel, 1977; Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Barker, 1999).

In fact, the extent to which the mental model or schema is developed and organized has been linked to an individual's expertise in a given area or with a given task (Green & Gilhooly, 1992). It has also been shown that the quality and organization of a mental model can have a major impact on how an individual performs when learning new information or a new way of completing a task (Chinnappan, 1998). Research on the "mental model hypothesis" echoes the findings and beliefs about the relationship between the fullness of a schema for a given task and expertise in that area (Baker & van Schaik, 1999; Barker, van Schaik, Hudson, & Tan, 1998). Because of the impact that such models can have on performance, they are often viewed as being integral to all of human activity (Barker, 1999).

The application of schema theory in applied cognitive psychology has reached into several areas specifically related to the issues of this study. For example, it has been proposed that appraisal ratings are based on a comparison with schemata for a job (Feldman, 1981; Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987). Likewise, Faulkner (1998) argued that users of computer systems form mental models of how things are supposed to work as they work on a task and that these models are used in the future by the user to improve performance on those tasks. These findings are believed to extend into the realm of SETs because students are repeatedly administered SETs leading to the development of a set of expectations based on those experiences. In most college and university settings, the SET process always occurs in essentially the same manner. This leads to students developing a schema that they use when approaching the SET task that is based on their prior experience and leads to a set of expectations for how it should

be done. Typically this process involves several components. In the vast majority of situations, SETs are conducted in class, during class time, with the professor absent from the room. Likewise, students generally know that they will be given a series of questions about the class and the instructor. They know that even if the list of questions spans multiple pages, all items are presented in a familiar format. Students come to expect to fill in the bubbles on a SET in the last couple of weeks of each class. With each successive administration of an SET, these expectations are reinforced, thereby making it easier to respond to SETs on future occasions since the general process is understood. However, if significant violations of the SET schema result from the move to a web-based system, performance on SETs could diminish. Research in the areas of human-computer interaction and web-site usability have begun to examine the impact that such moves to a web-based environment can have on surveys.

Issues of Usability in Web-Based Surveys

Lozar-Manfreda, Batageli, & Vehovar (2002) point out that at this time, the design and usability of a web survey are central to its success, though there is little agreement in the literature on what those design and usability elements should be. The majority of design elements and suggestions are actually based on assumptions drawing from research with paper-based questionnaires and are only recently being empirically examined within the web-based context. In their work, Lozar-Manfreda, Batageli, & Vehovar (2002) report the results of two studies conducted as part of the Research on Internet in Slovenia (RIS) project. By randomly presenting participants with either one-page or multi-page surveys, they concluded that because of load-time and transitions

required with multi-page surveys, the completion rate decreased as the number of pages increased. Lozar-Manfreda, et. al. also found that participants presented with a single page survey were less likely to skip items than when a multi-page design was used. They also discovered that participants were more satisfied with the single-page design, though no explanation for why was provided.

Heerwegh and Loosveldt (2002) conducted two studies examining the effect that the use of radio-buttons rather than drop-down lists had on survey respondents. Their primary focus was on the completion rate of participants and in the end they decided that for the general user radio-buttons would work better, depending on the item and survey being used. The two main factors influencing their decision related to task difficulty and similarity to paper. They concluded that drop-down lists were more difficult (two steps required to select an option versus one step required with radio-buttons) and that radio-buttons were closer to the paper-based format. Dillman and Christian (2002) noted that the use of blank text areas resulted in higher average ratings than "bubble" options for paper-and-pencil surveys, but no comparison has been done to drop-down lists or from within a computerized format.

Lozar-Manfreda and Vehovar (2001) also examined the response and completion rates achieved as the result of alternative design features. With follow-up interviewing of participants who did not complete the web based survey, they concluded that the length of a questionnaire was the best predictor of when a participant was likely to drop-out prior to completion. The second largest factor was a loss of interest in the survey, though it could also be argued that this loss of interest was a side effect of the length with longer

surveys not being able to hold the interest of participants through completion. Additionally, they discovered that the higher the percentage of open-ended items in the survey the more likely participants were to drop out of the survey without completing it.

Dillman, Tortora, and Bowker (1998) found that excessive use of graphics led to a decrease in completion rates, a decrease in written responses, and an increase in the item skip rate. Dillman, Tortora, and Bowker (1999) included this finding in a summary of design recommendations. One recommendation was to try and match web-based surveys to paper-based formats, and thus radio-buttons should be used instead of drop-down lists. A second recommendation was that a single page display should be used rather than a multi-page design, although no empirical evidence for these recommendations was provided. Dillman, Tortora, and Bowker (1999) also mentioned that a major guiding factor in their recommendation was the lack of computer literacy of the "typical" Internet user of the time.

This general thought was echoed again by Dillman (2000) when he noted that users were "confused" by drop-down lists, perhaps because of a lack of familiarity. Similarly, Dillman and Bowker (2001) noted that participants felt a loss of context when only one item was displayed at a time, but again little discussion of the impact this created beyond completion rates was noted. However, with the vast growth of the Internet over the past few years, many of these recommendations may no longer apply given that the "typical" Internet user is now more experienced than when these studies were conducted.

Reips (2002) examined the effect that the number of items displayed per page had on responses and found that when items were placed together on a page, participants responded differently to the items than when the items were placed on separate pages, though no definite pattern emerged. This is similar to the findings of Couper (2000) who found that participants often felt that items on the same page were "sets" and that the inter-item correlations between items increased when they were displayed on the same page.

Research Questions and Hypotheses

Despite a wide range of research on the usability of design aspects of web surveys, no study has fully examined the effects that an interaction of design elements might create for users, for example the combined effect that different item grouping options (i.e. single versus multiple page) and response format (i.e. radio-buttons versus drop-down lists) might have on user satisfaction, error rate, and time on task. Additionally, relatively few studies have attempted to provide an explanation of why certain design elements create usability problems for users, settling instead for simply describing the problem(s). Also, the vast majority of this research has been conducted using marketing surveys in which participants tended to have relatively light interest in the topic. Web-based measures of issues deemed more important by the participants, for example SETs completed by students, might lead to more tolerance for less usable systems.

Thus there are two gaps in the literature this study will begin to fill. First, by looking at how design elements or response type and item grouping interact with each

other, a more realistic impression of the usability of a web-based survey system can be achieved. Likewise, by using SETs, the results of this study can begin to fill the void in research on the usability of web surveys in which participants have a more vested interest as well as providing direction to colleges and universities on how best to design and implement a web-based SET system.

The overarching hypothesis is that the closer a web-based SET matches students' SET schema, the higher the usability rating will be. From this overarching hypothesis the following specific hypotheses will be examined in each of three key measures of usability:

Error Rate – As the differences in format between a paper-and-pencil SET and a web-based SET increase, participants will be less consistent in responding to items.

Completion time – The closer that a web-based SET format matches the more traditional bubble-sheet, the faster participants will complete the SET.

Satisfaction – Participants will be more satisfied with the design of a web-based SET system that matches their SET schema.

Additionally, the following secondary questions will also be addressed:

- Does the display of the SET affect the length of on-target written comments provided by students?
- Does the display of the SET affect the completion rate?

CHAPTER 3: METHOD

Participants

Students from 20 large section courses (sections with over 100 students enrolled) from a wide variety of departments at a large southern university were recruited for participation in this study. Courses from the following departments were included: Art, Biology, Computer Science, English, History, Mass Communication, Political Science, and Psychology. All participants were treated in accordance with the APA Ethical Guidelines. An email was sent to 4,372 students with a request to participate. Students were informed at the time of recruitment, and again when they began the study, that all information collected would remain confidential. They were also informed that the student evaluation of teaching (SET) completed as part of the study would be used for research purposes only and did not take the place of any in-class SET that was administered to them. Of the original number recruited, 936 (21.4%) began the study, and 856 began the SET portion of the study. A total of 816 completed the SET, and 791 completed all study related instruments. Demographic information for those that completed all instruments is provided in Table 1. The makeup of the sample closely matches that of the university in terms of ethnicity, but younger students were over-represented in relation to the university, most likely due to the fact that the large-section classes tend to be lower-level courses taken primarily by freshman and sophomore students. There was a slight under-representation of males in the sample when compared to the make-up of the university as a whole.

Table 1. Demographic Information

	N=791	%*
Gender		
Males	248	31.4
Females	534	67.5
Ethnicity		
African-American	35	4.4
Asian	18	2.3
Hispanic	136	17.2
White (non-Hispanic)	567	71.7
Other	24	3.0
Age Ranges		
18-19	285	36.0
20-21	302	38.2
22-23	96	12.1
24+	102	12.9
Classification		
First year	221	27.9
Sophomore	237	30
Junior	198	25
Senior	120	15.2
Other	5	.6

* Percents do not total to 100% because of missing data.

Materials

There were several measures used in this study. A web based SET, created specifically for this study, served as the primary instrument. The questions for the SET can be found in Appendix A. The format of the SET varied depending on condition, as described in more detail later, but for all conditions, participants responded on a 5-point scale ranging from Strongly Agree to Strongly Disagree for the first 23 items, followed by 4 open-ended comment questions, and 4 demographic items.

In addition to the SET, a 12-item user satisfaction scale (Appendix B), a 10-item Internet literacy and experience scale (Appendix C), and a 17-item experience and expectation measure (Appendix D) were also administered. The satisfaction scale is

based on items from the 50-item Software Usability Measurement Inventory (SUMI) which is a leading user satisfaction measure (Kirakowski, 2003). Since the full 50-item scale contains many items unrelated to the current system, a shortened version of the instrument is being used here. The Internet literacy measure is an updated version of the I-Net sub-scale of the Computer Interface Literacy Measure-Self Report (CILM-SR), which was designed to be a platform and software independent measure of basic Internet literacy (Turner, Sweeney, & Husman, 2000). The experience and expectation scale consists of two parts. The first part, displayed immediately after the participant accessed the system, consists of 12 items measuring past experience with SETs and 2 questions about expectations for web-based SETs. The second part, administered after completing the SET, consists of 3 items asking how close various aspects of the web-based SET matched the participant's expectations. For the question asking about expected response format in Part I, an actual example of a radio-button, drop-down list, and text-box were provided to help clarify what each option indicated.

The web-based SET system was created using the PHP scripting language in conjunction with the MySQL database system, and delivered via the Internet from a server located within a department at the University and maintained by the researcher. In using the web-based system, participants were required to sign-in using their university computing account and student ID number. Details of the web-based SET administration process are provided later. The system was designed to work with a wide variety of web browsers so that no group would be excluded based on their choice of web browser or operating system. The database used for storing the data contained 3 tables. The first

table contained a list of student ID, username, course enrollment, and information on which courses had been evaluated to verify that a participant was eligible to participate in the study and to determine which course (or courses) participants were allowed to evaluate. All study related data collected from participants were stored in the second table using a unique, randomly generated participant ID number that was not associated with the participant's student ID number or username to protect the anonymity of their responses. Only data from the first SET completed by the participant was stored in this table. Data from additional SETs completed by the participant were stored in the third table so that information from all SETs for a given course could be reported back to professors for their course.

Procedure

In the last few weeks of the semester, the researcher emailed instructors of the large section classes asking for their assistance with the project. For those that agreed, the researcher contacted students enrolled in the class via email and in person to inform them of the study. When students were recruited, they were told that the University was investigating the move to web-based SETs, and this study was designed to examine how such a system might work. They were also informed that the study was also being used as part of the dissertation research being conducted by the researcher. Additionally, they were told that their course had been selected since it is believed that the large sections would see the greatest benefit of moving to web-based SETs and the large class size makes the course more representative of the student body of the university. Written instructions containing a description of the study and sign-in procedures (with a direct

link to the study) were emailed to all students via their University email accounts. Two follow-up emails were sent to students who had not signed in as reminders about the study containing the same information provided in the original email. This allowed for the recruitment of students who were not present in class on the day when the announcement of the study was made. Data collection took place over a three week time period immediately prior to the end of the semester.

When participants first accessed the web-based SET system, they were required to sign-in using their University Computing Account username and their Student ID number. Instructions on how to obtain this information were provided for those that either did not have, or had forgotten, some of the information. This information was needed so that students could be presented with a list of the courses that they were enrolled in which were included in the study. The list of courses available for evaluation was presented as hyperlinks to the students. Participants enrolled in only one of the large section courses selected only had one course listed, while students enrolled in multiple large sections might see two or three courses listed. The majority of participants were enrolled in only a single large section course, though there was one student enrolled in five large section courses, four courses of which were included in this study. The system maintained a record of which courses a participant had evaluated so that a student could not complete multiple evaluations for the same course and would not be required to complete all evaluations at the same time.

