2009

The Impact of Questionnaire Design on Response Times and Responses to Questions

Marc Frey
University of Windsor

The Impact of Questionnaire Design on Response Times and Responses to Questions

by

Marc Frey

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

Windsor, Ontario, Canada
2009

Author's Declaration of Originality

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

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

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

Abstract

In the past, it has been noted that minor changes in questionnaire design can influence the responses submitted (Jackson, Ing, & Arseneault, 2007). This study looked to evaluate whether variations in item wording or response scale characteristics would influence the way individuals cognitively process and respond to questionnaires. To facilitate this a 3x2x2x2 randomized by repeated measure experimental design was implemented, where scale characteristics and item wording were manipulated. Multiple Analysis of Variance tests were conducted, and it was noted that variations in scale characteristics and item wording resulted in differences in cognitive processing as well as the responses submitted. Questionnaire characteristics interacted with the type of experiences being evaluated, suggesting that some experiences result in different types of cognitive processing than others. The results from this study suggest that researchers should be careful when creating questionnaires, as subtle variations can alter the way individuals process and respond to items.

TABLE OF CONTENTS

The Impact of Questionnaire Design on Response Times and Responses to Questions

When conducting social science research, self-report measurement tools are often required to obtain data that would otherwise be challenging to quantify. In addition to adapting previously developed questionnaires, researchers often create and use their own measurement tools. Furthermore, questionnaire data has become rather ubiquitous in social science research. Nearly every area of psychological inquiry makes use of questionnaire data to investigate and draw inferences based on the responses submitted by participants. As a result questionnaire construction has sweeping impact in both research and applied contexts, including but not limited to: health, industrial/organizational, and clinical settings (Eccleston, McCracken, Jordan, & Sleed, 2007; Hoffman, Blair, Meriac, & Woehr, 2007; Osberg, Haseley, & Kamas, 2008). One specific area where questionnaires are often used exclusively to assess individual opinions, is student evaluations of teacher effectiveness in post secondary education centers. This measurement is of critical importance in academic settings as it is often used for promotion and tenure decisions involving academic faculty (Chen & Hoshower, 2003; Zabaleta, 2007).

Due to the prevalence of questionnaire data in both applied and research settings, one would assume that ample research and thought goes into the creation, construction and revision of such measures. Despite past publications outlining practices in questionnaire development (Gray & Guppy, 1999; Rea & Parker, 1997), there are numerous ongoing issues in questionnaire design that require further investigation. In addition, some decisions made by researchers when developing questionnaires seem to occur due to convention, despite more recent evidence that such selections may be

detrimental to the accuracy, reliability and validity of the questionnaire (Schriesheim, Eisenbach, & Hill, 1991). As a result, further research is necessary to clarify which practices in questionnaire development work and which do not.

This study has been developed with the intention to evaluate and further understand the design characteristics of questionnaires, and to observe interactions between design characteristics, which may influence results. Specifically, the impact of variations in item wording, response anchor orientation, and scale length were evaluated. Furthermore, this study was designed to clarify the influence survey design variations have on the cognitive processes involved in responding to items. This was accomplished by experimentally manipulating the design characteristics of student evaluations of teaching effectiveness (SETE) scales. Use of this measurement tool allows for real world evaluation of the variables, in a setting that may benefit from the findings. Due to the nature of this research endeavor, literature involving the SETE, scale construction and cognitive psychology as it applies to scale construction will be reviewed. This will be done to allow for a better understanding of all facets involved in this research project, as well as assisting in the application of the results from this study.

Student Evaluations of Teaching Effectiveness

As of late, student evaluations of teaching effectiveness (SETE) have garnered large amounts of acceptance and usage within many academic institutions. This vast support has taken a nearly global perspective with academic institutions worldwide administering student evaluations of teaching effectiveness (Chen & Hoshower, 2003; Moore & Kuol, 2005; Zabaleta, 2007). SETE are used to provide instructors with information from students regarding their teaching effectiveness in a class setting. This is completed with the hope that the instructor will take the feedback provided and attempt to

improve their teaching skills, based upon suggestions and scoring in the individual categories. They are also used for personnel related decisions by academic institutions. By providing numerical representation of evaluations of teaching effectiveness, institutions are able to quantify instructor ability and make better judgments of the instructor's effectiveness in classroom settings. These are called, respectively, formative and summative evaluations (Chen & Hoshower, 2003; Crumbley, Henry, & Kratchman, 2001; Moore & Kuol, 2005; Sedlmeier, 2006; Zabaleta, 2007).

In many cases SETE ratings assist in administrative decisions that may be irreversible and influential in future academic employment opportunities. In addition, SETE provide useful information to both instructors and institutions regarding: curriculum development and revisions; as well as assisting in resource allocations (Crumbley et al., 2001). The benefits of SETEs have also been extended to identifying and assisting groups of students who are facing similar difficulties, which may otherwise have gone unnoticed (Moore & Kuol, 2005). With these functions in mind it is easy to see the plethora of benefits that SETE provide, as well as the great importance they have in academic settings.

Despite the usefulness of SETE, these measurement tools do seem to have some limitations. In particular, research has found that some sources of variability in SETE scores include: subject matter, students, instructor characteristics and class characteristics (Crumbley et al., 2001; Koh & Tan, 1997). It has been found that some teaching methods, which have been empirically supported but are not viewed favorably by students, may result in negative evaluations by students (Crumbley et al., 2001). As well, negative ratings may be even more critical if the student dislikes the course material that is taught (Crumbley et al., 2001). In addition to personal bias, it has been found that variations in

instructor type and class type can influence SETE scores. For instance, scores may differ based upon the age (Langbein, 1994), gender (Fandt & Stevens, 1991; Langbein, 1994; Lueck, Endres, & Caplan, 1993), and rank (Lueck et al., 1993) of the instructor. Furthermore, it has been found that smaller classes with higher attendance rates and greater frequency of responses provide for the most positive results (Koh & Tan, 1997).

Some researchers have found that student characteristics, as well as student preconceptions may bias the results of SETE scores. It has been found that students are more likely to endorse positive ratings for the instructor, when the instructor grades "leniently". As well, students who receive higher grades have been found to provide more favorable SETE scores for instructors than those students who received lower grades (Perkins, Guerin, & Schleh, 1990). These results suggest that students may view SETEs as a reciprocal and subjective scoring procedure as opposed to an objective evaluation of instructor performance. On the contrary it has been argued by Marsh and Roche (2000) that there are better explanations than grading "leniency", when examining student responses. Specifically, the authors found that student responses were not related to grades. Instead they note that the relationship between grades and student responses was largely a function of perceived learning, prior interest in the subject and the level of the course being evaluated.

Recently it has been found that variations in scale properties have an effect on instructor performance ratings submitted by students. In fact, even seemingly harmless variations in questionnaire construction can result in significantly different SETE scores. In some cases it has been found that variations in item wording and response scale formating can result in appreciably different rating scores provided by students (Jackson, Ing, & Arseneault, 2007a). In addition it has been noted by Sedlmeier (2006) that the

respondent's certainty in their answers may moderate the influence of response anchor characteristics (i.e. wording or orientation) on students' responses; where greater certainty in responses results in decreased influence from scale characteristics on responses.

The environment under which teacher evaluations take place may also be a concern in SETE development and administration. It has been noted by Barnette (2000) that educational assessment questionnaire settings may be prone to apathetic cognitive processing by participants. Lack of effort on the part of respondents could theoretically be due to the lack of guidance, motivation and or incentive for responding (Barnette, 1999). One large issue related to this is that the lack of cognitive effort by respondents may increase the threat of acquiescence bias (or "yea saying") by participants. This issue seems to be inflated in educational settings, and thusly it has been suggested that development of SETE questionnaires should take these concerns into consideration (Barnette, 1999; 2000; Knowles & Condon, 1999; Schriesheim & Hill, 1981).

Due to the nature of students' perceptions of instructor performance, it is typically assumed that multiple components play a role in the gestalt assessment of instructor performance. As a result, SETE surveys are often developed with the intention of measuring various factors that make up the construct of instructor performance. In an attempt to better understand the factors involved in SETE measures, numerous studies have incorporated factor analytic techniques. These pursuits have provided insight into the multidimensionality and the validity of SETE ratings (Burdsal & Bardo, 1986; Jackson et al., 1999; Marsh, 1991, 1984). Factor analyses conducted by Burdsal and Bardo (1986) and Jackson et al. (1999) noted six factors within the Student Perceptions of Teaching Effectiveness Scale (SPTE-I). The six factors found in these situations were

labeled: rapport with students; perceived course value; course organization and design; grading fairness; course difficulty; and workload.

Recent Investigations of Questionnaire Development using SETE

Due to the importance of SETE measurement tools and the lack of standardization in their design and usage, researchers have begun to use SETEs to evaluate how differences in questionnaire construction can cause differential responses by participants (Arseneault & Jackson, 2005; Barnette, 1999, 2000; Ing & Jackson, 2007; 2006; 2008). Recent research has examined the influence that response anchor type has on participant responses, while making use of SETEs. Initially, Arseneault and Jackson (2005) observed differences in responding based on the response anchors (agreement or evaluation anchors) and instructions (to submit an opinion or evaluation) that were presented to participants. Within their study it was noted that agreement response anchors elicited more negative responses from participants.

Ing and Jackson (2006, 2007) completed two studies in an attempt to replicate the findings from Arseneault and Jackson's (2005) prior study. In the first study, participants were also provided instructions prior to engaging in the questionnaire to either evaluate or give an opinion regarding teacher effectiveness. This study did not replicate the results of the initial study, but it was found that participants responded differently based upon the instructions they were provided (Ing & Jackson, 2006). In the second study, the separate factors of the SETE measure were used as dependent variables. The results of this study once again did not replicate the findings of Arseneault and Jackson's (2005) initial work. However, significant interactions were noted between questionnaire instructions and the SETE subscales.

Finally, Ing and Jackson (2008) attempted again to replicate the findings by Arseneault and Jackson (2005). In this attempt, an additional variable that consists of scale length was included in the design. Specifically, half of the participants were given a 5-point scale and the other half were presented with a 9-point scale. Although, this study did not replicate the past findings by Arseneault and Jackson (2005), interestingly enough there was found to be significant differences in participant responses based upon the length of the scale they were given. In particular, participants who were given a longer response scale provided more positive responses when compared to those who were given a shorter response scale. This finding may have relevancy to acquiescence bias (or "yea saying") in responding, as these results suggest that longer response scales may lead to inherently more positive responses from participants. Taken as a whole these studies provide evidence that not all questionnaire designs are equivalent.

Cognition of Responding

When responding to questionnaires, individuals undergo numerous cognitive processes that ultimately result in their reply (Tourangeau, 2003). The respondents' motivation to place cognitive effort in each component of the cognitive process can influence the responses given (Krosnick, 1991; Simon, 1979). Furthermore, it has been found that how questions are worded may influence the cognitive processes involved in responding, altering the content of the response that is submitted (Schriesheim & Hill, 1981). Researchers have attempted to find ways to quantify the cognitive processing involved in responding, with the hope of better understanding how aspects of questionnaire design influence the particular cognitive processes involved in responding.

In the context of questionnaire development and measurement strategies, researchers have become aware of the influence that underlying cognitive processes have

on responses provided by individuals. As such, many of the issues (and sources or error) when responding to items, are related to breakdowns in the underlying cognitive processes involved in responding (Tourangeau, 2003). Within the literature there is an agreed upon set of cognitive stages by which individuals generate responses to questions, illuminating areas where potential pitfalls may occur. This process is thought to begin when the individual reads and attempts to comprehend the question posed to them. Then the individual undergoes an information retrieval phase, which coincides with the question. Afterward, they formulate a judgment based upon the question and information possessed. Finally, they will attempt to encode their responses into the rubric (or responses options) provided by the measure (Schwarz, 1999, 2007; Sudman, Bradburn, & Schwarz, 1996; Tourangeau & Rasinski, 1988). Furthermore, each stage reflects an active process by the participant, where the cognitive resources used work in synchrony with one another to develop a response (Matlin, 2002; Tourangeau, 2003). By considering each cognitive step involved and the related issues at each stage of processing, ideally one may be able to generate a questionnaire that provides for optimal results.

Understanding the question. When considering the first stage of the response procedure (reading and comprehending the questions), how the questionnaire is worded and oriented can play an integral role in the final responses provided by individuals. The first concern the researcher may encounter at this stage is whether or not the respondent understands the questions in the same way that the researcher intended them to. If the item does not represent the same domain as the researcher had planned, the validity of the question may be called into question. Therefore, when considering how participants approach the first stage of responding, researchers would do well to generate clear and concise questions that are unambiguous to the respondent (Schwarz, 1999).

An additional concern that may arise at this stage of responding is the influence of positive or negative wording on the comprehension of the question. Within cognitive research it has been well established that human beings understand sentences and information better if it is presented in a positively worded fashion (Clark & Chase, 1972; Hearst, 1991; Matlin, 2002; Sherman, 1976; Wason & Jones, 1963; Williams, 1991). Research has found that when questions are phrased in a negative (no or not) or implied negative (denies or dislikes) fashion, respondents are less likely to understand the question posed to them (Sherman, 1976). Furthermore, this difficulty in understanding negatively worded questions is associated with a longer response time required to evaluate the questions presented (Clark & Chase, 1972; Williams, 1991). Despite these findings, researchers often utilize negatively worded items in questionnaires to curb potential response biases (such as acquiescence). However, it needs to be recognized that the inclusion of negatively worded items may inhibit an individual's ability to understand the question posed, and may result in less than accurate responses being submitted.

Another issue that may influence the way a respondent perceives a question is "context effects" of the questionnaire. Namely, individuals will often use previous questions presented when attempting to understand what a new question is asking of them. As a result, the individual may understand the question they are attempting to answer in a different way than the researcher had intended, based largely on the location of that item within the entire questionnaire. As well, contexts effects may have more broad implications, where participants respond to questions based upon previously inferred social norms within the questionnaire (Sudman et al., 1996). Therefore, researchers should attempt to create a coherent and relevant progression both within question wording and the location of questions within the questionnaire. It is important to

be aware of the influence context may have on participant responses and attempt to avoid such concerns when developing questionnaires (Schwarz, 1999).