After first signing into the system, Part I of the Experience and Expectations measure was presented to the participant. Based on the responses to the expectation

questions, the participant was randomly assigned to a condition that was believed to match their expectations to a varying degree based on two dimensions (item response format and item display format). There were a total of 6 possible similarity matches included, based on the responses to the two questions about item response and display formats:

- **Similarity Group 1:** matched the least expected choice for both questions
- **Similarity Group 2:** matched the second most expected for one question and the least expected for the other
- **Similarity Group 3:** matched the second most expected choice for both questions
- **Similarity Group 4:** matched the most expected choice of one question and the least expected for the other
- **Similarity Group 5:** matched the most expected choice for one question and second most expected for the other
- **Similarity Group 6:** matched the most expected choice for both questions

In making the assignment to similarity groups, the program would first randomly determine which of the 6 similarity groups to assign to the participant using a random number generator. For those participants assigned to Similarity Groups 2, 4, and 5 the system would then randomly select which of the two expectation dimensions (response method or number of items per page) to assign first. In other words, the system would vary whether expectations about response format or number of items displayed per page were violated and the degree of that violation. Then using the participants rankings for the first dimension, the system would determine which option from the second dimension

should be used. Four examples of this process are provided in Appendix E for clarification. This was done to ensure that there was no systematic bias created in the assignment of conditions and that a relatively equal number of participants would be assigned to all similarity conditions. Since the primary focus of this study was on the violation of expectations as determined by the similarity groupings, a lower degree of control was in place for assignment to the specific response method by presentation type conditions described below.

Three **methods of responding** to items, drawn from existing web-based SET forms found on the World Wide Web, were used as one of the presentation dimensions on which the conditions are based. For the radio-button condition, 5 columns of radio-buttons appeared after each statement, allowing the student to click on one radio-button for each item to indicate their level of agreement with the statement. In the drop-down list condition, a drop-down box allowed students to select one of the five ratings from a list. Each drop-down box defaulted to read "Please Rate". For the text-box condition, a blank text area was provided allowing the student to fill in their rating using the numbers 1 to 5. The response scale was provided on each page of the SET.

The second dimension of presentation to be varied was the **presentation type** which is determined by the number of pages and the number of items displayed on each page. The first condition was the format that is most common on existing web-based evaluations where all items are presented at the same time. The second condition presented groupings of items on separate pages, creating 4 pages for the SET items and a fifth page for the demographic items. The groupings consisted of items related to

evaluating the course, the instructor, overall ratings, and open-ended comments. The final condition presented participants with a single item per page, resulting in 31 pages (27 SET items plus 4 demographic questions). A set of screenshots with examples of the various pages for the entire web-based SET system can be found in Appendix F.

In most usability studies, participants are observed (and frequently video taped) in a controlled setting. This allows the researcher to know the true time on task that was required as well as the number of errors made while trying to complete a task. Since the logistics of this study did not allow for each participant to be directly observed, an alternative solution for measuring time on task and error rate was necessary.

When a participant began his/her first SET, the start time was recorded on the server. After the student had completed his/her first SET, the time at which they submitted the SET was recorded so that a total time to complete the SET could be calculated. Since it was possible for a participant to "walk away" in the middle of his/her SET, the times for participants which appear to be extreme outliers were not included in the analyses related to time on task. Further discussion of this issue and the measurement of error rate are provided later.

After the completion of his/her first SET, the system reminded the participant of the investigative nature of the study and presented them with the Internet literacy measure and user satisfaction instruments. These instruments were given after the first SET so that the number of SETs completed in the web-based format before completing the other measures remained constant for all participants. Next, Part II of the experience and expectation questionnaire was presented as a check to see how well the assignment to

Similarity Group worked in violating their expectations. Only data from the first SET completed were used in the analyses. After completing the measures, participants enrolled in more than one large section course were thanked for their participation and informed that any additional SETs completed would not be used as part of the study, but that they were free to complete the additional SETs if they choose to do so. They were then returned to the initial course evaluation selection screen. Participants in only one large section course were thanked for their participation and returned to the initial course evaluation selection screen.

After the data collection time period ended, the web-based SET was removed from the website and replaced with a brief description of the nature of the study, a list of the research questions being asked, a brief summary of the preliminary results, and other debriefing information.

CHAPTER 4: RESULTS

The analysis of the data was divided into three areas to address the various issues covered in this study. First, an analysis from a general usability standpoint was conducted to see if one combination of design characteristics produced a more "usable" system, regardless of participant's level of Internet literacy. Second, analyses examining the role that schema and expectations played in determining the usability of the web-based SET system were conducted in an effort to provide a further explanation of why one design might be preferable to the others. Finally, analyses related to the secondary questions were undertaken.

Before discussing the specific analyses which were conducted, it is important to address how error rate and time on task were calculated. Unlike typical usability studies which count the number of errors made while completing a series of tasks, there were two measures of "error rate" used in this study.

The first error rate in this study was measured by creating a disparity score by subtracting the average rating given to the instructor on individual items from the rating given on the overall rating question. The same process was used to create a disparity score for the rating of the course. It seemed that the lower the level of agreement between the ratings for individual items and the overall rating, the more error in the ratings existed. These disparity scores were used for the majority of the analyses described later.

The second error rate measure, used purely for descriptive purposes, was the coefficient alpha reliability for the SET items under the various display conditions. In

essence, it was assumed that the higher the internal consistency in responding to the items, the less error was included in the individual ratings.

As previously mentioned, both the beginning and end times for the SET were recorded for each participant. The difference in seconds between the two times was used as the measure of time on task. Since it is possible that a participant "walked away" during the SET process, those participants with unusually long completion times could unduly impact the average completion times. In order to avoid those with unusually extreme completion times (both above and below the mean) from skewing the analyses, only the middle 95% of completion times were used. This yielded a total of 773 participants for the analyses who completed the SET in 1.5 to 20 minutes.

General Usability

Currently the literature on the analysis of the components of usability is mixed on the whether error rate, time on task, and satisfaction should be analyzed in unison or separately (Frøkjær, Hertzum, & Hornbæk, 2000). As a result, some preliminary analyses were conducted to assess the most appropriate analysis for the data.

The first analysis was a correlation between the disparity scores for error rate, the time on task, and user satisfaction. This was done to help determine whether the data should be analyzed using a univariate or multivariate approach. Additionally, Internet literacy (as measured by the CILM-SR) was correlated with each of the 3 usability measures to assess its viability for use as a covariate in later analyses. Although all correlations were statistically significant ($p<.05$), the low correlation coefficients suggest that the significance could largely be an artifact created by the large sample sizes (see

Table 2.) Descriptive statistics by condition for each of the three usability measures can be found in Table 3.

Table 2. Correlations between usability measures and Internet Literacy

	Time	Error	Satisfaction	CILM-SR
Time	---	-0.096 (n=766)	0.096 (n=758)	-0.055 (n=745)
Error		---	-0.101 (n=752)	-0.049 (n=739)
Satisfaction			---	0.245 (n=743)
CILM-SR				---

Table 3. Mean, SD, and N by condition for each usability measure.

	Error Rate			Satisfaction			Time on Task		
	Mean	SD	n	Mean	SD	n	Mean	SD	n
Radio Buttons									
1 Page	-.51	.72	81	50.47	8.92	81	264.26	172.17	81
Grouped	-.37	.93	77	50.34	8.51	76	254.58	153.88	77
1/Page	-.47	.51	80	46.38	8.08	78	323.79	164.80	80
Drop Down									
1 Page	-.50	.59	94	46.99	9.03	92	302.08	187.69	95
Grouped	-.51	.39	78	48.16	8.33	76	324.94	197.95	78
1/Page	-.48	.39	75	47.09	9.00	75	355.64	208.14	75
Text Box									
1 Page	-.37	.72	88	50.09	8.14	89	254.07	171.70	89
Grouped	-.49	.77	88	48.09	8.55	87	297.67	183.91	90
1/Page	-.34	.67	105	47.86	7.89	107	368.47	217.16	108

A 3x3 ANOVA was used to analyze the error rates yielding no significant differences for response format ($F(2,757)=1.48$, $p>.05$), display format ($F(2,757)=.065$, $p>.05$) or their interaction ($F(4,757)=1.313$, $p>.05$). A similar analysis of the user satisfaction ratings also yielded no significant effect for the response format ($F(2,749)=2.52$, $p>.05$) or the interaction of response and display formats ($F(4,749)=2.07$,

p>.05). The satisfaction ratings did differ based on the display format (F(2,749)=4.38, p<.015, eta^2=0.012), with users rating the display of all items on a single page (M=49.12) higher than when each item was displayed on a separate page (M=47.20). The low effect size raises the question as to whether this difference alone is of any practical importance when designing a web-based SET system.

The analysis of the time on task data with a 3x3 ANOVA yielded what is perhaps more important information about the design of a web-based SET system. It was discovered that significant differences in time on task were caused by the display format (F(2,764)=11.59, p<.01, eta^2=.03) with each item being presented on a separate page (M=351.22 seconds) being statistically different from the conditions where all items were presented on the same page (M=274.40 seconds) and from the "grouped' item display condition (M=292.81 seconds). This is actually not too surprising given that one would expect it to take longer to go through 31 pages than 1 or 5 pages, especially when page load time is taken into account. Although the effect size is relatively small, it does translate into almost a minute difference in the average times between the grouped and single item displays. Since most students would be asked to complete 4 to 6 evaluations in a given semester, that one minute difference translates to several minutes which could be saved by not using the single item display format. Likewise, a statistically significant difference was discovered for the response format (F(2, 764)=3.81, p<.05, eta^2=.01) with the drop-down lists (M=351.22 seconds) taking significantly longer than when radio buttons were used (M=274.40 seconds). Again, although the effect size appears relatively small, it translates to over a minute difference per evaluation completed. This

combination of findings begins to suggest that the combination of radio-buttons with items grouped per-page might have an advantage over other combinations.

The final usability analysis of the system looks at the internal consistency for the web-based SET. When calculated using all available data (n=777), coefficient alpha = .9553, but a closer look by condition revealed some additional information that could be relevant in the design of a web-based SET system. As can be seen from the coefficient alphas found in Table 4, not all conditions functioned equally well. Perhaps the most notable trends in these data show that regardless of display format, text boxes always resulted in less consistent responses by participants. Likewise, regardless of response format, participants were less consistent in their responses when only a single item was presented per page. When drop-down lists were used, participants were more consistent when all items appeared on the same page. In general, participants appeared to be most consistent with radio-buttons when the items were grouped onto separate pages.

Table 4. Coefficient alphas by scale and condition.

	Instructor Items	Course Items	All Items
Across conditions	.9153	.9355	.9553
Radio-Buttons			
1 Page	.9329	.9451	.9629
Grouped Pages	.9610	.9701	.9818
1/Page	.8195	.9053	.9209
Drop-Down Lists			
1 Page	.9285	.9436	.9549
Grouped Pages	.9077	.9341	.9521
1/Page	.9008	.9222	.9483
Text Boxes			
1 Page	.8970	.9099	.9408
Grouped Pages	.8205	.9168	.9239
1/Page	.7824	.8988	.9148

Schemata and Expectations

The first question to answer in relation to the role that a SET schema might play was to verify that such a schema does appear to exist. Based on the expected response and display formats indicated on Part I of the Experience and Expectations instrument, a variable was created to indicate which combination the participant felt was most likely to be used. A Chi-Square for Goodness of Fit was used to test whether one (or more) combinations stood out as being more likely to be selected by participants. Since this expectation was not dependent on the time taken to complete the evaluation, data from all participants who answered the questions were included in the analysis. If no schema exists then there should be no pattern in the expected display. However, if a schema about SETs does exist, then one of the combinations should stand out from the rest as being selected by the vast majority of participants. This was in fact the case, with the radio-button response format and single page display being the most commonly selected combination ($\chi^2(8, n=856)=1214.778$, $p<.001$). Additionally, a chi-square used to assess the relationship between the expected response/display formats with previous paper-and-pencil SETs showed that a strong relationship was present, suggesting that expectations about web-based SETs was strongly influenced by previous exposure to paper-and-pencil SETs ($\chi^2(64, n=694)=142.61$, $p<.001$).

Given that evidence for a schema was found, the next step was to check to see if the assignment to Similarity Groups worked as expected in terms of violating that schema. To do this, an average "violation" score was created by reverse scoring the questions about response and display format from Part 2 of the Experience and

Expectations instrument (so that a higher score would indicate a greater mismatch with expectations) and then averaging the two responses. Descriptive statistics for similarity groups based on violation scores are provided in Table 5. Overall, the violation of expectations worked as expected with the exception that Similarity Group 2 ended up being rated as a higher violation than was found in Group 1, where a complete mismatch with expectations was assigned. A one-way ANOVA revealed that the assignment to similarity group did cause a statistically significant difference in violation of expectations ($F(5,740)=4.151$, $p<.001$, $eta^2=.027$). Using a Scheffe post-hoc procedure, Similarity Group 6 was found to be statistically different from Similarity Groups 1 and 2.

Table 5. Violation statistics by Similarity Group

Similarity Group	Mean	SD	N
1	2.66	.945	124
2	2.73	.874	130
3	2.63	.916	120
4	2.59	.883	128
5	2.46	.846	123
6	2.27	.931	121

The next series of analyses were conducted to examine differences in usability measures based on the violation of the SET schema. Based on information gained during the analysis of the general usability issues, ANOVAs were used for each of the analyses. Based on Similarity Group, there was no difference in the average errors ($F(5,760)=.987$, $p>.05$), user satisfaction ($F(5,752)=1.418$, $p>.05$), or time to complete the evaluation ($F(5,767)=1.47$, $p>.05$). Descriptive statistics for the similarity groups can be found in Table 6.

Table 6. Mean, SD, and N by Similarity Group.

	Error Rate			Satisfaction			Time on Task		
Similarity Grp	Mean	SD	n	Mean	SD	n	Mean	SD	n
1	-.360	.647	126	46.90	8.18	126	342.07	207.05	128
2	-.415	.657	137	47.80	7.96	135	317.45	201.07	138
3	-.501	.706	124	48.36	8.64	121	301.79	183.64	125
4	-.517	.603	130	49.12	8.57	129	291.14	194.59	132
5	-.459	.605	127	48.72	8.78	125	291.92	174.29	127
6	-.451	.709	122	49.37	9.21	122	292.71	170.31	123

Secondary Questions

A chi-square revealed that the combination of design elements did not lead to changes in completion rates (χ^2 (9, n=856) = 13.12, p>.05). Similarly, the violation of expectations did not appear to lead to any differences in completion rates (χ^2 (5, n=856) = 7.49, p>.05). Since the questions gathering demographic information about participants was part of the web-based SET, no information about gender, age, ethnicity, or classification is available for those that did not complete the web-based SET.

The final set of analyses examined differences in responses to the open-ended items on the web-based SET. For purposes of analysis, any text entered into the comment box for an item was considered a comment. An average comment length was calculated using the mean of the first three open-ended questions on the SET. The final open ended question ("Additional comments or suggestions") was not included since more than half of the participants did not respond to that question. A one-way ANOVA revealed that the similarity group did not create a difference in the average length of a comments made (F(5,772)=.865, p>.05). Likewise, a 3x3 ANOVA examining the effects of response and

display formats resulted in no significant effects with regards to average comment length ($F_{Response}(2,763)=1.92$, $p>.05$; $F_{Display}(2,763)=.01$, $p>.05$; $F_{RxD}(4,763)=.79$, $p>.05$).

Next, comparisons were made based on the number of comments made, again counting any text entered for an open ended question as a comment. The assignment to similarity group did appear to create a small difference in the number of comments made ($F(5,766)=2.72$, $p<.05$, $eta^2=.02$). Descriptive statistics for the number of comments made by Similarity Group can be found in Table 7. A 3x3 ANOVA examining the effects of response and display formats revealed no effect for the interaction ($F(4,763)=.24$, $p>.05$) or response format ($F(2,763)=.95$, $p>.05$). There was, however, a significant effect for display format, though with a small effect ($F(2,763)=8.23$, $p<.001$, $eta^2=.02$). When items appeared alone on a page (M=3.27), there were more comments made than when all items appeared on a single page (M=2.77) or when the items were grouped on a page (M=2.97). It is important to note that, all open-ended comments appeared on the same page when items were grouped and always followed all of the fixed response items. A different pattern of open-ended responses might be found if they were grouped individually with the corresponding fixed-response items.

Table 7. Average number of comments made by Similarity Group.

Similarity Group	Mean	SD	N
1 (Complete Mis-Match)	3.29*,**	1.29	127
2	3.15**	1.25	138
3	2.87*	1.47	125
4	2.80*,**	1.53	132
5	3.09	1.28	127
6 (Complete Match)	2.81*	1.50	123

* Group 1 is statistically different from Groups 3, 4, and 6.
** Group 2 is statistically different from Group 4.

CHAPTER 5: DISCUSSION

Usability Analysis

Although very small, the direction of the correlations between error and time on task and between time on task and satisfaction were in the expected direction. The less time a person took, the more errors they made, but at the same time the less time it took, the more satisfied they were with the system. This indicates a strong need for finding the right balance between needing students to take enough time to be careful in their responses, but at the same time not making the process take so long that satisfaction with it drops below a level where students will use it.

One explanation for the lack of relationship of usability measures with computer literacy could be because the measure used was initially designed to assess a very basic level of literacy, with over 80% of the respondents scoring towards the top of the scale (over 20% scored at the highest possible level on the instrument), a ceiling effect kept the measure from being a good fit in this study as a covariate. A better instrument, designed to measure across a wider range of literacy might be beneficial, though the trend of higher computer literacy leading to higher satisfaction makes sense in that those who use the computer more often will probably view the web-based SET in a more positive light. Additionally, the possibility of a schema about web-based surveys might have been activated in addition to, or in place of, the intended SET schema. This issue is discussed in more detail later.

Although differences on most measures of usability were minimal, overall the radio-buttons appeared to have a slight advantage as a response format over drop-down

lists and text boxes in terms of errors, satisfaction, and time on task. Also, when radio buttons are used, the data suggest that grouping items by topic on separate pages reduces errors (both in terms of the disparity scores and coefficient alpha reliabilities) and decreases the time necessary to complete the task, while maintaining a relatively high satisfaction level. Since radio-buttons are the response option most similar to the "bubbles" used on many automated scoring sheets, the finding that they worked best coincides with the notion that participants were trying to assimilate the web-based system into their existing schema about SETs, thus providing some evidence for the overarching hypothesis proposed earlier.

Schemas and Expectations

The facts that many students indicated the same combination of response and display format for the web-based SET and that these expectations closely paralleled their self-reported, prior experiences with web-based SETs, provide support for the notion that a schema about SETs does exist. Given the overall findings of the study, the expectations from paper-and-pencil SETs appear to be carried over by students to their expectations about web-based SETs. However other schema might also be activated by the web-based format that make assimilation and accommodation of the new format easier than initially anticipated. For example, if a schema about web-based surveys was activated at the same time as schema for SETs, the web-based schema could make responding to the SET easier. In future research, the addition of a better measure of computer expertise, including experience with web-based surveys, could help shed additional light on this issue.

Although significant differences in violation scores existed for the extremes of the similarity groupings, the small differences between similarity groups raises some question as to how well-defined the schema about SETs might be. Given that several participants would indicate a relatively high match with expectations when presented with a format previously indicated to be the least expected, questions about the strength of the SET schema seem warranted. It is possible that no major differences in expectations existed between the various response and display formats, and although a slight preference was indicated in the initial part of the study, this preference was not great enough to have a significant impact overall. Another explanation could be that participants realize that things on the web do not always match their expectations from paper-and-pencil experiences. So, even though one response option might be expected, it is possible that participants realized another option was equally likely to be used. Overall, participants seemed to adapt well to the web-based SET format, regardless of how well it matched their initial expectations.

Secondary Questions

It is reassuring that the response and display format options did not impact completion rates since all of these combinations are currently being used for web-based SET systems, as previously described. Likewise, the fact that violation of expectations did not lead to an increase in non-completion rates is somewhat reassuring in that we might not have to worry as much about students failing to complete an evaluation because the format violated their expectations. However, the reasons why students

continued to complete the SET are not clear, so caution should be used in drawing any major conclusions from these findings.

Although differences in the length of comments were not found for the various conditions, the difference in the number of comments made is interesting. It is possible that participants who had already taken the time to move from page to page for each new question were more willing to take the time to enter comments. Additional analyses of the nature of the comments made could be useful.

General Recommendations

Based on the findings, the following general recommendations are suggested for those either in the process of or considering the development of a web-based SET system. The ideal method of response for a web-based SET system appears to be a radio-button. Overall, the radio-button response format seemed to be the easiest for participants. Items about similar issues and topics should be grouped onto multiple pages. In the current study, items for the instructor were on a page separate from items about the course. This appears to have led to more consistent responses by participants as they focused on each of the general topics of the SET (instructor and course.) So, the combination of radio-buttons, with items grouped by area on separate pages appears to yield the best results from a usability standpoint. Given the time to complete the SET, it is recommended that the single-item per page format not be used. Likewise, text-boxes seemed to lend themselves to more error in response entry. Some participants entered multiple ratings in the same text box or used values outside of the range of the scale. Although safe-guards can be put in place to prevent this sort of data input error by participants, eliminating the

possibility to make the mistake in the beginning by using a fixed-response format seems the better solution.

Also, although not directly addressed in this study, a review of some comments made by participants (either in the open-ended comment section of the SET or via email to the researcher) seems to indicate that concerns over anonymity were not prevalent. A few commented that with paper-and-pencil SETs, they felt concerned over professors recognizing their handwriting and that the web-based SET removed that issue.

Limitations and Future Directions

As with any study, there are ways in which this study could be improved. For example, the measure of computer literacy used in this study (the CILM-SR), failed to adequately discriminate among participants in terms of their computer literacy skills. As a result, it was not possible to examine how participants at different levels of computer literacy responded to the various conditions. Although the CILM-SR is well suited for determining basic levels of computer literacy (its originally intended use), a better measure of computer literacy might be useful for determining the ways in which computer literacy and prior computer experiences factor into the use of a web-based SET system. Still, the fact that so many students were able to do well on the basic measure of computer literacy is reassuring since web-based SETs will need to be usable by all students, regardless of their level of expertise with computers. Additionally, rather than focusing on computer literacy in general, the focus should be on knowledge of and experience with web-based surveys and forms since those are the areas most closely related to use of a web-based SET system.

The recording of demographic information so late in the process here proved a limitation in that certain questions could not be answered about participants who dropped out of the study before completing that part of the study. Related to this is the potential non-response bias created by those who chose to discontinue participation or not to participate in the study. Almost 80% of those recruited for the study chose not to participate, and an additional 3% dropped out of the study after completing the web-based SET. There is a chance that the remaining 17% did not adequately represent the population. Such a difference might explain why only those with high scores on Internet literacy were found in the final sample. Had a larger percentage of participants taken part in the study, differences in the results might have been discovered.

Additionally, questions about the comparability of web-based to in-class SETs were not addressed in this study, and although there is growing research on this issue, no solid conclusions have been reached. This is especially true in examining the difference in response rates of different groups between the two modes of administration. Although some researchers have started to examine the potential for non-response bias (Thorpe, 2002), further research into these issues is encouraged, especially in comparison of the web-based to paper-and-pencil format.

APPENDIX A: STUDENT EVALUATION OF TEACHING QUESTIONS

Questions about the instructor

The instructor was well prepared for class.

The instructor encouraged student participation in class.

The instructor treated students with respect.

The instructor was open to diverse opinions.

The instructor effectively explained material in clear manner.

The instructor graded in a fair manner.

The instructor provided helpful feedback on course assignments.

The instructor appeared knowledgeable about the course material.

The instructor appeared interested in teaching the course.

The instructor was available to students outside of class.

Questions about the course

The course was well organized.

The course was intellectually challenging.

The course covered the content outlined in the syllabus.

The course provided meaningful insight into the subject.

The objectives and assignments for this course were clearly explained.

The course provided me with a valuable learning experience.

I feel this course has increased my knowledge about the subject.

As a result of this course I am motivated to become a better student.

As a result of this course, I am better able to think analytically.

As a result of this course, I learned the major concepts covered in this course.

Overall

On a scale of 1 (worst) to 5 (best), I would rate this course as a ___

On a scale of 1 (worst) to 5 (best), I would rate this instructor as a ____

On a scale of 1 (worst) to 5 (best), I would rate my experience in this course as a ___

Open ended

Best aspects of the course?

Worst aspects of the course?

Suggestions for improving the course?

Additional comments or suggestions?

Demographic

Gender

Age

Ethnicity

Classification

APPENDIX B: USER SATISFACTION MEASURE

For each of the following items, please respond using the following 5-point scale:

1	2	3	4	5
Strongly Disagree		Neutral		Strongly Agree

1. I found it easy to navigate between items.
2. I found it difficult to respond to items.
3. I found it frustrating to use the web-based system.
4. I found it easy to access the web-based system.
5. I found it easy to learn how to use the on-line evaluation system.
6. I often forgot where I was in the evaluation process.
7. I found the interface easy to understand.
8. I often felt confused by the response options available.
9. I am satisfied with the web-based based system.
10. I felt it took too long to complete the evaluation.
11. Overall, I enjoyed completing the web-based evaluation.
12. I would enjoy completing more evaluations using the web-based system.