Information retrieval. After an individual has understood the question appropriately, they are required to place cognitive effort into a retrieval process of confirming or disconfirming information relating to the question they are asked. Most researchers agree that this process involves a rigorous search of long-term memories that are relevant to the question at hand (Tourangeau & Rasinski, 1988). As well, the general consensus amongst memory researchers seems to be that memories are often cognitive reconstructions of events as opposed to exact replications (e.g., Matlin, 2002). Consequently, characteristics of the measure as well as cognitive heuristics used by respondents at this stage of the cognitive response procedure could alter the final response provided.

Williams and Hollan (1981) discuss this memory retrieval phase, as a multistage reconstructive process, which may be susceptible to errors or failure in memory recovery. This stage is thought to begin with an initial memory search based upon the descriptors presented in the question, then related memories are examined and a more detailed search is pursued until all necessary information is retrieved. However, due to the fact that memory is primarily a reconstructive endeavor, respondents may be prone to errors when attempting to derive their answers to questions. Participants may commit errors of commission or omission of information based upon how they are searching for relevant memories to answer the question. In particular, individuals will often organize and summarize their memories so that they fit in a coherent and logically consistent format. This manner of processing, based on mental shortcuts, may lead to errors and consequently inaccurate responses (Sudman et al., 1996).

As well, at this stage of responding, individuals who are not adequately motivated to engage in the task may be more inclined to rely on cognitive shortcuts (or heuristics). This can lead to mental shortcuts such as a confirmation bias or availability heuristic. When using a confirmation bias, the individual may become reluctant to actively pursue disconfirming evidence and instead rely primarily on confirming information to answer the question. As a result, participants may be more likely to merely agree with statements presented to them (Nickerson, 1998; Wason, 1960; Zuckerman, Knee, Hodgins, & Miyake, 1995). In the case of availability heuristics, individuals will use the most readily available related memory and only this information to base their decision. This may occur when individuals can easily recall a rare incident or when they have difficulty recalling frequently occurring events (Tversky & Kahneman, 1973). Beyond these specific instances of memory error, it is important for researchers to realize that memory can be influenced by the emotional state of the individual. As a result, questions that are worded in a particular tone within a questionnaire may not only influence how the individual understands the question, but also the memories they use to determine their response (Sudman et al.,1996).

Judgment formation. Once individuals have understood the question presented and have engaged in a thorough search for relevant information, they will undergo a judgment process. This stage of responding is very closely aligned with the relevant memories accumulated during the individuals' information search. In this particular stage the individual cognitively weighs the relevant memories and derives a sum or total. With this total in mind the individual will determine what position they will take when responding (Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000).

When individuals are generating a judgment, they may attempt to streamline the process by cognitively estimating what their response should be (Tourangeau et al., 2000). For example, when evaluating and recalling relevant memories, respondents may cognitively estimate their judgment by relying on generic information that is quickly recalled (Means & Loftus, 1991; Smith, 1991, as cited in Tourangeau et al., 2000). Blair and Burton (1987) provided evidence of this, when they noted that the more events necessary for recall, the more difficulty respondents had formulating their responses. The difficulty in evaluating large amounts of information may translate into decreased effort in the judgment phase of responding. In turn, respondents may be more likely to rely on cognitive estimation, as opposed to submitting the most accurate response.

Encoding of judgments. The final stage of the cognitive response process requires the individual to translate the judgment they have made into the scale provided. In particular, individuals must cognitively assess the meaning of the scale and then attempt to express themselves within the established parameters. As a result, the scale anchors or the amount of scale options provided in the questionnaire may impact the way the individual responds (Tourangeau et al., 2000).

Much like the wording of the question, if the response is not understood by the individual, or requires too much cognitive effort, they may opt to rely on shortcuts when encoding. It is conceivable that in these scenarios, participants may select the most socially acceptable or cognitively simple response. This issue is of particular relevance to SETEs as it has been noted that on average, student evaluations of teaching effectiveness are two scale points higher than the theoretical neutral point (when using 9-point scales) (Sears, 1983). As well it has been noted that longer scales (5-point scales vs. 9-point scales) contribute to more positive ratings in SETE (Ing & Jackson, 2008). Two potential

explanations come to mind when considering these results. Firstly, as noted by Zajonc (1968), humans seem to prefer positive opinions of their interpersonal relationships. As a result it is possible that when respondents are encoding judgments on SETEs, they opt to select more positive scale responses. Secondly, it is possible that increased scale lengths in SETE responding may require additional cognitive effort from respondents and as a result they opt to rely on cognitive shortcuts that ultimately lead to socially acceptable (or overtly positive) responding.

In many ways, the reply by the participant is constrained by the response options provided, such that they must contour their judgment into the limits of a scale. By allotting a numerical value to represent the opinion, much of the information possessed in the actual judgment can be lost. One obvious concern may be that individuals who do not have a strong response, or feel that they do not have a response (or that they "don't know") may have difficulty in transferring this judgment into a scale. Thus, it is possible that such judgments may be placed within the scale in such a way that the researcher is unable to discern what the respondent really meant (Beatty, Herrmann, Puskar, & Kerwin, 1998; Koriat & Goldsmith, 1994).

As well it has been noted that numerical values associated with response options may alter the way respondents perceive the scale provided. Specifically, if participants are provided a 0 to 10 scale it may be cognitively perceived as a progressive improvement where each scale point represents a more positive response. Alternatively a scale labeled from -5 to +5 may infer to participants that the low end of the scale is the polar opposite of the high end. Each type of numerical value may cognitively alter the way the individuals perceive the response and consequently how they choose to reply. Therefore,

it is important that researchers thoughtfully select the response options that they provide respondents (Schwarz, Knauper, Hipler, Noelle-Neumann, & Clark, 1991).

Despite all of the potential cognitive misplays that can arise when responding, it is important to realize that this is a remarkably efficient and intricate set of processes that take place in very short spans of time. Under optimal conditions, questionnaire responses can be relatively accurate (Schriesheim & Hill, 1981). As a result researchers are able to use questionnaire data to draw useful conclusions. However, in the worst-case scenarios, issues in the cognitive response process could lead to inaccurate, inconsistent and difficult to understand data (Clark & Chase, 1972; Schriesheim & Hill, 1981; Tversky & Kahneman, 1973; Williams, 1991). Therefore, in order to obtain optimal data, it is imperative that researchers are mindful of the cognitive processes individuals go through when responding to questions.

Motivation and its Impact on Cognition of Responses

When considering cognitive processing, it is important to take into account the impact motivation has on the cognitive effort put forth by respondents. Motivation is important because it influences effort output. If respondents are not motivated they will be less inclined to give the necessary cognitive effort in the response processes, consequently relying more heavily upon heuristics.

As human beings, we are often placed in situations where we have numerous demands and limited time and effort to give. As a means of achieving our goals we are required to prioritize and concentrate effort on the most salient tasks (Deci & Ryan, 1985). Extending from these observations related to motivation, Simon (1955; 1979) states that individuals may undergo what he terms as satisficing; where in an attempt to

conserve energy, tasks will be completed with the least amount of cognitive effort needed to adequately address the situation.

According to researchers, satisficing can occur even during the multiple cognitive stages involved in responding to questionnaires (Krosnick, 1991). Upon review of Krosnick's satisficing theory, individuals choose one of two response paths when selecting their reply to a question. The first route is referred to as "optimizing" and it is characterized by full engagement in the cognitive processes required to respond. This approach usually leads to the most accurate or "optimal" response. On the contrary, the second path is entitled "satisficing". When an individual is satisficing in the cognitive response processes, their final reply is often a superficial response that would appear reasonable or logical to an observer. Furthermore, satisficing is thought to be the product of decreased effort when responding; consequently the response submitted is not likely to reflect the individuals' actual opinions. Krosnick goes on to suggest that satisficing may be the result of: low motivation, low cognitive abilities and/or challenging tasks (such as difficult or poorly worded questions). Moreover, Krosnick (1991) notes that satisficing responses are often less reliable and less accurate as compared to optimized responses to questionnaires.

When examining a respondent's cognitive processing of individual questions, it is imperative to keep in mind the motivation they may or may not have when it comes to partaking in the questionnaire. This is necessary because individuals, who are not motivated to cognitively engage in the questions being asked, are not likely to provide the most accurate responses possible. Furthermore, such lack of motivation could lead to decreased cognitive effort in question responding and increased reliance upon response heuristics (such as acquiescence bias or "yea" saying).

Response Time and Response Cognition

Since the foundational years of psychology research, response time measurements have been used to make inferences about internal processes (Yan & Tourangeau, 2008). Furthermore, cognitive psychologists currently collect response time data to observe the intricacies of cognition (Matlin, 2002; Yan & Tourangeau, 2008). Survey researchers have become increasingly aware of the benefits response time data provides in understanding the cognitive processes involved in responding. Thus response time data may be a helpful ally in determining what constitutes ideal questionnaire characteristics (Bassili & Scott, 1996; Yan & Tourangeau, 2008). As well, some research has found that response times are better predictors of actual behavior than responses submitted (Bassili, 1993). As a result, response time to questions may be an important variable to consider in all questioning scenarios. Additionally, with the advent of computer generated testing, response time measures are much more easily collected and used by researchers.

Within survey research, response time data is often thought of as a proxy for other relevant variables in the cognitive stages of responding. Some researchers opt to think of response time as specific to recall ability or other individual characteristics that influence responding. While others have posited that response time latency may be related to questionnaire design characteristics (Yan & Tourangeau, 2008). One alternative approach to response time data is to consider it to be a proxy of cognitive effort put forth by the respondent. Within neuropsychological and cognition research settings, cognitive effort is often gauged using response time data, thus in survey responses this too may be applicable (Kellogg, 1986; Piolat, Olive, & Kellogg, 2005; Piolat, Olive, Roussey, Thunin, & Ziegler, 1999).

When examining response times, researchers have found that there are specific characteristics that may lead to shorter and longer response times. In particular, it has been noted that when individuals have stronger attitudes or more vivid memories associated with an occurrence, they often respond more quickly than those individuals with weaker attitudes (or less lucid recollections). As a result this has inspired some researchers to postulate that response time latency may be largely attributable to the memory retrieval process (Bassili & Roy, 1998). However, it is important to consider that the information retrieval stage can be expanded or constrained by the questionnaire design utilized.

Basilli and Scott (1996) have noted that response times significantly differ based upon the questionnaire's design characteristics and their influence on the cognitive processes inherent in responding. Although they do admit that memory retrieval does play a role in response time, their research suggests that response times can be strongly influenced by the wording of questions during the reading and understanding phase of responding. Specifically, it was found that poorly worded questions and double barreled questions generated longer response times from participants, as compared to the more concise and repaired versions of questions (Bassili & Scott, 1996). Additionally, Tourangeau, Rasinski, and D'Andrade (1991) have found that when questions of similar content are in contextual proximity, response time for the subsequent question decreases. As a result, it would appear that access to memories can be influenced by both the wording of the question as well as the items location in the questionnaire. Thus it has become apparent that response time may be used by researchers to determine questions or design choices that may challenge the individuals understanding of the questionnaire.

Additionally, Bassili and Scott (1996) found that unnecessary negative wording of items resulted in longer response times from participants, regardless of the question length. As well, it has been noted by Clark and Chase (1972) that when presented with positive information, individuals respond more quickly and with a lower error rate than when presented with negative information. When framed in the context of theory in cognitive psychology, this result parallels established literature where it has been noted that humans' cognitively process positively phrased information more efficiently than negatively phrased information. It is conceivable that respondents might take longer to respond to information that is phrased in a negative context, because more cognitive effort is required by the respondent to understand the question (Matlin, 2002; William, 1999). Thus phrasing of questions may be highly influential in the amount of elapse time required for individuals to respond.

As well, it is possible that the response options provided to individuals may influence the cognitive process, and result in increased (or decreased) response times. In particular it has been noted by Heerwegh and Loosveldt (2002) that the styles of response options (radio buttons vs. drop-down boxes) lead to differences in response times and in completion rates (where radio buttons had faster response times, and less attrition). As well it has been noted that response options that follow a logical order (e.g. from top to bottom) resulted in faster response times from participants, as compared to response options that did not follow a logical order (Tourangeau, Couper & Conrad, 2004). Consequently, it would seem that differences in response scale orientation might require increased cognitive effort on the part of respondents, resulting in increased response times.

Despite the overwhelming progress that has been made in understanding cognitive response processes through response time measurement, it is apparent that response time data is limited in some ways. Chiefly, response time data to date has only been used as an "elapse time" during the presentation of the question and the participants' response to the question. Whereby, a solitary numerical value is derived and used to understand a very intricate multi-stage set of cognitive processes (Yan & Tourangeau, 2008). As a result, only so much can be inferred from this piece of data, as there is no specific information that delineates cognitive possessing at each stage of responding. Researchers are currently limited to testing hypotheses by instituting controlled studies that manipulate specific stages of responding, and then checking for differences in the resulting response times. An additional limitation of response time data collection is the passive nature of response time measurement. As discussed by Yan and Tourangeau (2008), if researchers make participants aware that they are measuring response times, they may be priming participants and introducing demand characteristics. One of the most accessible ways of measuring response times is through the use of computer-generated questionnaires; which can gauge the elapse time of responding without the individual's awareness. Therefore, computer implementation of questionnaires is almost a necessity when collecting response time data. With these limitations in mind, if used appropriately response time data can be a very helpful tool in understanding the cognitive processes required when responding and in turn improving future questionnaire development.

Sources of Error and Biases

When developing questionnaires, researchers should be aware of some issues that may arise which could influence the data collected. Researchers should take into consideration potential areas of concern such as: middle positioning of responses and

"don't know" responses, social desirability, acquiescence bias, the influence of question order and type, as well the response options presented. Although in some cases no corrective measures may be possible, awareness of the sources of error and bias in measurement tools can allow for foresight regarding potential issues.