Items 1-12 are adapted in part from the Software Usability Measurement Inventory (SUMI) to reflect the web-based SET rather than general use software. Items 2, 3, 6, 8, and 10 are reverse scored.).

APPENDIX C: INTERNET LITERACY (ADAPTED FROM THE CILM-SR)

For each of the statements below, please rate how well each of the statements applies to you using the following scale:

1 2 3 4 5

not at all like me neutral or unsure very much like me

____ 1) I use the Internet on a daily basis.

____ 2) I often download programs and applications from the Internet.

____ 3) The Internet is my primary source of information about current events.

____ 4) I do not use the Internet.

____ 5) I often feel lost or confused when using the Internet.

____ 6) I think the Internet is a useful tool.

____ 7) I do not see myself using the Internet after I graduate.

____ 8) I prefer to use resources other than the Internet to find information.

____ 9) I can easily find the information I need on the Internet.

____ 10) I would rather search in the library than the Internet to find information.

11) Do you have Internet access where you currently live? _ Yes _ No

12) How often do you use the computer labs on campus in a typical month?

_ I never use the computer labs on campus.

_ Sometimes, but less than once a month

_ 1-4 times a month

_ 5-8 times a month

_ 9-12 times a month

_ More than 12 times a month

13) How often do you access the WWW in a typical month?

_ I do not use the WWW.

_ Sometimes, but less than once a month

_ 1-4 times a month

_ 5-8 times a month

_ 9-12 times a month

_ More than 12 times a month

14) How often do you complete web-based surveys in a typical month?

_ I never complete web-based surveys.

_ Sometimes, but less than once a month

_ 1-4 times a month

_ 5-8 times a month

_ 9-12 times a month

_ More than 12 times a month

* Items 4, 5, 7, 8, & 10 are to be reversed scored. Items 11-14 are additional items for this study.

APPENDIX D: COURSE EVALUATION EXPERIENCES AND EXPECTATIONS

PART I – Displayed after students first access the system, before completing the SET.

The following questions are designed to look at you previous experiences with course evaluations and your expectations regarding a web-based course evaluation.

Have you ever completed a course evaluation before? ___ Yes ___ No

Have you ever completed a web-based course evaluation before? ___ Yes ___ No

For the following 8 items, please indicate whether course evaluations you have previously have asked questions about the indicated topic.

	Yes	No
Questions about the instructor's interactions with students	___	___
Questions about how prepared the instructor was for class	___	___
Questions about how assignments were graded	___	___
Questions about the instructor's ability to explain the material	___	___
Questions about the organization of the course	___	___
Questions about how challenging the course was	___	___
Questions about how much you learned in a course	___	___
Questions about value to you of a course	___	___

Please rank how common the following 3 response formats have been on previous course evaluations you have completed: (*Please use each number only once.*)

1- Most common 2- Second most common 3- Least common

___ Bubble-sheet (e.g., Scantron format)

___ Written response (e.g., short answer)

___ Fill in the blank using a provided scale

Please rank how common the following 3 descriptions of length have been on previous course evaluations you have completed: (*Please use each number only once.*)

1- Most common 2- Second most common 3- Least common

___ All items appeared on a single page.

___ Items appeared on 2-3 pages (or one page front & back).

___ Items appeared on more than 3 pages.

Please rank the following 3 response formats in terms of your expectations for a course evaluation given via the web: (*Please use each number only once.*)

1- Most expected 2- Second most expected 3- Least expected

___ Radio buttons

___ Drop-down list

___ Text box for entering a number from a provided scale

Please rank the following 3 options for how items might appear in terms of your expectations for a course evaluation given via the web: (*Please use each number only once.*)

1- Most expected 2- Second most expected 3- Least expected

___ All items appear on a single page.

___ Items appear on 2-3 pages.

___ Items appear one at a time.

PART II – Displayed after completion of the first SET and the Satisfaction measure. Please rate how similar the following aspects of this web-based course evaluation matched your expectations using the following scale:

1	2	3	4	5
Not at all as expected		Somewhat as expected		Exactly as expected

____ Way of responding to items

____ Number of items displayed per page

____ Content of items

APPENDIX E: ASSIGNMENT TO CONDITION PROCESS EXAMPLES

Example 1 – Participant assigned to Similarity Group 6

Since Similarity Group 6 consists of both dimensions matching the participant expectations, the system will check the participant's responses to the two expectation items from the experience and expectation questionnaire. For example, if the participant indicated that they most expected to see radio buttons, then radio buttons will be used for the response format. Likewise, if the participant ranked 2-3 pages as the most expected option for the number of pages, then that condition for the display type will be used. This creates a perfect match between the participant's expectation and the way the SET will be displayed.

Example 2 – Participant assigned to Similarity Group 1

Since Similarity Group 1 consists of both dimensions being the least expected by the participant, the system will check the participant's responses to the two expectation items from the experience and expectation questionnaire. For example, if the participant indicated that they least expected to see text boxes, then text boxes will be used for the response format. Likewise, if the participant ranked one item per page as the least expected option for the number of pages, then that condition for the display type will be used. This creates the greatest disparity between the participant's expectation and the way the SET will be displayed.

Example 3 – Participant assigned to Similarity Group 4

Similarity Group 4 consists of matching the most expected choice for one dimension with the least expected for the second dimension. Therefore, the system must

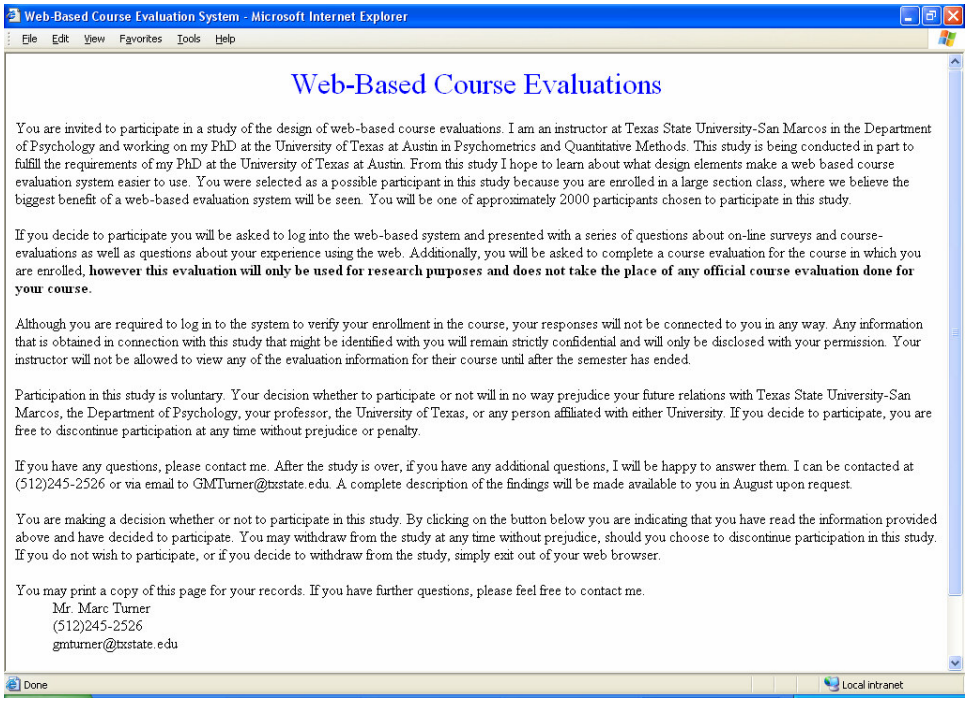
randomly determine which dimension to consider first. For example, the system might decide that the display type is the first dimension to consider. So, the system would then see which display type the participant marked as the most expected (e.g., all items on one page). Next it would select the response method option ranked as the least expected (e.g., drop-down lists), thus creating a perfect match on one dimension and a perfect mismatch of expectations on the second.

Example 4 – Participant assigned to Similarity Group 5

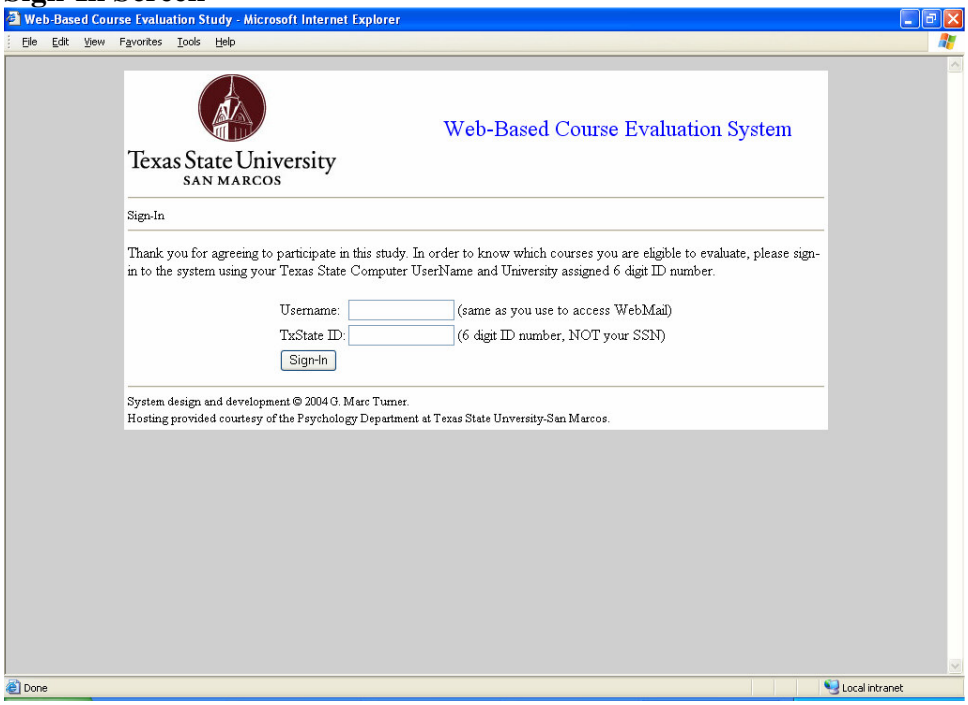
Similarity Group 5 consists of matching the most expected choice for one dimension with the second most expected option for the second dimension. Therefore, the system must randomly determine which dimension to consider first. For example, the system might decide that the response method is the first dimension to consider. So, the system would then see which response method the participant marked as the most expected (e.g., radio-buttons). Next it would select the response method option ranked as the second most expected (e.g., drop-down lists), thus creating a perfect match on one dimension and a partial mismatch with expectations on the second.

APPENDIX F: WEB SET SCREENSHOTS

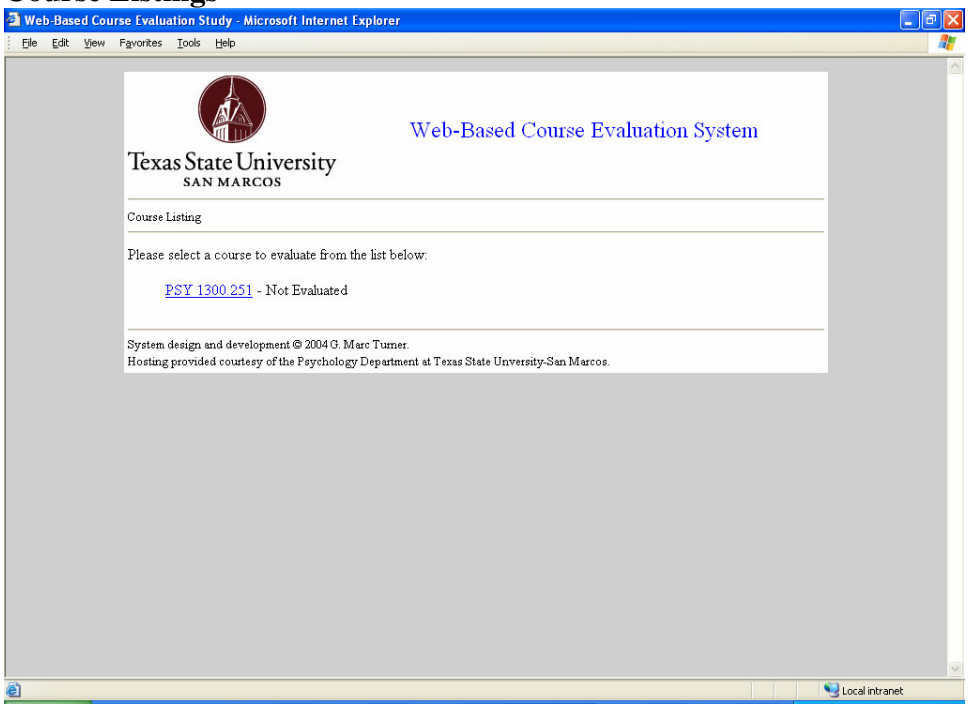
Welcome Screen



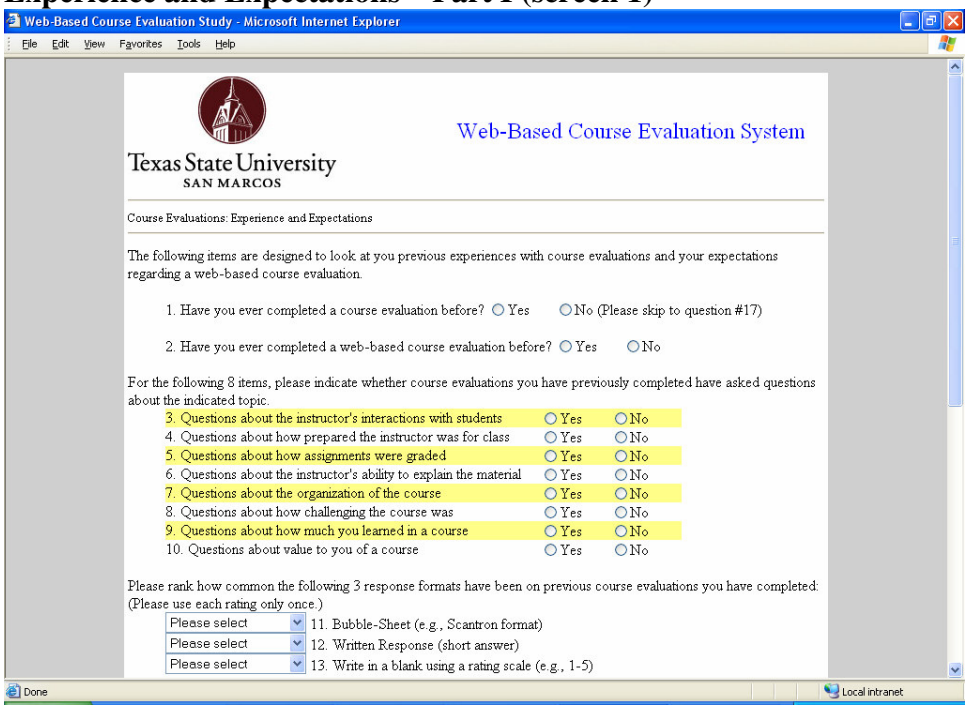
Sign-In Screen



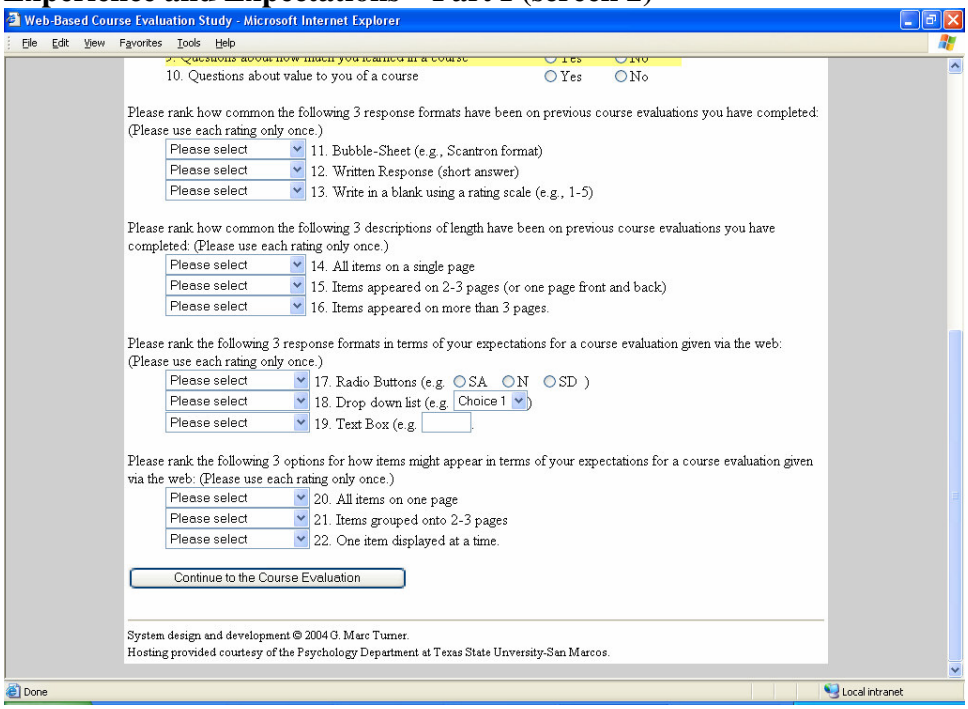
Course Listings



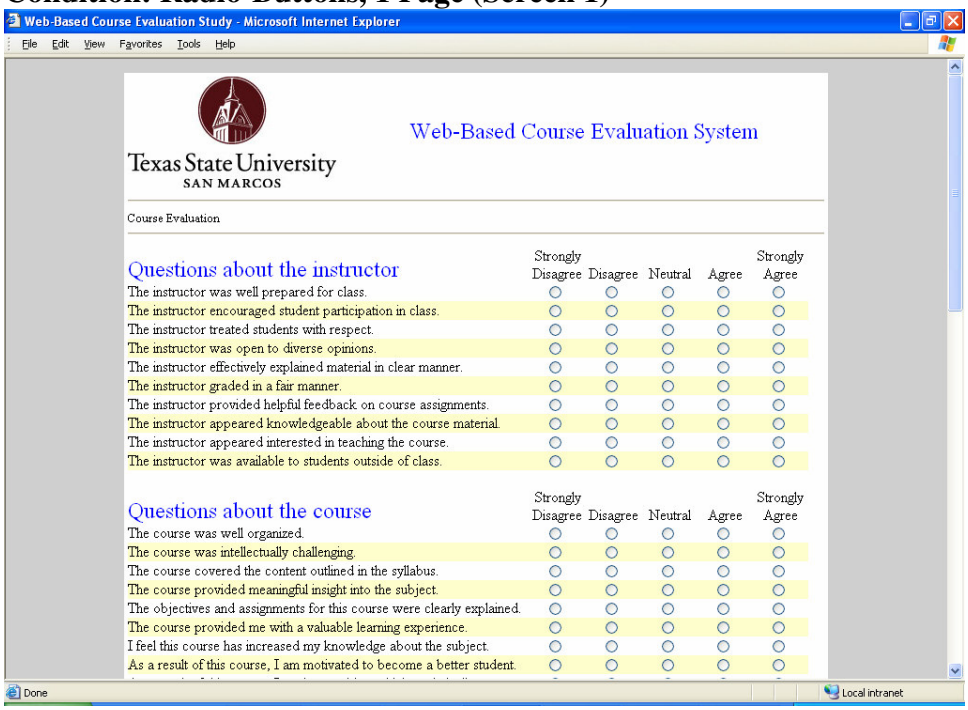
Experience and Expectations – Part I (screen 1)



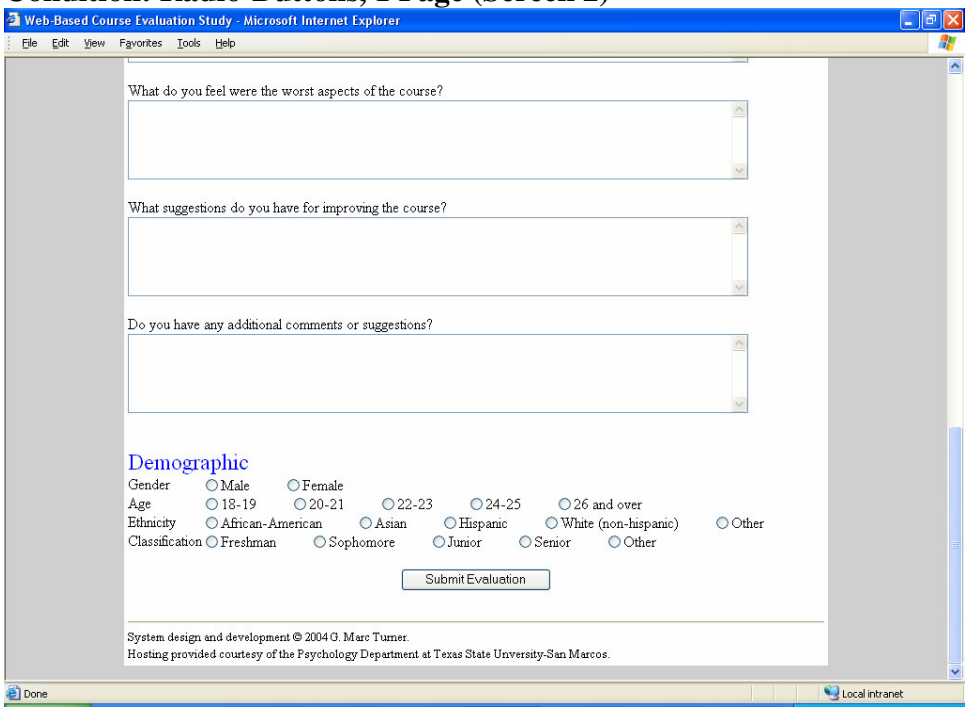
Experience and Expectations – Part I (screen 2)



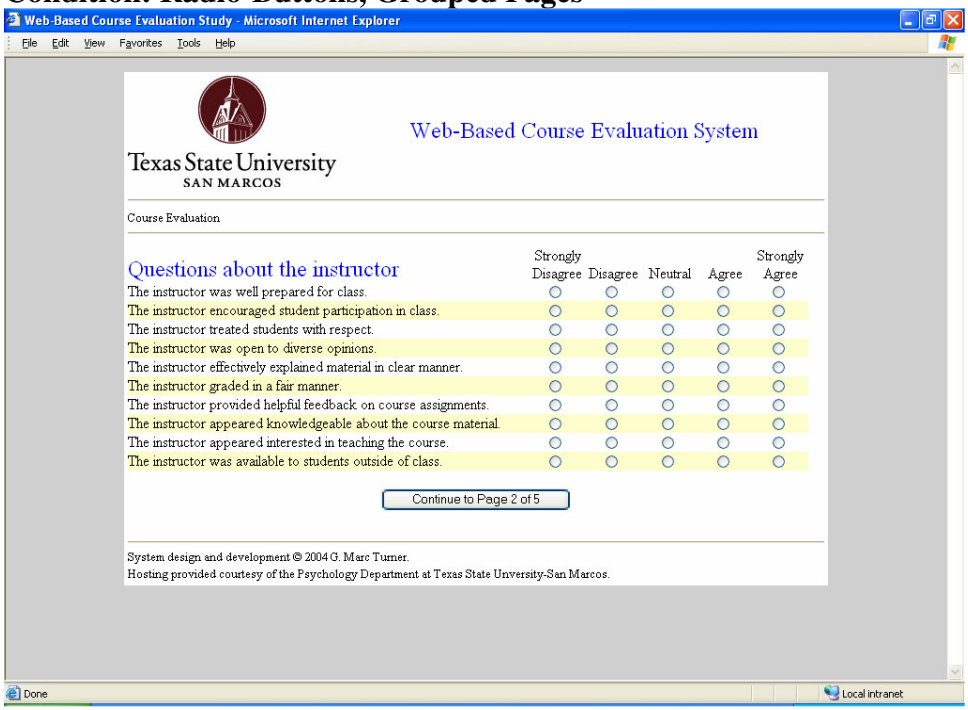
Condition: Radio-Buttons, 1 Page (Screen 1)