Whether or not to include a middle response option in scale construction is a challenging issue that researchers often face. A middle response option allows participants, who do not feel that they have a substantial opinion about the question, to avoid providing a response weighted in a specific direction (Presser & Schuman, 1980). Furthermore, it has been suggested by Presser and Schuman (1980) that if a middle response is not allotted to participants it may result in increased error in measurement as participants may not feel able to accurately respond to questions provided (as some individuals may have no specific feelings about the item). As well it has been found that not providing a middle response option may lead to increased variability in responses and decreased central tendencies, in certain cultures (Si & Cullen, 1998).

Alternatively it has been suggested that the quality of data in North American cultures may not be affected by the presence (or absence) of a middle response (Andrews, 1984; Si & Cullen, 1998). As well, Schuman and Presser (1980, 1981) have found that when a middle response option is provided there are typically no significant effects involving the distribution of responses on either side of the scale. However, it was noted that a participant's perceptions of what was expected from questions could be influenced by not providing a middle response alternative. For instance, no middle response may suggest to participants that you must be on one side or the other of an issue. Despite debate in the literature involving middle responses, it would seem that the inclusion of a

middle response is beneficial in alleviating any undue difficulty in responding for participants.

A similar debate found in the literature involves whether or not participants should be permitted to say, "I don't know", to questions provided. It has been noted that when participants are permitted to say, "I don't know", the amount of "don't know" responses decreases if a middle response option is provided (Bishop, 1987; Schuman & Presser, 1981). Consequently, some may view middle response options as a non-response, while others may see it as an outlet for fence sitting (when drawn to both ends of the scale), and finally some may view it as an "I don't know" (or absence of information response) response option. Furthermore, some participants struggle with the third cognitive stage of responding (which involves formulation of a judgment) thus providing a middle response or "don't know" response may alleviate this cognitive burden when responding. However, many researchers have expressed difficulty interpreting what a "don't know" response means in the context of data analysis. Additionally, some researchers might worry that allowing "don't know" responses may result in less data for analysis (Beatty, Herrmann, Puskar, & Kerwin, 1998; Koriat & Goldsmith, 1994;). When considering whether or not to permit participants to say they "don't know" to responses, researchers should evaluate the type of questionnaire (and data) being used. In some cases it may suffice to merely utilize a middle response option; however, other situations may require both middle and "don't know" response options.

Another source of error in questionnaire development involves the effect a specific ordering of questions may have on results. Due to the serialized nature of question responses, the questions that preceded the one presented may influence the way participants choose to respond. Ordering may lead to either primacy (first) or recency

(last) effects where participants' responses are contingent upon past information they have cognitively processed. This source of error is most readily apparent when two or more items are closely related in subject matter or wording. In addition, questions that are particularly salient to respondents may have a sizeable influence on later questions that are broad in nature, yielding results that may not actually represent the individual's true opinions (Schuman & Presser, 1981).

One issue that is rather pervasive in questionnaire data collection is the threat of social desirability influencing participant responses. Social desirability is described as a propensity to present oneself in a favorable light, in an attempt to be approved by others. This typically involves individuals responding in such a way that they appear to prescribe to norms and requirements suggested by society (King & Bruner, 2000). This behavior may seriously jeopardize the validity of questionnaire data, as individuals may opt to alter their initial (or true) response so that they may fit better within perceived social norms. As well, this issue is of particular concern when attempting to gain evaluative information from individuals. It has been noted that participants opt to hedge negative information and provide socially desirable responses even when they are assured that the information will remain confidential (Sudman et al., 1996; Thomas & Kilmann, 1975). This may have particular relevance to student evaluations of teacher effectiveness, as participants are required to evaluate an individual who has greater power than they do. Thus participants may be reluctant to provide information that does not fit within the social norms of the situation. Beyond the impact of social desirability on scale validity, others have noted that relationships between other variables of interest may be influenced by social desirability, causing spurious or suppressed relationships (King & Bruner, 2000). As a

result, measurement tools have been developed and are often utilized to control any potential effects as a result of social desirability (Crowne & Marlowe, 1960).

Acquiescence Bias

One of the paramount concerns that face researchers when developing questionnaires is the threat of acquiescence bias. Acquiescence bias is referred to in the literature as a respondent's increased propensity to agree with statements provided to them, despite the actual content of the item. An acquiescence response pattern is characterized by more "yes" responses than one would expect in typical responding; that is, if a participant is properly evaluating questions and answering honestly (Johanson & Osborn, 2004; Knowles & Nathan 1997; Schuman & Presser, 1981).

Interestingly, specific personality types seem to be more prone to acquiescence than others. Couch and Keniston (1960, 1961) found that there were significant correlations between response patterns ("yea-sayers" and "nay-sayers") and personality characteristics as noted in clinical interviews. Specifically they found that "yea-sayers" were typically more extroverted, impulsive, emotional and under-controlled. Alternatively, "nay-sayers" were found to be more introverted, cautious, rational and over-controlled. These findings have been verified by subsequent research, where it has been noted that specific personality types are associated with increased acquiescent responses (Knowles & Nathan, 1997; Ray, 1983).

Within the literature there are multiple perspectives regarding "why" participants opt to indiscriminately acquiesce to questions. The first approach implies that motivational and impression management issues plague "yea-sayers" when they are in the process of responding (Knowles & Condon, 1999). The historical understanding of motivation's relationship with acquiescence assumes that participants desire to please the

researcher and thus paint a positive image of themselves to the researcher. As a result, when responding, participants evaluate their initial answer and then attempt to contour the answer to fit within the perceived social expectations (Couch & Keniston, 1960, 1961). On the contrary, modern motivation theorists posit that individuals do not undergo an introspective process when responding. Instead they suggest that participants opt to provide the most readily available response that is socially desirable (Leary & Kowalski, 1990). This process may suggest that participants are involved in a form of cognitive satisficing when responding, whereby they are motivated to select the response with the least effort required while providing socially desirable information.

The second approach to acquiescence found within the literature involves primarily a cognitive perspective. Cronbach (1942, 1950) felt that acquiescent responses might be characterized by apathetic cognitive processing of questions that results in uncritical acceptance of the item. When applying Cronbach's view to Krosnick's (1991) cognitive satisficing model it seems that acquiescence may involve a lack of cognitive evaluation in the first stage of responding (understanding and interpreting the question) which then leads to "yea-saying" response habits. Under this approach, individuals who respond too quickly may not be providing adequate time to evaluate the questions and thus would likely be uncritically acquiescing to the question asked.

Other cognition researchers have suggested that acquiescence may be the result of apathetic information searches by participants. Specifically where participants neglect to pursue contraindicating information about the item (in memory) while only cognitively searching for confirmatory evidence in memory (Bassok & Trope, 1983; Zuckerman, Knee, Hodgins, & Miyake, 1995). Once again, using Krosnick's (1991) satisficing theory, under this approach it would appear that acquiescence issues could also stem from

the second stage of responding (retrieving the relevant information from memory). Furthermore, it has been noted that cognitive approaches by acquiescing respondents may be reflective of heuristics and peripheral route processing. Whereas respondents who utilize central route processing may be more contemplative of the confirming and disconfirming evidence involved and would be likely to select their response based upon thorough cognitive evaluation (Cacioppo, Petty, Kao, & Rodriguez, 1986). As well, it has been postulated that some individuals may not have the cognitive flexibility to adequately evaluate the information they possess. Consequently, such individuals may be more reliant upon heuristics when responding to items presented (Knowles & Nathan, 1997). Under this approach, respondents who exhibit a pattern of acquiescence would also be likely to respond quickly to questions; due to decreased cognitive effort when searching for relevant information.

According to Gilbert (1991), acquiescent responding may be a function of apathetic cognitive processing both in understanding the question as well as in the metacogntive evaluation of information possessed. Hence, an individual who provides an acquiescent response likely has difficulties understanding the question yet effortlessly agrees to the item presented. In this scenario, instead of reconsidering the response and evaluating if there is information that disconfirms their initial "gut instinct" the respondent opts to move on without additional cognitive effort. On the contrary, under normal reply conditions, one may initially agree with the statement presented, but then under further evaluation and cognitive effort decide not to agree with it. Hence, "yea-sayers" (participants with an acquiescence bias) would engage in quick cognitive responses to items, while not taking the necessary time to reevaluate the information they possess prior to giving a "final" reply.

Lastly it is important to consider that an interaction between effort and motivation may play an integral role in acquiescent responding. Motivation theorists often suggest that individuals will attempt to provide the least effort required to obtain satisfactory results (Simon, 1955/1979; Krosnick, 1991). Therefore, it has been suggested by researchers that apathetic cognitive processing could be related to motivational factors. Hence, individuals who lack motivation may be less inclined to cognitively engage in the questionnaire and as a result may be more likely to acquiesce to questions (Krosnick, 1991). Consequently, acquiescent responding may be less a function of innate cognitive ability and instead may be reflective of the individuals' motivation to cognitively engage in the requirements of the task.

It has been suggested by some that the type of items and response options implemented may also influence the frequency of acquiescence bias. In particularly, if items are perceived to be ambiguous by participants it may result in increased "yea-saying" by respondents (Hurd, 1999). Also, if participants do not properly understand the items, it can result in difficulties interpreting the participants' response patterns, because the responses may not reflect the individuals' actual opinions (Ray, 1983). Accordingly, many researchers have suggested implementing a method of controlling and evaluating acquiescence in the design of questionnaires. One way of measuring acquiescence is to provide each trait indicator question (positively worded items) with a trait contraindating item (negatively worded items). Consequently researchers are able to assess acquiescence as the total number of "yes" response to both positively worded items and negatively worded items (Knowles and Nathan, 1997).

Reverse Wording of Stems

Within questionnaire design, researchers often opt to use a mix of positively and negatively worded questions. Negatively worded questions are items that are constructed in the opposite semantic direction as positively worded items. This is often done in an attempt to deter respondents from simply agreeing with statements presented, by forcing respondents to critically evaluate the items first (Nunnally, 1978; Schriesheim & Hill, 1981). This tactic has been accepted as convention for many years, and originated when it was noted that individuals' responded more often in agreement than disagreement with statements presented (Barnette, 2000; Cronbach, 1950; Schriesheim, Eisenbach & Hill, 1991).

The positive ramifications of using negatively worded items have been well documented, with such benefits as: decreased acquiescence, as well as forcing participants to be more astute when responding. However, some research has found that opting to use negatively worded items can potentially attenuate the psychometric characteristics of the measure being used. Explicitly, it has been found that using negatively worded questions can lead to decreased internal consistency, as well as problems with factor structures and other related statistics associated with the measure (Barnette, 2000; Nunnally, 1978; Rossi, Wright, & Anderson, 1983). Consequently many measurement tools may implement negative wording, when in fact such a procedure does not benefit the researcher (Schriesheim, Eisenbach, & Hill, 1991; Schriesheim & Hill, 1981).

In a contrary view, Nunnally and Bernstein (1994) suggest that extremely high reliability in measurement tools may confound factor loadings and relationships amongst variables of interest. They suggest that reverse wording of items is often necessary,

despite the risk of potentially reducing reliability. They argue that items worded in a positive fashion result in more homogenous responses, which inflate reliability as a result of method variance, ultimately at the expense of measurement validity. Thus, to Nunnally and Bernstein, decreased reliability due to mixed wording of items is a necessary sacrifice to augment the utility of the measurement tool.

Beyond issues of scale reliability, Schriesheim et al. (1991) explicitly warn against the indiscriminant use of negatively worded questions, as it may result in decreased questionnaire validity. One study completed by Schriesheim and Hill (1981) found that when items were negatively worded, responses were less accurate when compared to responses to positively worded questions. This result suggests that researchers who choose to use negatively worded items may be unwittingly degrading the validity and accuracy of their findings. Based on research regarding questionnaire reliability and validity, Schriesheim et al. (1991) conclude that polar opposite or negative items should not be used as a control mechanism for acquiescent response patterns. Furthermore, they discuss their disapproval with the fact that psychometricians continue to advocate the use of alternating item wording, despite a body of literature that suggests otherwise.

More recently, it has been argued by Barnette (1999; 2000) that based upon the detrimental characteristics of negatively worded stems, researchers should only make use of such tactics if absolutely necessary. In particular, Barnette notes that most research does not require negatively worded items, and that negative wording is primarily needed when participants are not cognitively engaged in the task and/or are not motivated to complete the task. However, Barnette (1999; 2000) cautions that in some settings participants may be more likely to "non-attend" resulting in greater measurement issues.

"Non-attending" participants are labeled as those individuals who do not seem to be engaged in the cognitive response processes, as characterized by acquiescent or "deviant" response patterns. Specifically, he cites that educational evaluation settings may have a higher tendency toward "non-attending" response patterns. Furthermore, it is noted that even small occurrences of "non-attending" response patterns may lead to large differences in Cronbach's alpha coefficients and may result in different responses from individuals (Barnette, 1999).

Reversal of Response Scale Orientation

For researchers looking to obtain data via questionnaires, Likert-type scales have become the most prevalent method used within social science research (Dawis, 1987; Weng, 2004). When making use of Likert–type scales, researchers conventionally arrange the response anchors in a left to right orientation, where the most favorable response is presented first (on the left) and the least favorable response is presented last (on the right) (Chan, 1991). This formatting suggests that individuals would cognitively process their response options in a positive (furthest left) to negative (furthest right) manner when responding to questions. Thus, it has been noted by some that an individual's cognitive processing and ultimately their response could be manipulated by variations in arrangement of scale anchors.

The inclusion of bidirectional response anchors in questionnaire design has been suggested by some as an alternative to negative wording when attempting to deter respondents from responding in an acquiescent manner (Barnette, 2000; Robinson, Shaver, & Wrightsman, 1991). This approach arose from the aforementioned concern that negative item wording may detrimentally impact the responses provided (Barnette, 2000; Schriesheim & Hill, 1981). According to Barnette (2000), the use of bidirectional

response orientation may deter acquiescent responding while avoiding the use of negatively worded questions and problems associated with them. This assertion stems from recent findings, where positively worded questions combined with bidirectional response options resulted in greater reliability cofficients than measures using alternating item wording combined with unidirectional response options (Barnette, 2000).