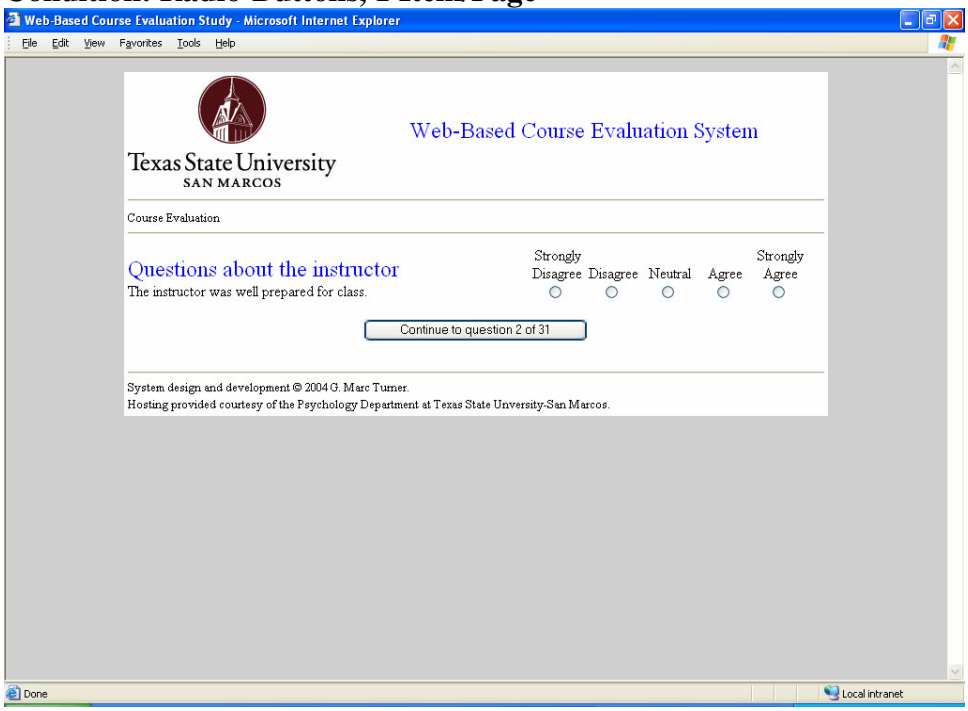
Condition: Radio-Buttons, 1 Page (Screen 2)



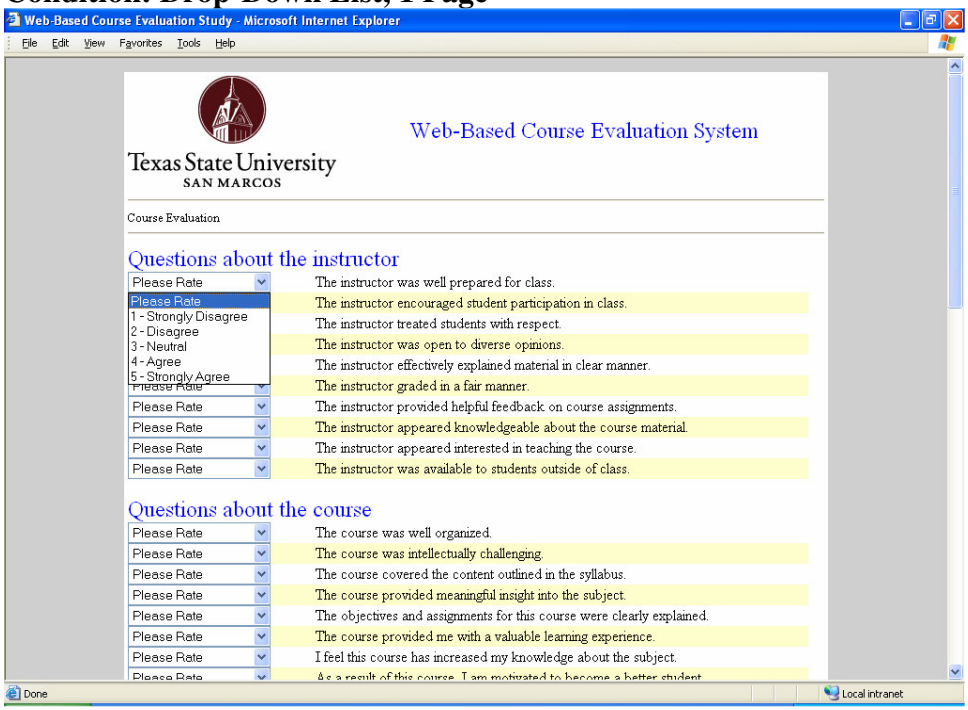
Condition: Radio-Buttons, Grouped Pages



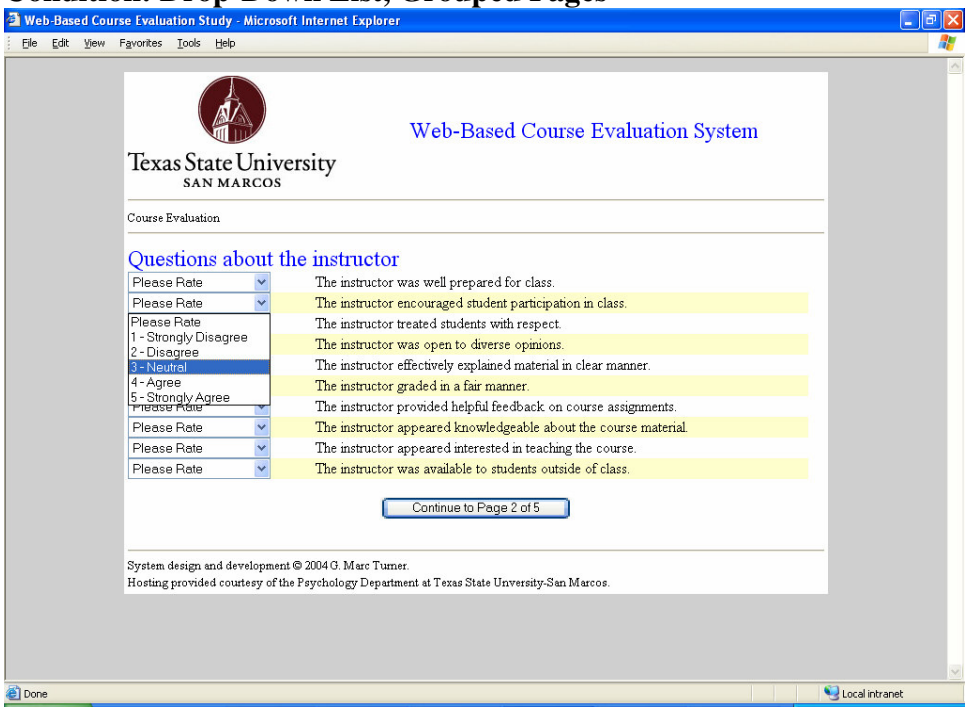
Condition: Radio-Buttons, 1 Item/Page



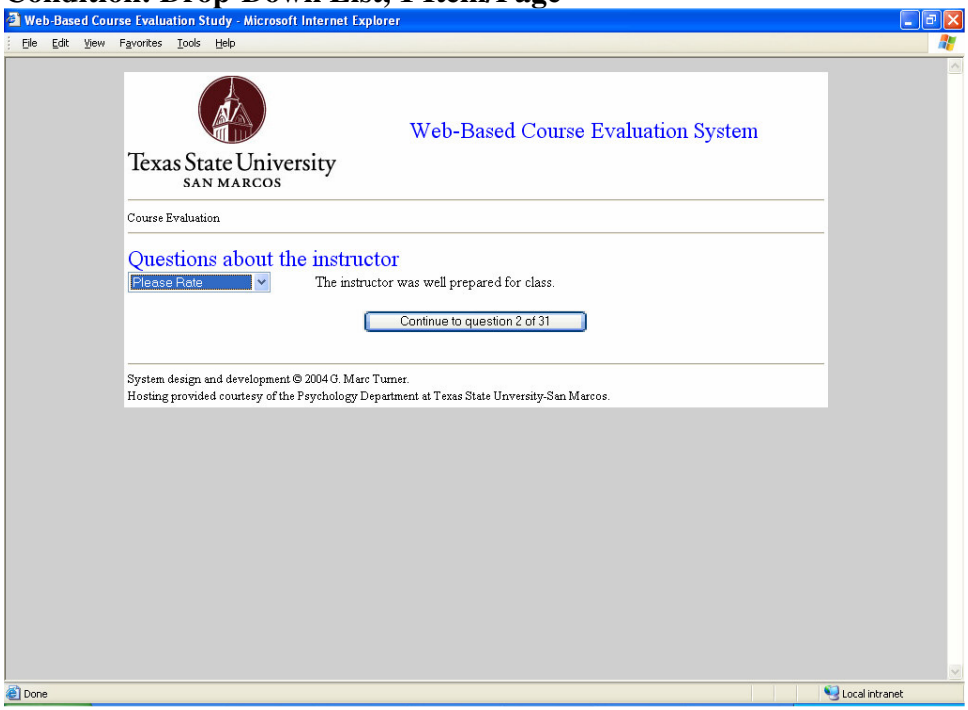
Condition: Drop-Down List, 1 Page



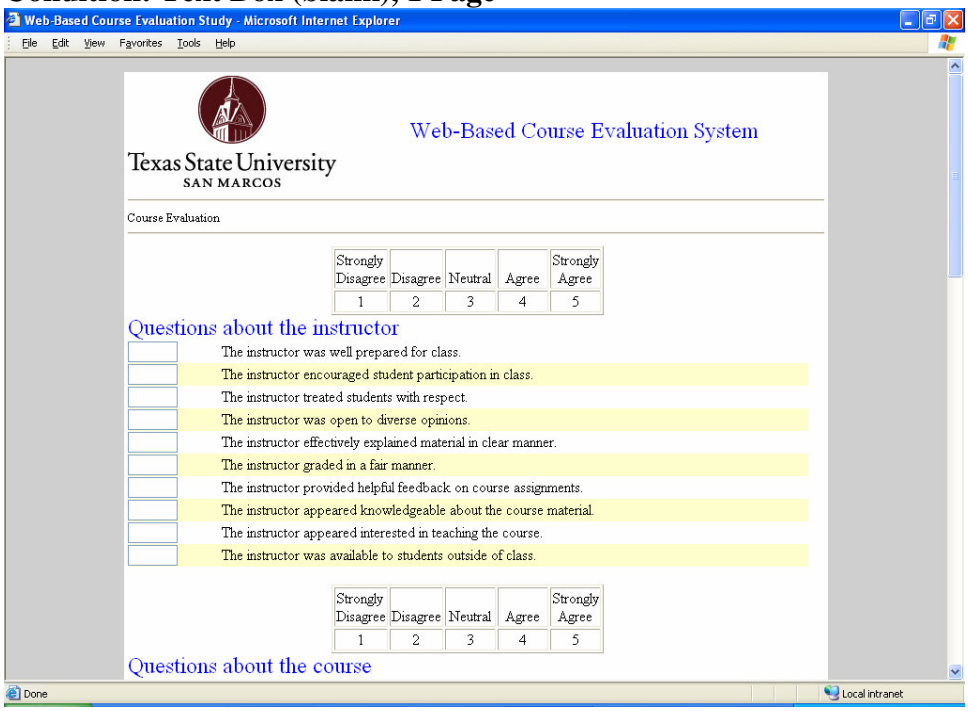
Condition: Drop-Down List, Grouped Pages



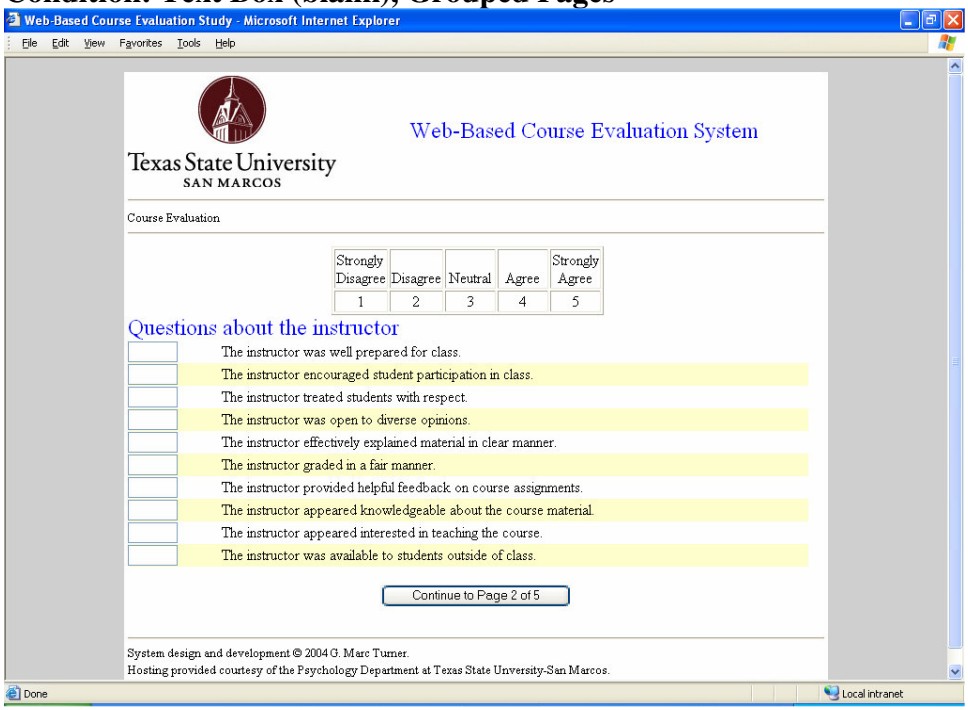
Condition: Drop-Down List, 1 Item/Page



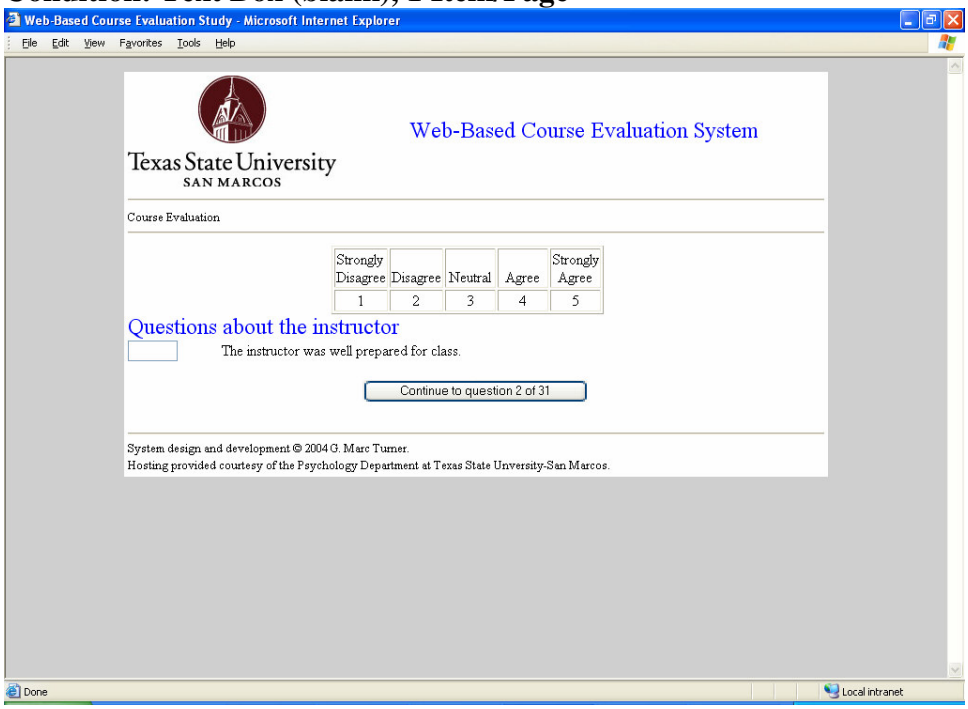
Condition: Text Box (blank), 1 Page



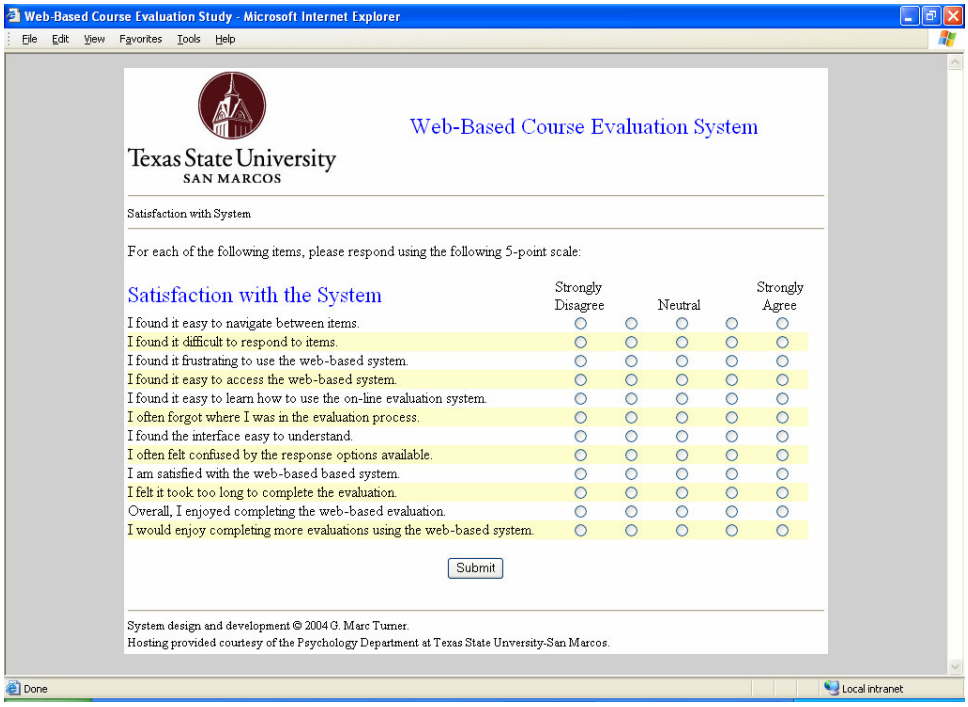
Condition: Text Box (blank), Grouped Pages



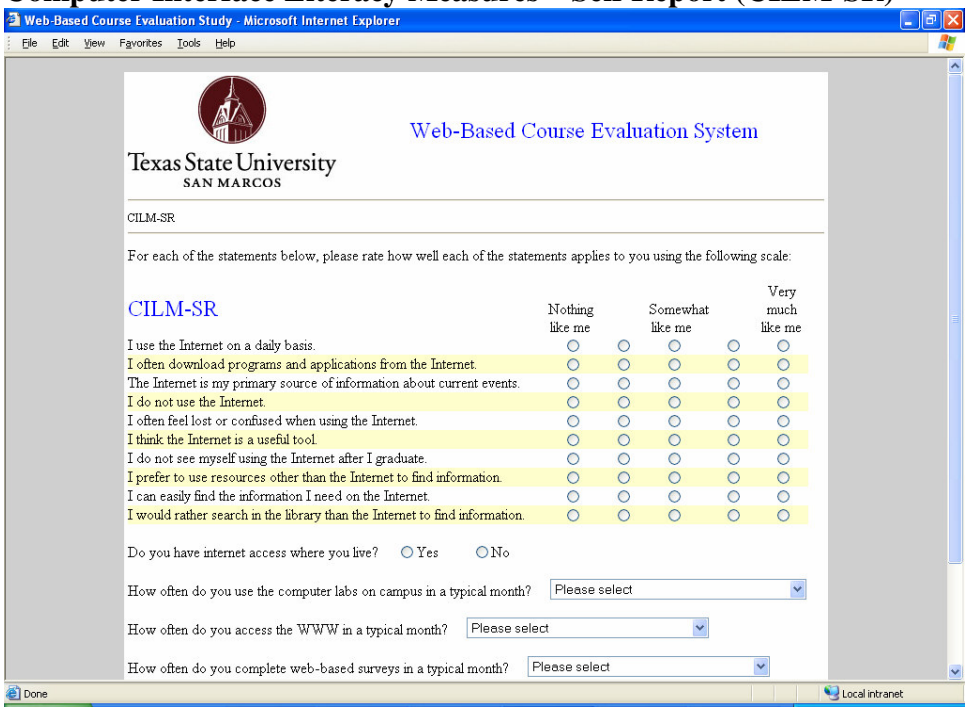
Condition: Text Box (blank), 1 Item/Page



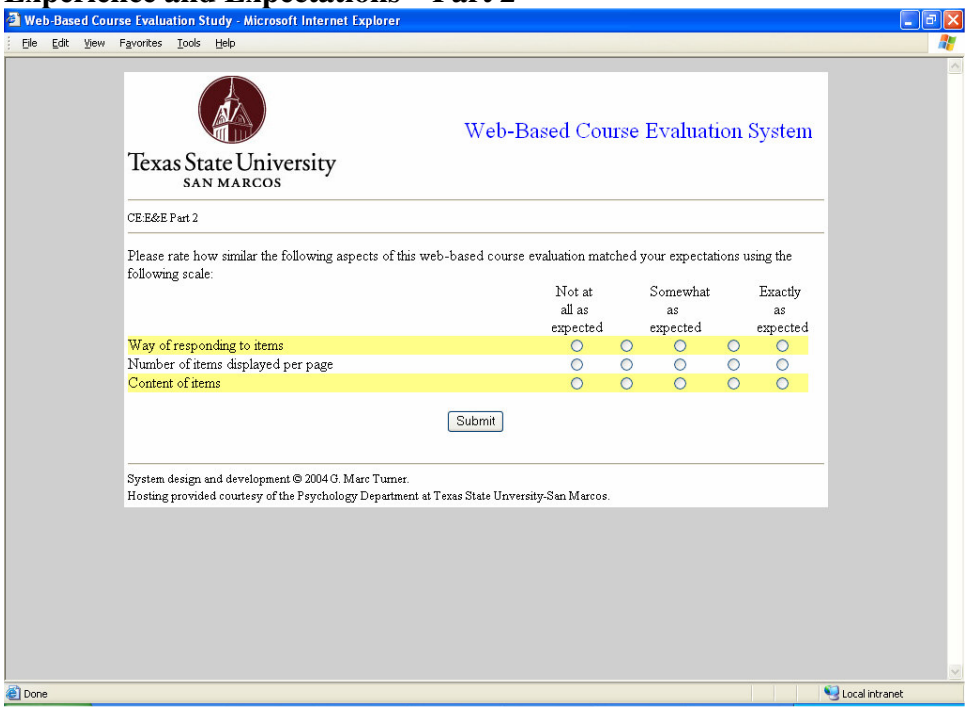
Satisfaction Measure



Computer Interface Literacy Measures – Self Report (CILM-SR)



Experience and Expectations – Part 2



REFERENCES

Andersen, K. & Miller, E. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics, 30*(2)*,* 216-219.

Barker, P., van Schaik, P., Hudson, S.R.G. and Tan, C.M., (1998). Mental Models and Their Role in Teaching and Learning, in Ottman, T. & Tomek, I. (Eds.) *Proceedings of ED-MEDIA & ED-TELECOM 98 - 10th World Conference on Educational Multimedia, Hypermedia and Telecommunications, June 20-25, Freiburg*, Germany, Charlottesville, Virginia, USA: Association for the Advancement of Computing in Education, 105-110.

Barker, P. (1999). Mental Models and Network Pedagogy. Enable99 Enabling Network-Based Learning. http://www.enable.evitech.fi/enable99/papers/barker/barker.html

Barker, P.G. and van Schaik, P. (1999). Mental Models and Their Implications for the Design of Computer-Based Learning Resources, *Proceedings of the 4th International Conference on Computer-Based Learning in Science (CBLIS '99),* University of Twente, Enschede, The Netherlands, 1-6 July, 1999.

Basow, S. & Spilberg, N. (1987) Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308-314.

Braskamp, L. & Ory, J. (1994). *Assessing Faculty Work: Enhancing individual and institutional performance*. San Francisco, CA: Jossey-Bass.

Booth-Kewley, S., Edwards, J., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77*(4), 562-566.

Boyer, E. (1990). *Scholarship reconsidered: Priorities of the professorate*. San Francisco: CA, The Carnegie Foundation for the Advancement of Teaching.

Bransford, J. (1979). *Human cognition: Learning, understanding, and remembering*. Belmont, CA: Wadsworth Publishing Company.

Brandsford, J. (1985). Schema activation and schema acquisition. In H. Singer & R.B. Ruddell (Eds.), *Theoretical models and processes of reading*, 3rd ed. Newark, DE: International Reading Association, 385-397.

Cantor, N. & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality & Social Psychology*, *35*(1), 38-48.

Cardy, R., Bernardin, H., Abbott, J., Senderak, M., & Taylor, K. (1987). The effects of individual performance schemata and dimension of familiarization on rating accuracy. *Journal of Occupational Psychology, 60*(3), 197-205.

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

Chalmers, P. (2003). The role of cognitive theory in human-computer interface. *Computers in Human Behavior, 19*(5)*,* 593-607.

Chang, L. & Hocevar, D. (2000). Models of Generalizability Theory in analyzing existing faculty evaluation data. *Applied Measurement in Education, 13*(3), 255-276.

Chinnappan, M. (1998). Schemas and mental models in geometry problem solving. *Educational Studies in Mathematics, 36*(3), 201-217.

Chonko, L., Tanner, J. & Davis, R. (2002). What are they thinking? Students' expectations and self-assessments. *Journal of Education for Business, 77*(5), 271-280.

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

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

Couper, M. (2000). Web Surveys. *Public Opinion Quarterly*, 64(4), 464-494.

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

Dahlin, B. (2001). Critique of the Schema Concept. *Scandinavian Journal of Educational Research, 45*(3), 287-300.

Davis, C. & Cowles, M. (1989). Automated psychological testing: Method of administration, need for approval, and measures of anxiety. *Educational & Psychological Measurement*, *49*(2), 311-320.