As an alternative to negatively worded items, bidirectional response options allow the item presented to be worded in a positive fashion, and forces participants to evaluate their response in the encoding phase of responding (Barnette, 2000). As a result, it is possible that bidirectional response options may encourage participants to appropriately understand the question provided as well as more accurately determine the relevant information, while forcing them to consider their response before formally providing it. Therefore, bidirectional response options may be a worthwhile alternative to negative wording, as it may deter participants from acquiescing without altering the meaning or wording of the questions provided.

Scale Length

When considering scale construction there is seemingly no definitive consensus as to the scale characteristics that should be implemented. Namely, determining what length of scale or what type of scale anchors should be used can provide questionnaire developers with a difficult dilemma. This quandary is readily apparent when examining the process by which teacher evaluation instruments are constructed, and the lack of consensus regarding one approach over another (Alwin, 1997; Jackson et al., 2007b; Sedlmeier, 2006).

Past research involving the number of response options used within questionnaires has lead to some contention amongst researchers (Alwin, 1997; Reber, 1996; Weng,

2004). In some instances researchers have noted that the number of response options given to participants has no significant impact on the internal consistency of the measure, nor the concurrent validity when compared to similar measurement tools (Reber, 1996). To the contrary, Lozano, García-Cueto and Muñiz (2008) using Monte Carlo methodology found that reliability and validity of measures increased as the number of response options increased. This result suggests that the subtle difference of increasing the number of response options in measurement tools may bolster the validity and reliability of the measure.

Even with no definitive direction presented regarding the number of scale points to be used in questionnaire research, most researchers agree that some characteristics are necessary to obtain optimal data. It has been suggested that increased response options may provide for improved power and accuracy in reporting from respondents (Alwin, 1997; Ing & Jackson, 2006; Weng, 2004). Researchers often utilize between 2 and 11 response options when structuring questionnaires. However, as evidenced by the literature, questionnaires should contain a minimum of 5 response options. By providing 5 response options, individuals are able to differentiate between the intensity and directionality of their opinions, while still having a middle option available (Alwin, 1997; Weng, 2004).

Despite equivocal findings regarding the psychometric benefits of increased response options in questionnaires, past research in this area seems to suggest that differences in the number of response options presented may cause for differences how individuals respond (Reber, 1996; Ing & Jackson, 2008). For example, it has been noted that when responding to SETEs participants were more likely to provide positive responses to longer scales (using 5-point vs. 9-point likert scales) (Ing & Jackson, 2008).

Other researchers have found similar results, where fewer response options resulted in significantly different responses, despite scales possessing statically equivalent psychometric properties (Reber, 1996).

One explanation for these findings may be related to satisficing theory, specifically how respondents perceive and cognitively translate memories (or related opinions) into the scale provided. In particular, research pertaining to the grain size of response options may assist in understanding the impact that scale length has on responses. In this line of research grain size of responses is defined as the amount of detail required in responding. For example, a fine grain response would be one that provides a great amount of precision (e.g. the exact number of people in a room) at the risk of accuracy (as it may be challenging to know exactly how many people there are in a room). On the other hand, a coarse grain response would provide less detail (e.g. 50 – 100 people in a room) but is more likely to be accurate (as the number of people is more likely to fit into a range). Ackerman and Goldsmith (2008) discuss how fine grain responses are more cognitively taxing to derive and often require respondents to be more confident in the information. This avenue or research applied to response scale length suggests that increased granularity of response options (i.e. longer response scales, requiring finer granularity) may require more precision and consequently more cognitive effort from the participant (Goldsmith, Koriat, & Weinberg-Eliezer, 2002). As a result, respondents may be more inclined to rely on heuristics at this stage of responding and consequently be more likely to acquiesce (providing a positive or agreement response).

However, it has been noted that for individuals who are not cognitively well equipped or are not of mature cognitive ability (i.e. children, adolescence) increasing the number of scale options may have detrimental results (Borgers, Hox, & Sikkel, 2004;

Weng, 2004). In one study Borgers et al. (2004) found that increasing the number of response options to 7 or more when providing questionnaires to children leads to a decrease in the reliability and stability of the measurement tool. It is also possible that increased scale length could interfere with a participant's ability to appropriately translate their opinions into the presented scale if they are cognitively apathetic (despite age or innate ability). Increasing the number of response options available may require more cognitive effort from participants when translating judgments into the scale provided (Krosnick, 1991). Therefore, in scenarios where participants are cognitively unmotivated it is possible that results could be less accurate if a longer scale is utilized. Thus in spite of evidence that increasing the number of response options may improve results, it is imperative that the researcher considers the population in question. In particular, researchers should be mindful of the cognitive abilities as well as the motivational characteristics of those who will be responding, as they can influence the impact that scale length has on participants (Alwin, 1997; Weng, 2004).

Recent Investigation Combining Reverse Wording and Scale Orientation

One study conducted by Barnette (2000) examined the benefits and drawbacks of negatively worded items and response option reversals. This study incorporated a 2x3 experimental design, where there were two levels for the wording condition (positive wording; mixed wording) and three levels of response option orientation (left to right, right to left and mixed). The primary objective of this research was to determine if there were any significant differences in reliability coefficients (Cronbach's alpha) based upon item wording or response orientation (as well as examining any potential interaction effects between these variables). This approach was based on past recommendations by researchers, regarding the usefulness of altering response alternative orientation to curb

acquiescence by participants (Robinson, Shaver, & Wrightsman, 1991). Furthermore, this was completed to determine if altering response option orientation would allow for a reliable measure, while potentially deterring participants from indiscriminately acquiescing to questions. The results of this study showed that positively worded items, regardless of response orientation, provided for the most reliable measurement tool. Interestingly, there were no significant differences in reliability between response orientation conditions. As a result the author suggests that positively worded items in combination with mixed (bidirectional) response orientations may allow researchers to control for acquiescence by participants without jeopardizing the reliability of the measurement tool (Barnette, 2000).

Barnette's (2000) study provided for a novel understanding and approach, by manipulating both the response scale orientation, as well as the item wording. However, the study was somewhat limited in that it was primarily focused on the reliability of the measurement tool. Although reliability is an important characteristic in measurement, some researchers have voiced concern that Cronbach alpha scores that are too high may be the result of decreased validity in the measurement tool. This may be particularly relevant in the case of acquiescence bias, as agreement responses often present a highly homogenous response pattern, which should result in high Cronbach alpha coefficients (Hulin, Netemeyer, & Cudeck, 2001). As well, this study did not include a completely negative wording condition, which means that the results of this study may be characteristic of differences between mixed (positive and negative) and positive worded measurement tools. Thus, it is difficult to generalize the results specifically to the effect of negatively worded questions, without including a condition where participants must respond exclusively to negatively worded questions.

The current research project was designed to replicate and expand on Barnette's (2000) findings by including a positive wording condition, a negative wording condition and a mixed wording condition. This additional condition should elucidate the influence negative wording has on questionnaire characteristics. Furthermore, in this study the number of dependent variables observed were expanded to include both mean response scores (based on different past classroom experiences) and mean response times. The addition of response time means as a dependent variable allows insight into the influence questionnaire characteristics have on the cognitive processes involved in responding.

Hypotheses

In addition to specific hypotheses, scale reliability was examined based on the *item wording*, *scale orientation* and *scale length* of the questionnaire. For the sake of this study, *item wording* was defined as the semantic orientation of questions implemented. This independent variable consisted of three levels: *positive wording of items, negative wording of items and mixed wording of items. Scale Orientation* was defined as the visual depiction of response options provided to participants. This independent variable consisted of two levels: *unidirectional response scale* (options arranged in a left to right fashion) and *bidirectional response scale* (response options varied from left to right, and right to left). Finally, in this study *scale length* was defined as the number of response options provided to participants. This independent variable was comprised of two levels: *5-point scale* (short scale) and *11-point scale* (long scale). Based on past research involving questionnaire development and cognitive theories involving how individuals process and respond to questions, the following hypotheses have been developed:

Hypothesis 1. It was expected that there would be a significant relationship between response time means when answering questions and response ratings submitted.

As this aspect of the study was exploratory in nature, correlations between response means and response time means were examined across all of the teaching effectiveness factors, separately for each type of course (liked; disliked). Also, no specific prediction regarding the directionality of the relationships was made. These relationships are expected based on prior research, where it has been inferred that the amount of thought individuals put into responding to a question should be related to the responses that are submitted (Bassili, 1993; Yan & Tourangeau, 2008).

Hypothesis 2. A two-way interaction was expected between the *item wording* and *scale orientation* on responses. More precisely, it was expected that positive wording with unidirectional response scales (left to right or right to left) would yield more positive responses from participants when compared to positively worded items with bidirectional response scales (mixed). However, no differences were expected between response means based on response option orientation when participants were given mixed or negatively worded items. This hypothesis was based on the assumption that bidirectional response options would curb the acquiescence response tendencies of participants (Barnette, 2000).

Hypothesis 3. A main effect was expected for *scale length* on responses. In particular it was anticipated that long response scales (11 point) would result in more positive responses from participants, as compared to short response scales (5 point). This hypothesis was based on past research completed by Ing and Jackson (2008), where it was noted that longer response scales resulted in more positive responses from participants.

*Hypothesis 4. T*wo-way interactions were expected for *wording of items* and *scale orientation* on response time. It was expected that participants who received mixed

worded items (half positively worded, half negatively worded) in combination with bidirectional response scales would take the longest time to respond, where positively worded items would yield the shortest response times. Furthermore, it was expected that unidirectional response scales would result in shorter response times for all wording conditions when compared to those posted using bidirectional response scales. This hypothesis was based on cognitive theory, whereby greater difficulty within the stages of responding should result in longer response times from participants (Schwarz, 1999, 2007; Sudman, Bradburn & Schwarz, 1996; Tourangeau & Rasinski, 1988).

Hypothesis 5. A main effect was expected for *item wording* on response times. Specifically, it was anticipated that negatively worded questions would result in longer response times than positively worded questions. This hypothesis was based on the concept that negative information is more difficult for individuals to cognitively process (Clark & Chase, 1972; Hearst, 1991; Sherman, 1976; Williams, 1991).

Hypothesis 6. A main effect was expected for *scale length*, such that long response scales (11 point likert) would result in longer response times by participants. This hypothesis was based on findings that suggested that response times are influenced by the scale characteristics implemented (Heerwegh & Loosveldt 2002; Tourangeau et al., 2004).

Methods

Participants

The University of Windsor's Ethics Committee approved all stages and components of the methods. The sample used for this study was comprised of 459 students who were enrolled in an undergraduate psychology course at the University of Windsor at the time of participation. Participants were recruited through the online

participant pool at the University of Windsor, where they personally selected and registered to participate in this study. Participants were required to have completed at least one full semester of courses at the University of Windsor or at another Canadian University, to verify that they had experience with the SETE forms. Sessions were conducted with up to 16 participants per time slot and were approximately 30 minutes in duration. Data from 21 participants were omitted in the final analyses, as they did not properly adhere to the instructions in the study (for more detail see the results section below). Of those who participated, 82.8% were female, 16.3% were male and .6% did not disclose their sex. Academically, most participants were either in their second (30.1%), third (33.6%) or fourth year (24.6%) (6.3% were in their first year and 5.4% responded 'other'). Despite requiring participants to have completed at least one full semester of courses in the past, 5% stated that they had not completed any previous course evaluations.

Measures

Participants were asked to complete the Students' Perceptions of Teaching Effectiveness scale, second edition (SPTE II). This measure is comprised of 58 items (19 items involve demographic information of the professor and student and were not included in the study), 39 of which are used for evaluation of professor effectiveness by students. The 39 items are typically anchored on a 5-point agreement scale (ranging from "Strongly Agree" to "Strongly Disagree"). Agreement anchors were used because they are frequently used in research testing variations in questionnaire design (Arseneault & Jackson, 2005; Ing & Jackson 2006, 2007, 2008). Also, it has been noted in past research that different scale anchors (i.e. evaluative anchors, or agreement anchors) do not appear to result in differences in the responses submitted by participants (Ing & Jackson, 2008).

Past research has found that these items load on one of six factors: Rapport with Students, Course Value, Course Organization and Design, Fairness of Grading, Course Difficulty, and Workload (Burdsal & Bardo, 1986; Jackson et al., 1999). As this study looked to build upon Jackson et al.'s (2007b) and Ing and Jackson's (2008) prior work, four of the six factors were chosen to be used in the study because they had been used exclusively in the preceding research. Specifically, the four factors used were: rapport with students; course value; course organization and design; and fairness of grading. Because the wording conditions were comprised of three levels (all positive; all negative; and mixed items), each item that was used included both a positively worded version, as well as a negatively worded version.

The first factor, *Rapport with Students,* is comprised of 7 items and focuses on the ability of the instructor to develop and maintain rapport with the students for the duration of the class. The second factor, *Course Value*, is comprised of 4 items and encompasses the students' perceived value in the course based upon: knowledge gained; expected retention; enthusiasm for attending class; recommendation of the course to others; and further interest in the subject matter, resulting from taking the course. The third factor, *Course Organization and Design*, is comprised of 7 items and involves assessment of skills and competencies the instructor may or may not possess, which include: organization; preparation; clarity and suitability of presentation in conveying course concepts and objectives; and answering questions. The fourth factor, *Grading Fairness*, is comprised of 4 items and incorporates the participants' perceptions of grading practices in three categories: quantity of evaluations, clarity of evaluations, and validity (appropriateness) of evaluation methods (Appendix C). In past research reliability coefficients for the four factors have ranged between .63 and .88 (Rapport with Students

.68 - .84; Course Value .65 - .79; Course Organization and Design .66 - .88; Fairness of Grading .63 - .74). However the inter-item reliability coefficients varied based on the questionnaire characteristics implemented (Ing & Jackson, 2008).

Exclusive use of these factors made for a shortened version of the SPTE II, where 22 items were given to participants. The shortened SETE was administered to assess the participants' evaluation of both a liked class experience and a disliked class experience. This was accomplished by providing two identical versions of the SETE to each participant (as a within subjects variable); one version was given to assess a liked class, and another was given to assess a disliked class. Prior to responding to each section, participants were provided with one of the two following instructions (which will be counterbalanced to control for potential order effects):

> Please answer the following questions for a class that you completed last semester that you **liked**.
>
> Please answer the following questions for a class that you completed last semester that you **disliked**.

In order to verify that participants had read and understand the questions given to them, multiple validity checks were incorporated in the questionnaire. Firstly, participants were asked whether they were assessing a liked or disliked course after being administered each version of the questionnaire. This was used to verify that participants had read and understood the instructions given to them for each version of the questionnaire. Also, as participants were responding via computer and software, in two instances within each version of the questionnaire they were asked to *'leave this question blank and click next'*.

Furthermore, demographic information was collected from each participant, which included: whether the class that was rated was liked or disliked by the participant; participant's degree of certainty regarding the accuracy of responses (overall, for both versions of the questionnaire); participant gender; grade received in the course that was assessed; participant year in university; age of participant; participant ethnicity; ethnicity of the instructor for the rated course; whether the participant had been taught by the instructor more than once; instructor gender; level of course that was assessed and whether English was the participants' first language.

In order to evaluate the effectiveness of response orientation in questionnaire usage, there were two levels of *scale orientation* incorporated in the questionnaire. In one condition all response anchors were oriented in a unidirectional (right to left) manner: *agree/disagree*. In the second condition, response anchors were oriented in a bidirectional manner, where half of the response anchors were oriented right to left (*agree/disagree*) and the other half were oriented from left to right *(disagree/agree*). As well there was an additional factor involving *scale length*, which was comprised of two levels: in the first condition responses were anchored on a 5-point scale, and in the second condition responses were anchored on an 11-point scale. An 11-point scale was used, as opposed to the 9-point scale used in earlier research, to maximize variability in responding based on scale length.

To gain insight into the participants' cognitive processes while responding to questions, response time was recorded as participants responded to each question. As this study was conducted using computer based data collection, all reaction time data was calculated via software. Response times were computed by recording the initial time when the question was presented as well as recording the time after the user had selected

a response option and clicked next. An elapse time was calculated from these two time points; this provided a numerical response time for each individual question, for each participant. This data was collected from the user side (as opposed to server side) of the web based software application; meaning that server latency did not interfere with response time calculations (all data was collected in real-time and then transferred to the web server).

Procedures

All sessions were conducted in a computer laboratory, where each participant was assigned a computer to be used for the duration of the study. When participants arrived for the study they were asked to select an open desk with a computer and were asked to read and respond to a consent form. Participants were then provided with a login name and password. The version of the questionnaire that the participants completed was randomly assigned, and was administered based on the login that was used to access the website. The web based application that was utilized was specifically developed for this study and was written using Adobe ColdFusion 8. The application was presented as a web-based questionnaire, with instructions given prior to starting and between each administration of the questionnaire. All responses were collected via the Internet and stored on a remote server.

Participants were then instructed to carefully read and follow the instructions given to them on the computer. Additionally, participants were informed electronically (within the application), that they were unable to return to prior questions, and that they should consider and select their responses accordingly. Each version of the study included a series of demographic questions (Appendix C). Participants were asked to complete the questionnaires for both past class experiences (liked and disliked; presented in

counterbalanced order). After completing both versions of the questionnaire participants were provided debriefing information and thanked for their participation.

Methodology

An experimental design was used to address the research hypotheses posed; this design included one within-subjects variable (counterbalanced to control for the effect of order) and three between-subjects variables. The within-subjects variable involved all participants responding to the questionnaire for a class they liked and for a class that they disliked. All of the participants received survey items created to measure each of the four dimensions of teaching effectiveness. The *item wording* provided to participants was manipulated as a between-subjects variable. There were three levels of this variable, where one third of the participants were given all positively worded questions, another third of the participants were given all negatively worded questions and the final third were given half of the questions positively worded and the other half of the questions negatively worded (mixed wording). The *scale orientation* provided to participants was also manipulated as a between-subjects variable. There were two levels of this variable, where one half of the participants were provided with response anchors oriented from right to left (*agree/disagree*) (unidirectional), and the other half were given half of the response anchors oriented right to left (*agree/disagree*), and the other half oriented left to right (*disagree/agree*) (bidirectional). Finally, the *scale length* provided to participants was manipulated as a between-subjects variable. Half of the participants were provided with a 5-point response scale, and the other half were provided with an 11-point response scale. The between subjects-variables used were randomly assigned to the participants.

Results

Data Analysis

A total of 459 participants completed the questionnaires for both class types. Of the participants who completed the study, 21 were removed because they had violated the embedded validity checks. In particular, 20 participants incorrectly stated the type of class they had responded to, when compared to the type of class they had been asked to respond to. One additional participant responded to items when they were asked not to.

According to central limit theorem, due to the relatively large sample size, the sampling distribution of means should be approximately normal. Outliers were defined as absolute values greater than 3.0 standard deviations away from the mean, using scale responses as the dependent variables. A total of 17 participants were classified as outliers and excluded from the analyses, resulting in a final sample size of 421. Alternatively, for response times as the dependent variables, a log transformation was conducted and no additional participants were excluded from the analyses. The log transformation of the response times posted was completed because it was noted through visual inspection that response times for each factor were skewed in a positive direction and outliers (as defined above) were predominantly characterized by excessively long response times. Thus, use of a log transformation was an ideal solution as it provided a correction for extreme durations of responding without decreasing the sample used for the analyses (Field, 2005). After compensating for extreme values, visual inspection of histograms and related statistics (i.e. skewness and kurtosis) demonstrated that the assumption of normality was satisfied for each dependent variable.

The assumption of homogeneity of variance was assessed using Levene's test of the homogeneity of error variance. When testing this assumption using scale responses as

the dependent variable, the one violation noted was for disliked responses on the '*Course Value*' factor. When testing this assumption using response times as the dependent variable only liked responses to '*Rapport with Professor*' violated this assumption. When examining the assumption of homogeneity of covariance matrices for scale responses, only the '*Course Value*' factor provided a violation. When examining this assumption for response times, only the '*Rapport with Professor*' factor violated this assumption. However, because the data were normally distributed, the cell sizes were approximately equal across the cells (33 - 38 per cell) and all cells exceeded 20, the assumption of homogeneity of variance-covariance matrices was thought to be tenable for the analyses (Stevens, 2002).

Bivariate correlations were conducted between the scale response means and the scale response time means for each of the four teaching effectiveness dimensions based on the type of class evaluated (liked and disliked). Only correlations pertinent to understanding the relationship between response means and response time means were explored, the resulting correlation matrices can be found in Table 4, Table 5, Table 6 and Table 7. In addition, the data were analyzed using four (one for each teaching effectiveness dimension) separate 2 (class type) x 3 (item wording) x 2 (scale orientation) x 2 (scale length) mixed-randomized by repeated measure analyses of variance (ANOVA), for each dependent variable (response means; and response time means). The *within-subjects* variable was *class*, with two levels: (1) liked and (2) disliked. The *between-subjects* variables included: *item wording*, evenly divided between (1) positively worded items, (2) negatively worded items, and (3) mixed worded items; *scale orientation*, evenly divided betwen (1) left to right orientation (unidirectional scale options), (2) and mixed orientation (bidirectional scale options); and *scale length*

separated into (1) five-point (short) and (2) eleven-point (long) scales. Responses measured with five-point scales were converted to values on an eleven-point scale; this was done so that comparisons between the two levels of *responses length* could be conducted. All tests were conducted using an alpha level of .05, including any contrasts used to evaluate the *a priori* hypotheses. However, due to the number of analysis involved, any further post hoc evaluations were conducted using a Bonferroni correction to maintain an alpha of .05.

Scale means, standard deviations, and reliability estimates were computed for each of the four teaching effectiveness dimensions based on the experimental conditions implemented and are presented in Table 1, Table 2 and Table 3. In addition, scale response time means and standard deviations were computed for the four teaching effectiveness dimensions based on the experimental conditions used. For the most part, the internal consistency reliability estimates were good for all of the teaching effectiveness dimensions, where 'good' was defined as values greater than .70 (Kaplan & Succuzo, 2005). When examined as a whole, it appears that variations in questionnaire design may attenuate reliability, but for the most part the Cronbach α coefficients were relatively consistent across the cells.

Table 1 *Means, Standard Deviations and Reliability Estimates for Scale Length*

	Teaching Effectiveness Dimension							
	Rapport with Students		Course Value		Course Organization and Design		Fairness of Grading	
Five-Point Scale	Liked Course	Disliked Course	Liked Course	Disliked Course	Liked Course	Disliked Course	Liked Course	Disliked Course
n	189	189	189	189	189	189	189	189
M	9.48	6.34	9.38	4.75	9.87	6.59	9.27	6.03
SD	1.16	1.72	1.36	1.74	1.06	1.63	1.48	1.90
α	.66	.77	.60	.67	.76	.76	.71	.65
Eleven-Point Scale								
n	205	205	205	205	205	205	205	205
M	9.50	6.03	9.34	4.10	9.80	6.39	9.03	5.70
SD	1.11	2.17	1.82	2.11	.95	2.19	1.53	2.16
α	.76	.84	.70	.78	.82	.86	.76	.72
Combined Scale								
n	394	394	394	394	394	394	394	394
M	9.49	6.18	9.36	4.41	9.83	6.49	9.14	5.85
SD	1.13	1.97	1.84	1.97	1.00	1.94	1.51	2.05
α	.72	.82	.65	.74	.80	.82	.74	.69
Corrected Item-Total Correlation	.66-.74	.78-.81	.53-.64	.67-.75	.75-.80	.78-.81	.62-.75	.55-.68

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

Table 2 *Means, Standard Deviations and Reliability Estimates for Scale Orientation*

	Teaching Effectiveness Dimension							
	Rapport with Students		Course Value		Course Organization and Design		Fairness of Grading	
Uni-directional Scale	Liked Course	Disliked Course	Liked Course	Disliked Course	Liked Course	Disliked Course	Liked Course	Disliked Course
n	198	198	198	198	198	198	198	198
M	9.64	6.15	9.44	4.42	9.92	6.42	9.22	5.77
SD	1.01	2.01	1.35	2.00	.97	1.94	1.50	2.07
α	.74	.84	.69	.75	.84	.82	.75	.71
Bi-directional Scale								
n	196	196	196	196	196	196	196	196
M	9.34	6.21	9.28	4.41	9.74	6.56	9.06	5.94
SD	1.23	1.93	1.36	1.94	1.04	1.95	1.52	2.02
α	.70	.80	.62	.73	.76	.82	.73	.67

Table 3 *Means, Standard Deviations and Reliability Estimates for Item Wording*

	Teaching Effectiveness Dimension							
Positive Wording	Rapport with Students		Course Value		Course Organization and Design		Fairness of Grading	
	Liked Course	Disliked Course	Liked Course	Disliked Course	Liked Course	Disliked Course	Liked Course	Disliked Course
n	134	134	134	134	134	134	134	134
M	9.61	6.19	9.50	4.73	9.82	6.70	9.35	6.19
SD	1.08	1.99	1.23	1.94	.94	1.97	1.34	1.98
α	.82	.84	.68	.77	.76	.85	.75	.68
Negative Wording								
n	127	127	127	127	127	127	127	127
M	9.57	6.27	9.32	4.13	9.92	6.45	9.06	5.86
SD	1.14	2.04	1.46	1.98	1.02	2.04	1.51	1.94
α	.70	.82	.68	.75	.86	.83	.76	.64
Mixed Wording								
n	133	133	133	133	133	133	133	133
M	9.29	6.08	9.27	4.36	9.76	6.33	9.02	5.52
SD	1.16	1.88	1.38	1.94	1.05	1.80	2.74	2.17
α	.63	.79	.60	.70	.76	.79	.70	.74

Bivariate Correlations

For the *Course Design* factor, there was a significant negative relationship between response means and response time means for liked classes (Table 4). This result suggests that as participants put less time (and likely thought) into their responses for a liked class, the responses given were more positive. However, there were no significant relationships for disliked classes.

Table 4 *Correlations for Course Design*

Measure	1	2	3	4
1. 'Liked' Score	--	.050	-.139(**)	.046
2. 'Disliked' Score	--	--	.083	.022
3. 'Liked RT'	-	--	--	-.134(**)
4. 'Disliked RT'	--	--	--	--

** Correlation is significant at less than a 0.01 level (2-tailed).

For the *Fairness of Grading* factor, there was found to be a significant negative relationship between response means and response time means for liked classes (Table 5). Similarly to *Course Design* responses, as participants put less time into their responses for a liked class, the responses given were more positive. Also, there were no significant relationships for disliked classes.

Table 5 *Correlations for Fairness of Grading*

Measure	1	2	3	4
1. 'Liked' Score	--	.121(*)	-.219(**)	.067
2. 'Disliked' Score	--	--	-.004	.024
3. 'Liked RT'	-	--	--	.432(**)
4. 'Disliked RT'	--	--	--	--

* Correlation is significant at 0.05 (2-tailed).
** Correlation is significant at less than a 0.01 level (2-tailed).

For the *Course Value* factor, there was found to be a significant negative relationship between response means and response time means for a liked course. In addition, for disliked classes there was found to be a significant positive relationship between response means and response time means (Table 6). This result suggests that as the participant put less time into their responses for liked course, the responses given were more positive. Whereas, when respondents put more time into their responses for disliked courses, the responses more positive.

Table 6 *Correlations for Course Value*

Measure	1	2	3	4
1. 'Liked' Score	--	-.090	-.259(**)	-.020
2. 'Disliked' Score	--	--	.000	.198(**)
3. 'Liked RT'	-	--	--	.381(**)
4. 'Disliked RT'	--	--	--	--

** Correlation is significant at less than a 0.01 level (2-tailed).

Finally, for the *Rapport with Students* factor, there was found to be a significant negative relationship between response means and response time means for liked classes.

As well, for disliked classes there was a significant positive relationship between responses and response times (Table 7). Again, this result suggests that as the participant put less time into their responses for a liked course, the responses given were more positive. Whereas, when respondents put more time into their responses for disliked courses, the responses were more positive.

Table 7 *Correlations for Rapport with Students*

Measure	1	2	3	4
1. 'Liked' Score	--	.032	-.171(**)	.008
2. 'Disliked' Score	--	--	.093	.153(**)
3. 'Liked RT'	-	--	--	.228(**)
4. 'Disliked RT'	--	--	--	--

* Correlation is significant at 0.05 (2-tailed).
** Correlation is significant at less than a 0.01 level (2-tailed).