Dillman, D. (2000). *Mail and Internet surveys: the tailored design methods*. Second edition. New York: John Wiley & Sons, Inc.

Dillman, D. & Bowker, D. (2001). The Web questionnaire challenge to survey methodologists. In Reips, U. & Bosnjak, M. (Eds.), *Dimensions of Internet Science*. Lengerich: Pabst Science Publishers.

Dillman, D. & Christian, L. (2002). The influence of words, symbols, numbers, and graphics on answers to self-administered questionnaires: results from 18 experimental conditions. Available on-line: http://survey.sesrc.wsu.edu/dillman/papers/single_space_fig_table.pdf [last accessed 9-11-2004].

Dillman, D., Tortora, R., & Bowker, D. (1998). Principles for constructing Web surveys. SESRC Technical Report 98-50, Pullman, Washington Available on-line: http://survey.sesrc.wsu.edu/dillman/papers/websurveyppr.pdf [Last accessed on October 11, 2004]

Dommeyer, C., Baum, P., Chapman, K., & Hanna, R. (2002). Attitudes of business faculty towards two methods of collecting teaching evaluations: Paper vs On-Line. *Assessment & Evaluation in Higher Education, 27*(5), 455-462.

Dwight, S. & Feigelson, M. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational & Psychological Measurement, 60*(3), 340-360.

Erdman, H., Klein, M., Greist, J. (1983). The reliability of a computer interview for drug use/abuse information. *Behavior Research Methods & Instrumentation, 15*(1), 66-68.

Evan, W., & Miller, J. (1969). Differential effects on response bias of computer vs. conventional administration of a social science questionnaire. *Behavioral Science, 14*(3), 216-227.

Eysenck, M. W. (1993). *Principles of Cognitive Psychology*. Hillsdale, NJ: Lawrence Erlbaum.

Faulkner, C. (1998). *The Essence of Human-Computer Interaction*. Prentice-Hall, London.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127-148.

Feldman, K. (1976). Grades and college students' evaluation of their courses and teachers. *Research in Higher Education, 4*(1), 69-111.

Feldman, K. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, *10*(2), 149-172.

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

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

Fiore, S., Cuevas, H., & Oser, R. (2003). A picture is worth a thousand connections: The facilitative effects of diagrams on mental model development and task performance. *Computers in Human Behavior*, *19*(2), 185-199.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*. The Hague, The Netherlands. April 1-6, 2000. [Available on-line: www.akira.ruc.dk/~mhz/Research/ Publ/CHI2000_preprint.pdf]

Goldman, L. (1985). The betrayal of the gatekeepers: Grade inflation. *Journal of General Education, 37*(2)*,* 97-121.

Green, A. J. & Gilhooly, K. J. (1992). Empirical advances in expertise research. In: Keane, M. T. & Gilhooly, K. J. (Eds.). *Advances in the psychology of thinking (Vol. I).* London: Harvester Wheatsheaf.

Greenwald, A. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186.

Greenwald, A. & Gillmore, G. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.

Hall, R. (2004). *Web Objects and User Schemas*. Available on-line: http://web.umr.edu/~rhall/commentary/web_objects.htm [last access 9/15/2004]

Hart, R. & Goldstein, M. (1985). Computer-assisted psychological assessment. *Computers in Human Services, 1*(3), 69-75.

Heerwegh, D. & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in Web surveys. *Social Science Computer Review, 20*(4), 471-484.

Hicken, S. (1991). Learning in tasks that incorporate natural versus unnatural mappings: An example using aircraft instrument comprehension. *Educational Technology Research & Development*, *39*(1), 63-71.

Hobson, S. & Talbot, D. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49*(1), 26-31.

International Organization of Standards (1998). *ISO 9241-11 Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability.* Geneva, Switzerland: International Organization of Standards.

Kantor, J. (1991). The effects of computer administration and identification on the Job Descriptive Index (JDI). *Journal of Business and Psychology, 5*(3)*,* 241-255.

Katz, L. & Dalby, J. (1981). Computer-assisted and traditional psychological assessment of elementary-school-aged children. *Contemporary Educational Psychology, 6*(4), 314-322.

Keisler, S. & Sproull, L. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50*(3), 402-413.

Kember, D., Leung, D., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education*, *27*(5), 411-425.

Kerin, R., Peterson, R. & Martin, W. (1975). Teaching effectiveness: How do we rate? *Combined Proceedings: American Marketing Association,* 694-698.

King, W. & Miles, E. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*(6)*,* 643-651.

Kirakowski, J. (2003). *Background notes on the SUMI questionnaire*. Human Factors Research Group University College Cork, Ireland. Originally 1994. WWW

version http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html. Feb. 2003. [Last Accessed 9-10-2004.]

Koch, W., Dodd, B., & Fitzpatrick, S. (1990). Computerized adaptive measurement of attitudes. *Measurement and Evaluation in Counseling and Development, 23*(1), 20-31.

Kolitch, E. & Dean, A. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies of Higher Education, 24*(1), 27-42.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311-328.

Krallman, D., & Holcomb, T. (1997). *First-year student expectations: Pre- and post-orientation.* Paper presented at the Annual Meeting of the Association of Institutional Research, Buena Vista, FL.

Langbein, L. (1994). The validity of student evaluations of teaching. *PS: Political Science and Politics, 27*(3)*,* 545-553.

Lautenschlager, G. & Flaherty, V. (1990). Computer administration of questions: More desirable or more social desirability? *Journal of Applied Psychology, 75*(3), 310-314.

Layne, B., DeCristoforo, J., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*(2), 221-232.

Lozar-Manfreda, K., Batageli, Z., & Vehovar, V. (2002). Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer Mediated Communication, 7*(3), [Available on-line only at: http://www.ascusc.org/jcmc/vol7/issue3/vehovar.html].

Lozar-Manfreda, K. & Vehovar, V. (2002). Survey Design Features Influencing Response Rates in Web Surveys. Paper presented at International Conference on Improving Surveys. [Available on-line at: http://www.websm.org/uploadi/editor/survey design features influencing.pdf]

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

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

Marsh, H. & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.) *Higher Education: A Handbook of Theory and Research, Vol. 8*. New York: Agathon.

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

Martin, C. & Nagao, D. (1989). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology, 74*(1), 72-80

Matheson, K. & Zanna, M. (1988). The impact of computer-mediated communication on self-awareness. *Computers in Human Behavior, 4*(3), 221-233.

McKeachie, W. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.

Mead, A. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Millea, M. & Grimes, P. (2002). Grade Expectations and Student Evaluation of Teaching. *College Student Journal, 36*(4), 582-590.

Obenchain, K., Abernathy, T., & Wiest, L. (2001). The Reliability of students' ratings of faculty teaching effectiveness. *College Teaching, 49*(3), 100-104.

Park, O., & Gittelman, S. (1995). Dynamic characteristics of mental models and dynamic visual displays. *Instructional Science*, *23*(5-6), 303-320.

Piccinin, S., Cristi, C., & McCoy, M. (1999). The impact of individual consultation on student ratings of teaching. *International Journal for Academic Development, 4*(2), 75-88.

Potosky, D. & Bobko, P. (1997). Computer versus paper-and-pencil administration mode and response distortion in noncognitive selection tests. *Journal of Applied Psychology, 82*(2), 293-299.

Price, E. & Driscoll, M. (1997). An inquiry into the spontaneous transfer of problem-solving skill. *Contemporary Educational Psychology, 22*(4), 472-494.

Ramsden, P & Martin, E. (1996). Recognition of good university teaching: Policies from an Australian study. *Studies in Higher Education, 21*(3), 299-316.

Reips, U. (2002). Context effects in web-surveys. In Batinic, B., Reips, U., Bosnjak, M., & Werner, A. (Eds.) *Online Social Sciences*. Seattle: Hogrefe & Huber.

Rezmovic, V. (1977). The effects of computerized experimentation on response variance. *Behavior Research Methods & Instrumentation, 9*(2), 144-147.

Rosenfeld, P., Booth-Kewley, S., Edwards, J., & Thomas, M. (1996). Responses on computer surveys: Impression management, social desirability, and the Big Brother syndrome. *Computers in Human Behavior, 12*(2), 263-274

Rosenfeld, P., Booth-Kewley, S., & Edwards, J. (1993). Computer-administered surveys in organizational settings: Alternatives, advantages, and applications. *American Behavioral Scientist*, *36*(4), 485-511.

Rozensky, R., Honor, L., Rasinski, K., Tovian, S., & Herz, G. (1986). Paper-and-pencil versus computer-administered MMPIs: A comparison of patients' attitudes. *Computers in Human Behavior, 2*(2), 111-116.

Scissons, E. (1976). Computer administration of the California Psychological Inventory. *Measurement and Evaluation in Guidance, 91*(1), 22-25.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, and M. Scriven, (Eds). *Perspectives of Curriculum Evaluation*, Chicago: Rand-McNally, 39-83.

Shelvin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of Student Evaluations of Teaching in Higher Education: Love me, Love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405.

Skinner, H. & Allen, B. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting & Clinical Psychology, 51*(2), 267-275.

Slack, W., & Slack, C. (1977). Talking to a computer about emotional problems: A comparative study. *Psychotherapy: Theory, Research, and Practice, 14*(2), 156-164.

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

Spencer, K. & Schmelkin, L. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education, 27*(5), 397-409.

Sproull, L. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal, 29*(1), 159-169.

Stanfel, L. (1991). An experiment with student evaluations of teaching. *Journal of Instructional Psychology, 18*(1), 23-28.

Stanfel, L. (1998). Accuracy of teaching evaluation information systems: A case study with analyses and redesign. *Journal of Instructional Psychology, 25*(2), 82-93.

Stanton, J. (1998). An empirical assessment of data collection using the internet. *Personnel Psychology, 51*(3), 709-725.

Steinbart, P. (1989). A comparison of computer-administered and paper questionnaires: Effects on student evaluations of teaching effectiveness. *Issues in Accounting Education, 4*(1), 120-125.

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

Stumpf, S. & Freedman, R. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71*(3)*,* 293-302.

Thorpe, S. (2002). Online student evaluation of instruction: An Investigation of non-response bias. Paper presented at the 42nd Annual Forum of the Association for Institutional Research. Toronto, Canada. Available on-line at: http://www.drexel.edu/provost/ir/conf/bias.pdf [Last Accessed on 9-17-2004]

Turner, G., Sweeney, N., & Husman, J. (2000). Development of the Computer Interface Literacy Measure. *Journal of Educational Computing Research, 22*(1), 37-54.

White, D., Clements, C., & Fowler, R. (1985). A comparison of computer administration with standard administration of the MMPI. *Computers in Human Behavior, 1*(2), 153-162.

Williams, W. & Ceci, S (1997) How'm I doing? Problems with students' ratings of instructors and courses. *Change, 29*(5), 12-23.

Wilson, F., Genco, K., & Yager, G. (1985). Assessing the equivalence of paper-and-pencil vs. computerized tests: Demonstration of a promising methodology. *Computers in Human Behavior, 1*(3-4), 265-275.

Wilson, W. (1999). Students rating teachers. *The Journal of Higher Education*, *70*(5), 562-571.

Winn, W. & Snyder, D. (1996). Cognitive Perspectives in Psychology. In D H. Jonassen (Ed), *A Handbook of Research for Educational Communications and Technology* New York: Simon and Schuster MacMillan, pp. 112-142.

VITA

George Marcus Turner was born in Longview, Texas on December 15, 1972, the son of Mary Elizabeth Turner and John Wayne Turner. After completing his work at Trinity School of Texas, Longview, Texas, in 1991, he entered Trinity University in San Antonio, Texas. He completed his Bachelor of Science degree in Psychology at Texas State University-San Marcos in San Marcos, Texas in May of 1995. In August of 1995 he entered the Graduate School of The University of Texas in pursuit of his Doctor of Philosophy in Educational Psychology.

Publications

Ogletree, Shirley, M., Turner, G. Marc, Desai, R., Hernandez, R, & Stafford, T. (2003). Attitudes toward housecleaning: Gender-related differences. *American Association of Behavioral and Social Sciences Journal*, *6*, 82-88.

Ginsburg, H.J., Ogletree, S. M., Silakowski, T.D., Bartels, R.D., Burk, S. L, & Turner, G. M. (2003). Young children's theories of mind about empathic and selfish motives. *Social Behavior and Personality*, *31*, 237-244.

Turner, G., Sweaney, N., & Husman, J. (2000). Development of the computer interface literacy measure. Journal of Educational Computing Research, 22(1), 37-54.

Permanent Address: 8303 Red Willow Drive, Austin, Texas 78736

This dissertation was typed by the author.