Taken as a whole, it appears that the weak negative relationship between response score means and response time means for liked courses is rather pervasive, as it occurs for all four teaching effectiveness dimensions. Contrarily, for disliked courses the weak positive relationship between response means and response time means only occurs for two of the teaching effectiveness dimensions. As a result, it appears that when responding to disliked classes, the type of information being requested also influences the relationship between responses and response times.

Significant Within-Subjects Main Effects for Responses

Across all four dimensions, there was a significant effect for the *type of course* being evaluated on the responses submitted: Course Design ($F_{[1,406]} = 984.30$, $p < .001$, $\eta^2 = .72$, $\omega^2 = .71$); Fairness of Grading ($F_{[1,406]} = 779.80$, $p < .001$, $\eta^2 = .66$, $\omega^2 = .66$);

Course Value ($F_{[1,406]} = 1654.81$, $p < .001$, $\eta^2 = .81$, $\omega^2 = .80$); and Rapport with Students ($F_{[1,406]} = 911.708$, $p < .001$, $\eta^2 = .70$, $\omega^2 = .69$). These tests can be thought of as a manipulation check for the study, as participants were asked to respond to one class that was liked and another that was disliked. These differences were in the desired direction (liked classes resulted in more positive responses than disliked classes), thus this provides validity to the assumption that participants were properly attending to the instructions given during the study (the scale means for each condition can be found in Table 1).

Significant Between-Subjects Main Effects for Responses

There was found to be a significant main effect on the *Fairness of Grading* factor for *scale length*, yielding an *F* statistic of $F_{[1,406]} = 4.58$, $p < .05$, $\eta^2 = .01$, $\omega^2 = .01$. Further examination of this result showed that shorter response scales, elicited more positive responses from participants (the scale means for each condition can be found in Table 8). In addition, a significant main effect was found for *scale length* on the *Course Value* factor ($F_{[1,406]} = 10.64$, $p < .001$, $\eta^2 = .03$, $\omega^2 = .02$); however, this result was qualified by an interaction with the type of course being evaluated (see the '*Significant Interactions for Responses'* section below). These results run contrary to the *a priori* hypothesis presented, where it was expected that long response scales would result in more positive responses.

Table 8 *Response Means for Scale Length by Course Type*

		Dimension of Teaching Effectiveness	
		Course Value	Fairness of Grading
5-point Scale	M	7.06	7.65
	SD	1.07	1.29
11-point Scale	M	6.71	7.36
	SD	1.18	1.37

On the *Fairness of Grading* factor there was found to be a significant main effect for *item wording* ($F_{[1,406]} = 4.31$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$). As this result was hypothesized, relevant linear contrasts were conducted. The results showed that participants in the positive wording condition submitted more positive responses when compared to those who had been given mixed wording condition (the scale means for each condition can be found in Table 9). However, there were no differences between the negative wording condition and the other two groups (positive wording; and mixed wording). In addition, a main effect for *item wording* was found on the *Course Value* dimension ($F_{[1,406]} = 3.37$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$), as well, the relevant linear contrasts were conducted. The results showed that positive wording yielded more positive responses when compared to negative wording (the scale means for each condition can be found in Table 9). However, there were no differences between mixed wording and the other two groups (positive wording; and negative wording).

Table 9 *Response Means for Item Wording by Course Type*

		Dimension of Teaching Effectiveness	
		Course Value	Fairness of Grading
Positive Wording	M	7.09	7.76
	SD	1.12	1.29
Negative Wording	M	6.73	7.48
	SD	1.10	1.28
Mixed Wording	M	6.82	7.27
	SD	1.16	1.41

Significant Interactions for Responses

A significant interaction was found between the *type of course* and *scale length* for the *Course Value* dimensions ($F_{[1,406]} = 8.43$, $p < .01$, $\eta^2 = .02$, $\omega^2 = .02$); as a result, a

Post Hoc analysis was conducted using a Bonferroni correction. It was noted that for disliked classes, the 5-point scale condition resulted in more positive responses when compared to the 11-point scale condition. However, for liked classes there were no differences in response means based on response scale length. The scale means for each condition can be found in Table 1.

Significant Within-Subjects Effects for Response Times

There was a significant effect for the *type of course* being evaluated on all of the dimensions, except *Rapport with Students*, on the response times posted: *Course Design* ($F_{[1,409]} = 17.42$, $p < .001$, $\eta^2 = .04$, $\omega^2 = .04$); *Fairness of Grading* ($F_{[1, 409]} = 30.87$, $p < .001$, $\eta^2 = .07$, $\omega^2 = .07$); and *Course Value* ($F_{[1, 409]} = 22.86$, $p < .001$, $\eta^2 = .05$, $\omega^2 = .05$). For all of these dimensions, disliked classes resulted in longer response times, when compared to liked classes (the response time means for each condition can be found in Table 10). These results suggest that on average participants put more time into their responses when they were asked to respond to a disliked course. However, for all of these dimensions, the results were qualified by interactions with other experimental conditions implemented (see the '*Significant Interactions for Response Times'* section below).

Table 10 *Response Time Means in Milliseconds for Course Type*

		Dimension of Teaching Effectiveness			
		Rapport with Students	Course Value	Course Organization and Design	Fairness of Grading
Liked Course	M	3966.55	3828.07	3925.74	3887.78
	SD	390.86	150.16	175.48	157.72
Disliked Course	M	3979.85	3866.70	3979.40	3930.45
	SD	157.85	147.11	171.93	140.22

Significant Between-Subjects Main Effects for Response Times

There was a significant effect on all dimensions, except *Fairness of Grading,* for *scale orientation* on the response times posted: *Course Design* ($F_{[1, 409]} = 12.16$, $p < .001$, $\eta^2 = .03$, $\omega^2 = .03$); *Course Value* ($F_{[1, 409]} = 19.55$, $p < .001$, $\eta^2 = .05$, $\omega^2 = .04$); and *Rapport with Students* ($F_{[1, 409]} = 10.81$, $p < .001$, $\eta^2 = .03$, $\omega^2 = .02$). For all of these dimensions, unidirectional response options elicited shorter response times, when compared to bidirectional response options (the response time means for each condition can be found in Table 11). These results suggest that on average participants took longer when responding if they were presented with bidirectional response options. However, for all of these dimensions, the results were qualified by interactions with other experimental conditions that were utilized (see the '*Significant Interactions for Response Times'* section below).

Table 11 *Response Time Means in Milliseconds for Scale Orientation*

		Dimension of Teaching Effectiveness			
		Rapport with Students	Course Value	Course Organization and Design	Fairness of Grading
Unidirectional Scale	M	3947.61	3829.92	3955.24	3906.51
	SD	133.02	127.87	121.19	132.85
Bidirectional Scale	M	4033.72	3882.02	3990.34	3927.45
	SD	499.63	124.35	106.28	119.80

For two of the teaching effectiveness dimensions, there were significant main effects for *scale length* on the latency of responding: *Fairness of Grading* ($F_{[1, 409]} = 5.15$, $p < .05$, $\eta^2 = .01$, $\omega^2 = .01$) and *Course Value* ($F_{[1, 409]} = 4.29$, $p < .05$, $\eta^2 = .01$, $\omega^2 = .01$). For all of the factors, it was noted that 11-point response scales resulted in longer response times, as compared to 5-point response scales (the response time means for each

condition can be found in Table 12). This result coincides with the *a priori* hypothesis presented, that longer response scales would result in longer response times. In the case of the *Course Value* dimensions, the result was qualified by an interaction with other included variables (see the '*Significant Interactions for Response Times'* section below).

Table 12 *Response Time Means in Milliseconds for Scale Length*

		Dimension of Teaching Effectiveness			
		Rapport with Students	Course Value	Course Organization and Design	Fairness of Grading
5-point Scale	M	3989.72	3843.44	3965.33	3902.66
	SD	502.18	126.59	117.67	132.47
11-point Scale	M	3990.17	3867.46	3979.56	3930.75
	SD	129.17	121.21	112.78	119.80

For all of the teaching effectiveness dimensions, there were significant main effects for *item wording* on response times: *Course Design* ($F_{[1, 409]} = 3.46$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$); *Fairness of Grading* ($F_{[1, 409]} = 6.11$, $p < .01$, $\eta^2 = .03$, $\omega^2 = .01$); *Course Value* ($F_{[1, 409]} = 3.83$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$); and *Rapport with Students* ($F_{[1, 409]} = 7.88$, $p < .001$, $\eta^2 = .03$, $\omega^2 = .02$). Despite, wording interacting with other included variables (see the '*Significant Interactions for Response Times'* section below), these results were interpreted to test the *a priori* hypothesis that negative wording condition would result in longer response times than the positive wording condition. Linear contrasts for each the four dimensions revealed that for the *Course Value* factor, positive wording resulted in shorter response time means, when compared to negative wording. Also, for the *Fairness of Grading* factor, it was noted that positive wording resulted in shorter response time means when compared to negative and mixed wording conditions. Taken together these results provide support for the hypothesis that negative wording

results in longer responses times than positive wording of items. For the *Course Design* and *Rapport with Students* dimensions, positive wording resulted in shorter response time means when compared to the mixed wording condition. For all of the teaching effectiveness dimensions there were no differences between negative wording and the mixed wording condition for response time means (the response time means for each condition can be found in Table 13). However, for all of these dimensions, the results were qualified by interactions with other experimental conditions that were utilized (see the '*Significant Interactions for Response Times'* section below).

Table 13 *Response Time Means in Milliseconds for Item Wording*

		Dimension of Teaching Effectiveness			
		Rapport with Students	Course Value	Course Organization and Design	Fairness of Grading
Positive Wording	M	3939.62	3836.50	3956.74	3888.46
	SD	143.96	130.31	119.69	129.08
Negative Wording	M	3983.64	3874.38	3975.80	3926.79
	SD	127.28	126.17	118.49	126.22
Mixed Wording	M	4046.67	3856.66	3985.21	3935.78
	SD	595.61	114.03	106.37	121.04

Significant Interactions for Response Times

For the *Fairness of Grading*, *Course Design* and *Rapport with Students* dimensions there were significant interactions between the *type of course* being evaluated and the *scale orientation* on response times: *Course Design* ($F_{[1, 409]} = 6.37$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$); *Fairness of Grading* ($F_{[1, 409]} = 9.16$, $p < .01$, $\eta^2 = .02$, $\omega^2 = .02$); and *Rapport with Students* ($F_{[1, 409]} = 4.75$, $p < .05$, $\eta^2 = .01$, $\omega^2 = .01$). Post Hoc analyses (using a Bonferroni correction) showed that for all three dimensions, bidirectional response options resulted in longer response times than unidirectional response options when participants were responding for liked courses. However, there were no significant

differences in response times for response format, when participants were responding to a disliked class (the response time means for each condition can be found in Table 14).

Table 14 *Response Time Means in Milliseconds for Scale Orientation by Course Type*

			Teaching Effectiveness Dimension		
			Rapport with Students	Course Organization and Design	Fairness of Grading
Unidirectional Scale	Liked Course	M	3911.22	3890.83	3865.10
		SD	160.20	184.37	161.76
	Disliked Course	M	3964.68	3976.24	3930.38
		SD	158.23	174.73	149.32
Bidirectional Scale	Liked Course	M	4023.74	3961.84	3911.24
		SD	527.72	158.29	150.24
	Disliked Course	M	3995.53	3982.66	3930.52
		SD	156.30	169.35	130.51

For all of the teaching effectiveness dimensions there were significant interactions between the *type of course* being evaluated and the *item wording* on response times: *Course Design* ($F_{[1, 409]} = 3.39$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$); *Fairness of Grading* ($F_{[1, 409]} = 13.25$, $p < .001$, $\eta^2 = .06$, $\omega^2 = .03$); *Course Value* ($F_{[1, 409]} = 8.30$, $p < .001$, $\eta^2 = .04$, $\omega^2 = .02$); and *Rapport with Students* ($F_{[1, 409]} = 3.15$, $p < .05$, $\eta^2 = .02$, $\omega^2 = .01$). Contrasts were conducted using a Bonferroni correction to explore the interactions. For the *Course Design, Fairness of Grading* and *Course Value* dimensions, it was noted that positive wording resulted in shorter response times, when compared to the negative and mixed wording conditions for liked classes. There were no significant differences in response times between negative and mixed wording conditions. Also, there were no significant differences in response times based on wording conditions for disliked classes. Similar to the other factors, for the *Rapport with Students* factor positive wording resulted in shorter response times when compared to mixed wording for liked classes. However, response

times did not significantly differ between positive wording and negative wording for liked classes. Also, there were no significant differences in response time means between negative and mixed wording conditions. Like the other results, *item wording* response time means did not significantly differ for disliked classes. Taken as a whole, these results suggest that positive wording for liked classes are susceptible to quicker response times, when compared to other wording conditions. But when individuals respond to disliked classes, the wording of items does not seem to have a significant effect. The relevant response time means for each condition can be found in Table 15.

Table 15 *Response Time Means in Milliseconds for Item Wording by Course Type*

			Teaching Effectiveness Dimension			
			Rapport with Students	Course Value	Course Organization and Design	Fairness of Grading
Positive Wording	Liked Course	M	3893.80	3787.86	3884.30	3831.87
		SD	176.20	153.51	177.05	149.00
	Disliked Course	M	3964.71	3866.76	3981.89	3927.60
		SD	163.95	155.13	183.98	146.20
Negative Wording	Liked Course	M	3971.55	3867.73	3935.17	3919.39
		SD	145.80	153.19	178.28	152.71
	Disliked Course	M	3981.48	3866.28	3980.39	3923.25
		SD	152.56	140.97	165.05	133.12
Mixed Wording	Liked Course	M	4035.01	3830.57	3958.46	3913.82
		SD	627.17	133.43	163.75	156.82
	Disliked Course	M	3993.53	3867.04	3975.95	3940.22
		SD	156.34	145.61	166.92	141.13

On both the *Course Value* ($F_{[1,406]} = 10.02$, $p < .01$, $\eta^2 = .02$, $\omega^2 = .02$) and *Course Design* ($F_{[1, 409]} = 4.3$, $p < .05$, $\eta^2 = .01$, $\omega^2 = .01$) dimensions, there were significant three-way interactions between the *type of course* evaluated, *scale length* and the *scale orientation* for response times. Visual representation of the interaction for *Course Value* can be found in Figure 1 and the interaction for *Course Design* can be found in Figure 2.

Via visual interpretation and contrasts (using a Bonferroni correction), for both teaching effectiveness dimensions it was noted that for liked classes with short response scales (5-point), bidirectional response options resulted in longer response times from participants when compared to unidirectional response options. Also, there were no significant response time differences between scale orientation conditions when long response scales (11-point) were implemented. For disliked classes, long response scales (11-point) with bidirectional response options resulted in longer response times when compared to unidirectional response options. Furthermore, there were no significant response time differences between scale orientation conditions when short response scales (5-point) were implemented. These results suggest that for liked classes bidirectional response options may cause participants to put more time into their responses when a short scale is implemented, but not for a longer scale. These results change though for disliked classes, where bidirectional response options only result in more time spent responding when a longer scale is used. The relevant response time means for each condition can be found in Table 16. Due to the large number of hypotheses and analyses conducted, a brief summary of all hypotheses and the relevant findings (including effect sizes) can be found in Table 17.

Table 16 *RT Means in Milliseconds for Scale Length and Format by Course Type*

			Course Value		Course Organization and Design	
			Unidirectional Scale	Bidirectional Scale	Unidirectional Scale	Bidirectional Scale
Liked Course	5-point Scale	M	3771.37	3863.45	3868.82	3960.18
		SD	154.51	153.88	194.89	165.51
	11-point Scale	M	3818.39	3861.33	3912.84	3963.45
		SD	140.92	133.17	171.31	151.73
Disliked Course	5-point Scale	M	3848.24	3858.42	3986.29	3959.61
		SD	156.64	125.55	178.13	166.54
	11-point Scale	M	3849.06	3911.51	3966.19	4005.06
		SD	150.11	146.13	171.51	169.35

Table 17 *Summary of Hypotheses and Corresponding Results*

Hypothesis	Expected Result	Actual Result
1. Significant relationships between responses and response times (RTs).	Due to exploratory nature of the hypothesis no specific directionality was proposed.	Significant relationships found between RTs and responses, based on the type of class being evaluated (-.139 to -.259).
2. Two-way interaction between *item wording* and *scale orientation* for responses.	*Positive wording* with *unidirectional scales* would result in more positive responses than other conditions.	No significant interaction found. Main effect found for *item wording.*
3. Main effect for *scale length* on responses.	*11-point scales* would result in more positive responses than *5-point scales*.	*5-point scales* resulted in more positive responses than *11-point scales* (ω^2= .01 to ω^2= .02).
4. Two-way interaction between *item wording* and *scale orientation* for response times (RTs).	*Mixed wording* with *bidirectional scales* would result in longer RTs than other conditions.	No significant interaction found. Main effects found for *item wording* and *scale orientation.*
5. Main effect for *item wording* on responses times (RTs).	*Negative wording* would result in longer RTs than *positive wording*.	*Positive wording* resulted in shorter RTs than *negative* and *mixed wording* (ω^2= .01 to ω^2= .02).
6. Main effect for *scale length* on response times (RTs).	*11-point scales* would result in longer RTs than *5-point scales*.	*11-point scales* resulted in longer RTs than *5-point scales* (ω^2= .01).

Figure 1 *Scale Length by Response Scale Format for Liked and Disliked 'Course Value'*

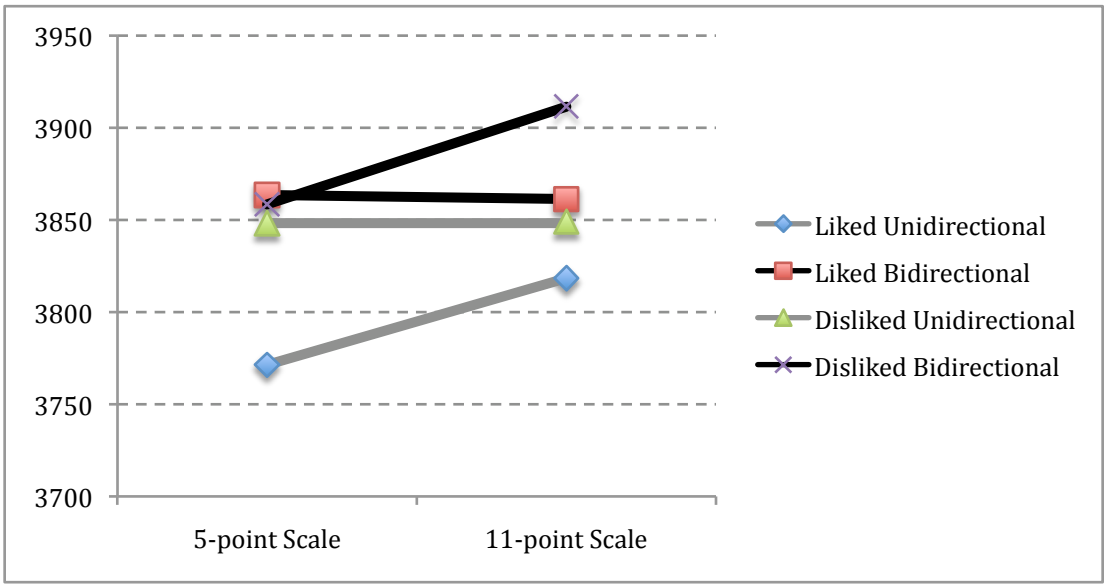
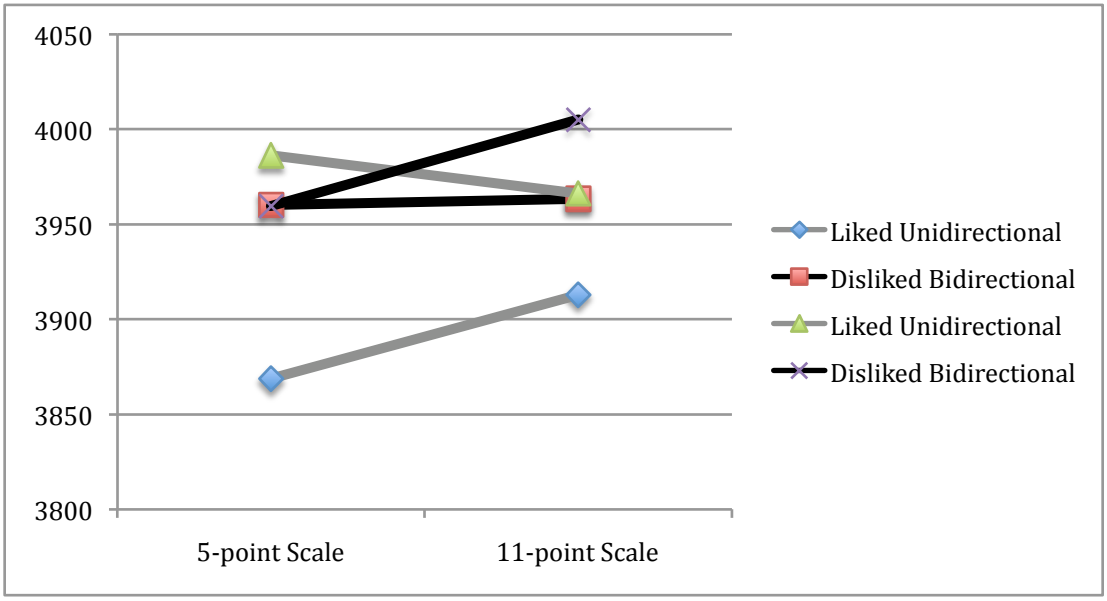


Figure 2 *Scale Length by Response Scale Format for Liked and Disliked 'Course Design'*



Discussion

As expected, responses provided were significantly related to the amount of time participants took when responding to items (hypothesis #1). Exploration of these results showed rather clearly that for liked classes the more time taken to respond, the less positive the responses were. Also, for some of the teaching dimensions, there was a positive relationship between the amount of time spent responding and the responses provided. Specifically, for disliked experiences more thought may lead to more positive (or less critical) responses from participants. Although these relationships were relatively weak (ranging in magnitude from -.139 to -.259), they do provide novel evidence that response times are related to how individuals' respond to questions and that the type of experience being evaluated can influence this relationship.

Although the second *a priori* hypothesis was not supported (there were no two-way interactions between *item wording* and *scale orientation*), there was found to be a significant main effect for *item wording* on the types of responses submitted. In particular, it was noted that for one factor (*Fairness of Grading*) positive wording resulted in more positive responses when compared to both negative wording and mixed wording conditions. A similar but not identical result was found for another factor (*Course Value*), where positive wording elicited more positive replies only when compared to the negative wording condition. Although these results were not predicted, they do coincide with previous research, where it has been suggested that negative and mixed wording of items results in less positive responses when compared to positive wording (Nunnally & Bernstein, 1994). Further thought is given to why the hypothesized interaction did not occur later in this paper.

Across all of the teaching effectiveness dimensions, there was a main effect for the *scale length* implemented (hypothesis #3). Contrary to the outlined *a priori* hypothesis and prior research, it was noted that shorter scales (5-point) resulted in more positive responses than longer scales (11-point). Although this result diverges from prior research (Ing and Jackson 2007; 2008), it does seem to logically coincide with the results found in other areas of this study (in particular the response time results) and it is discussed in more detail later. The fact that an 11-point scale was utilized, does deter from the ability to directly compare these results to previous research (that used 9-point scales), nevertheless the results suggest that very long scales may result in participants providing more negative (or critical) responses overall. This may be due to the granularity requirements (Ackerman & Goldsmith, 2008; Goldsmith et al., 2002) of the longer response options, which in turn results in less positive responses.

Also, an interaction was noted between the *scale length* used and the *type of course* being evaluated. Exploration of this interaction showed that shorter scales (5-point) resulted in more positive responses than longer scales (11-point), for disliked classes (but not for liked classes). This result provides evidence that for disliked experiences, shorter response scales may result in more positive responses from individuals. Thus, a longer scale, with increased response granularity may exacerbate the negativity of responses provided by participants for disliked experiences.

An unanticipated effect was noted for the *type of class* being evaluated, where disliked classes resulted in longer response times, for three of the teaching effectiveness dimensions (*Course Design*, *Fairness of Grading* and *Course Value*). Interestingly, this result suggests that the participants put more thought (and cognitive processing) into their responses when they were replying to a disliked experience. Furthermore, the type of

classes evaluated interacted with other variables that were included in the analyses. This provides evidence that the type of experience being evaluated can influence how individuals process their responses.

There were no significant interactions found between *item wording* and the *scale orientation* on response times (hypothesis #4). However, as predicted there was found to be a main effect for the type of *item wording* implemented on the response times posted (hypothesis #5). As predicted, for two of the teaching effectiveness dimensions there were significant differences in response times between positive wording and negative wording conditions, where positive wording resulted in shorter response time means. In addition, there was a significant interaction between *item wording* and the *type of course* being evaluated. Examination of this interaction effect elucidated the impact that item wording had on response times. In particular it was noted that in most cases (for the *Course Design*, *Fairness of Grading* and *Course Value* factors) positive wording resulted in shorter response times than both negative wording and mixed wording conditions for liked but not disliked classes. In one case (*Rapport with Students*) positive wording resulted in shorter response times than mixed wording for liked classes. These results seem to suggest that for liked experiences, individuals put less thought into their responses when items are worded in a positive manner when compared to negative or mixed item wording.

Despite the lack of an interaction between *item wording* and *scale orientation* (hypothesis #4), there was found to be a main effect for *scale orientation* on all of the teaching dimensions, where bidirectional response options resulted in longer response times than unidirectional response options. Furthermore, in some cases the orientation of response options interacted with other variables included in the analyses. In particular,

there was an interaction between the *scale orientation* implemented and the *type of course* (for *Course Design*, *Fairness of Grading* and *Rapport with Students*) being reviewed, where bidirectional response options resulted in longer response times for liked classes but not disliked classes.

As expected, a significant main effect was found for *scale length* based on response times (hypothesis #6) on two of the teaching effectiveness dimensions (*Course Value* and *Fairness of Grading*). Exploration of these results confirmed that longer response scales resulted in longer response times being posted. This result supports the idea that longer response scales, resulted in participants providing more thought when providing their responses. In two instances (*Course Value* and *Rapport with Students*) a three-way interaction was found between the *type of course* being evaluated, the *scale orientation* and *scale length*. This result showed that for liked experiences using shorter response scales (5-point), bidirectional response options resulted in longer response times. However, for disliked experiences and long response scales (11-point), unidirectional response options resulted in longer response times from participants. This result seems to suggest that for liked experiences, short scales may result in less thought from respondents, unless bidirectional response options are used. However, for disliked experiences using long response scales, individuals are already placing more thought in their responses, but bidirectional response options requires further contemplation prior to responding.

Broad Implications

Taken as a whole the results of this study provide for some rather intriguing implications. Firstly, it is important to consider that response time data is often thought of as a proxy for cognitive processing or cognitive effort (Bassili & Scott, 1996; Matlin,

2002; Yan & Tourangeau, 2008). With this in mind, interpretation of the response time data can allow insight into how cognitive processing can be influenced by variations in questionnaire design, and how this may be related to the actual responses submitted.

At this point one may wonder, why it is important to observe when respondents place greater amounts of time into their responses. More precisely, why is it that a longer response time should be viewed as superior to a shorter response time? It is important to note that longer response times may not always be a desirable outcome when individuals are responding to questions. Conceivably, if one has a very clear recollection of relevant events and is already sure of their responses, their response time would not be as long as someone who is having difficulty recalling relevant memories and is not as sure of their response. Such response time differences may be attributable to individual and experiential differences. However, overly quick response times, such that the individual did not even have enough time to read or understand the question presented, could represent blatant patterns of response bias (Barnette, 1999; 2000; Schwarz, 1999). Furthermore, it has been discussed that evaluations in educational settings are often a victim of unattending response biases, and consequently are likely to be characterized by overly fast response patterns (Barnette, 1999; 2000: Krosnick, 1991). Therefore, differences in response times of approximately 10s of milliseconds, potentially provides an ecologically valid way of gauging the impact of questionnaire design on cognitive processing. Furthermore, longer response times would be particularly desirable in the case of this study, as it was the hoped that variations in question design may force participants to 'attend' more when responding to SETE.

Thus, the fact that response time data was related to responses submitted is an important step in measuring the cognitive processes individuals undergo when responding

to questionnaires. Due to relatively weak correlations, it would be overzealous to suggest that response times are a substantial predictor of responses, as other extraneous factors are likely influencing the response times submitted (e.g. motivation to respond, cognitive apathy, other environmental distracters etc.). Suffice to say, that despite weak relationships, the fact that response times and responses were significantly related suggests that response time data can be useful in understanding why and how people respond the way that they do.

Additionally, the fact that this relationship appears to change based on the type of experience being evaluated, suggests that the context of the experience (liked or disliked) being evaluated influences how humans process and respond to questionnaires. This finding is substantiated by the finding that disliked classes resulted in longer response times when compared to liked classes. Furthermore, this difference in cognitive processing based on the type of experience seems to play an influential role in the way that questionnaire design characteristics impact respondents. That is, for the most part liked classes seem to be the most susceptible to the influences of questionnaire design. This is made evident by the fact that variations in response scale design and item wording interacted with the type of experience being evaluated and in most cases differences were found only when liked experiences were being evaluated. In theory, this may occur because participants address responding to each experience differently. In the case of a disliked class, the individual may feel that they must be critical of their own thinking when criticizing or providing negative information about others. Thus when responding to a disliked experience more thought is given to each answer, in an attempt to be 'fair' and 'accurate'. On the other hand, for liked experiences, individuals may feel that submitting errantly positive information is less detrimental and as a result they are more susceptible

to cognitive heuristics and acquiescence type response patterns. Further, this may be attributable to a 'halo effect', where respondents recall one positive experience and in turn judge all other aspects of the experience to be positive (Asch, 1946; Thorndike, 1920). Consequently, respondents may not feel that 'positive' experiences require as much cognitive critique and therefore their response time is comparatively shorter than that of negative experiences.

One result that is better understood by examining response time data is the effect for scale length on responses and response times submitted. As noted early, longer response scales resulted in less positive responses from participants when compared to shorter response scales. In previous studies the opposite was found, where longer response scales yielded more positive responses than short scales. However, when the response time data and prior cognitive theory is taken into consideration, a clearer depiction of the influence long response scales have on responding is made available. In particular, longer response scales resulted in longer response times from participants. This result coincides with recent research in cognition which suggests that increasing the grain size of response options should result in more thought from the individual (Ackerman & Goldsmith 2008; Goldsmith et al., 2002). Under this line of thought, it stands to reason that individuals exposed to longer response scales, should provide more thought to their responses, deterring response heuristics and acquiescence bias. Consequently, the finding that longer response scales yields longer responses times and more negative responses seems to fit well with recent cognitive theory.

Additionally, when the type of experience being evaluated and variations in response formatting are thrown into the mix, it appears that liked experiences with shorter scales require more thought when bidirectional options are provided. This result seems to

suggest that short scales when evaluating liked classes may be more susceptible to cognitive heuristics, as the information is liked and the scale is comprised of coarse granularity, which requires less thought. But if the response options alternate in direction, participants are forced to put more thought into the responses they submit. In the case of disliked experiences longer scales required more thought when bidirectional options were provided. This result seems to suggest that disliked classes, with long bidirectional response scales present a constellation of factors that require considerably more thought from participants when compared to other conditions. It terms of the impact that these differences in thinking have on responding, it appears that disliked experiences combined with longer response scales, results in participants submitting more negative responses. This is likely due to the combined cognitive influence of evaluating a disliked experience while dealing with the increased granularity of responses options provided. When taken as a whole these results seem to provide evidence that response scale variations impact the way people think about their responses in a different way than other aspects of questionnaire design (e.g. item wording, item order etc.).

Continuing with the theme of separation between cognitive processes utilized when responding, the predictions that item wording and response orientation would interact when it came to responses and response times was not supported. Interestingly enough, item wording acted as a main effect and interacted with the type of course being evaluated, for both responses and response times. The results involving the impact of wording seem to support the prior literature in this area, but with an additional caveat: the type of experience being evaluated makes a difference in how individuals think about the questions. That is, positively worded items seem to result in more positive responses than other wording options. However, in terms of cognitive processing, for liked experiences

use of negative or mixed wording seems to require individuals to provide more thought than they would using positively worded items. This may be because participants are more susceptible to cognitive heuristics when they reply to a liked experience, thus variations in item wording requires increased thought from the individual for these types of experiences. Alternatively, as noted in previous research, when words are incongruent with the emotional affect of an experience, there is often increased response latency by individuals (Duscherer, Holender, & Molenaar, 2008). Although it is unlikely that this study caused participants to experience variations in affect based on recalled evaluation of a past course, this does provide a potential alternative explanation for this finding.

Despite the lack of any interactions between item wording and response item characteristics (as hypothesized), with further contemplation this finding fits well with previous theory. Specifically, despite the fact that each stage of cognitive processing works together to generate a reply, the processes themselves seem to be unique and separate from one another. Conceivably then, response scales should not interact with how items are read and understood, in the same way that item wording should not interact with how responses are encoded. As a result, the prediction that item wording and response scale characteristics would interact seems to be a nearsighted perception of the cognitive processes involved in responding. The evidence in this study suggests, that one could alter the way an item was worded without concern that it would influence the way response scale arrangement would be perceived by respondents. This is not to say that item wording or scale variations are more or less important when it comes to processing and responding, instead that each is a distinct and unique part of the process that should be considered when designing a questionnaire.

Practical Implications

When examining the results as a whole, this study provides practical information for both those designing questionnaires, and those looking to understand how individuals think and respond to questionnaires. Regarding the ongoing debate in the literature pertaining to the use of mixed wording of items, the results of this study support the theory that positively worded items result in less thought (and potentially increased cognitive heuristics) and consequently more positive responses from individuals. However, this study added a level of nuance to this debate that was previously not discussed in the literature. In particular, that using negative wording or mixed wording may be most helpful in causing increased cognitive processing when a liked experience is being evaluated. Thus, based on these findings and past literature, it appears that using mixed wording (or even negative wording) is advisable, to deter individuals from relying on apathetic cognitive processing techniques.

The results of this study pertaining to response scale length, diverge from previous findings, but provide pragmatic directions for those designing questionnaires. The finding that longer response scales resulted in more thought from individuals and less positive responses suggests that increase response granularity in questionnaires may deter individuals from depending on cognitive heuristics when responding. Consequently, when the results of this study and cognitive theory pertaining to responding are considered, it seems prudent for researchers to consider longer responses scales when developing questionnaires.

In the case of bidirectional response options, the results paint a murky image. When cognitive processing is considered, it appears that bidirectional response options seem to serve a useful function, as they result in more cognitive processing from

participants (i.e. longer response times). This is most notably effective for liked experiences, where individuals seem most vulnerable to cognitive shortcuts. In an ideal situation, respondents will avoid response heuristics and provide an appropriate amount of time to fully and accurately answer each question (especially in the case of SETE). Thus, bidirectional response options may assist on this front, by forcing participants to give more thought (or increased response time) to each individual item. However, this increase in cognitive processing did not seem to translate into differences in how the individual responded to items. One possible explanation for this is that participants may have already determined their response to the item and altering the response format does cause them to think, but not necessarily to reflect on the substance of the answer they are going to provide.

When bidirectional scales are compared to scale length (another manipulator of response characteristics), the way an individual encodes their response seems to be altered by increasing the amount of granularity in response options and not necessarily by changing the direction of the options. This study provides preliminary evidence that there is no difference in the responses provided based on the response option orientation implemented; however more thought seems to be given when finding the appropriate response, based on the direction of the scale. These results do not necessarily rule out the utility of bidirectional response options, but it does clarify the role it has on how individuals think about and respond to questions. As it stands, it would be difficult to recommend the use of bidirectional response options to deter response biases. However, future research is needed to better establish the influence this design variation has on cognition and responding.

Limitations

Despite significant findings and ties with past literature, there are some limitations to this study that should be noted. In particular, the results of this study for the most part yielded small effect sizes. This result is somewhat unusual as the effect sizes noted, even for expected effects, were markedly smaller than the effect sizes reported in prior research (Ing and Jackson, 2008). Consequently, the results and the implications of the findings should be considered in the proper context. That is, with smaller effect sizes, variations in questionnaire design may not have a marked effect on the way individuals cognitively process or respond to questionnaire items. Alternatively, it is possible that the smaller effect sizes noted might be the result of design characteristics that were implemented in the study. Thus further research is required to substantiate and clarify the findings of this study.

One design characteristic that may have influenced the way individuals responded to items is the use of computer-based administration. Some preliminary research has found that whether questionnaires are administered via the Internet or paper does not seem to result in differences in how individuals respond to questions (Puklavetz, Rodzon, & Howell, 2009). However, there does not seem to be a consensus as to whether the type of administration (online or paper) interacts with the effects of questionnaire design characteristics. Thus it is possible that the effects of questionnaire characteristics may have been different if this study were administered by paper. Furthermore, although the study was administered via computer it was conducted in a laboratory setting with multiple participants per session. As a result, it is possible that external factors (e.g. additional noise) could have influenced the way individuals processed and consequently responded to the items presented. However, this approach does emulate actual scenarios

where SETE are conducted, providing an ecologically valid way of testing the research questions.

Also because this study implemented 11-point responses scales as a 'long' scale, it is challenging to compare this finding to that of prior research, which implemented 9-point response scales (Ing & Jackson, 2008). The discrepancy between the current findings and those reported in past research is a curious one. A possible explanation for these differences may be that 11-point scales are less frequently implemented in questionnaires as compared to 9-point scales. Thus, it is possible that the 11-point scale resulted in more cognitive processing and consequently less positive responses because it was novel to participants. However, this result does seem better explained by cognitive research which suggests that increased granularity of response options should require more thought from individuals (Ackerman & Goldsmith 2008; Goldsmith et al., 2002). Nevertheless, further research is required to fully understand why this discrepancy occurred.

An additional limitation within this study is how the type of course within-subjects variable was administered to participants. In particular, participants were asked to evaluate a 'liked' class and a 'disliked' class experience. This open-ended structure may have made participants feel required to provide an overtly negative response (for a disliked class) or positive response (for a liked class) regardless of their actual experience. This component may have acted as a demand characteristic, whereby participants felt that the class experiences must fit into a homogeneous grouping of what is a good a bad class. In the contrary, this approach does appear to allow for ecological validity, as participants were able to determine the class they wished to review based on their own past experiences.

Finally, the inability to differentiate the amount of time devoted to each individual component of cognitive processing limits the ability to interpret some of the findings in this study. In particular, this may be a concern for the results that involved item wording, as the negatively worded items often included an additional word (e.g. 'not') in the phrasing of the question. Interestingly, the positive wording condition, in most cases, resulted in shorter response times than negative and mixed wording conditions; where negative and mixed wording conditions did not differ. This result may suggest that the shorter response times are characterized by the semantic directionality of the items, and not necessarily the number of words in the sentence. That is, because the mixed wording condition contained an even number of negatively worded and positively worded items, and positive wording resulted in shorter response times; it appears that the semantic direction of the items may be the reason for these differences. This conclusion is further supported on theoretical grounds, as respondents also provided more positive responses to questions when they were positively worded. Thus, it appears that participants put less thought into positively worded items and in turn they appear to be more reliant upon acquiescent type response patterns. However, further research is needed to fully understand this phenomenon.

Future Directions

As noted earlier, despite differences in response times based on specific design characteristics, there is no way to know 'exactly' how participants were cognitively processing items. This study does provide preliminary evidence, by using response times as a proxy for cognitive processing, that design characteristics have separate influences on specific cognitive processes that are involved in responding. Thus future research, that is able to measure response times for the different processes in responding may be

extremely useful in understanding exactly how variations in design influence the way individuals process questions.

One interesting direction that could be taken from this study would be to test the validity of responses based on a controlled and experimentally manipulated experience (e.g. using videos or vignettes). Through this approach it would be possible to see if questionnaire design characteristics influence the accuracy of the responses provided. Furthermore, this would allow insight into the impact cognitive processing has, based on questionnaire variations, on the accuracy of responses that are submitted. This could also be extended to other avenues, where actual behaviour could be measured as an outcome variable on attitude or opinion scales. That is, the relationship between response time, responses and questionnaire variations could be examined as they relate to the actual behaviours that individuals take.

In addition, it was noted from this study that how the individual felt about their experience (liked or disliked), interacted with the questionnaire design characteristics they were given as it related to the responses and the response times that were submitted. An interesting approach may be to include a single item in questionnaires to distinguish the type of responses that would follow. For example, by adding an item that asks respondents to initially state whether the experience they will be evaluating was a 'liked one' or a 'disliked one' may prove to be useful to researchers. This approach may then allow researchers to identify those who are more prone to response biases (typically those responding to liked experiences), thus allowing them to potentially manipulate the type of questionnaire that will be given to the participant based on this information. However, future research would need to be conducted to determine the utility of this approach in real world circumstances.

In conclusion, it appears that variations in questionnaire design can impact the way people think and respond to questions asked of them. As a result, researchers should be diligent when creating measurement instruments, so that they obtain the most accurate information possible. Further research involving how questionnaire design and implementation variations affect cognitive processing and responses submitted would greatly benefit future questionnaire development endeavors.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting-with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1224-1245.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A Structural Modeling Approach. *Public Opinion Quarterly, 48*, 409-42.

Allport, G. W. (1937). The functional autonomy of motives. *American Journal of Psychology, 50*, 141-156.

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

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

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290.

Bassok, M., & Trope, Y. (1983-1984). People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic. *Social Cognition, 2*, 199-216.

Barnette, J. J. (1999). Nonattending respondent effects on internal consistency of self-administered surveys: A monte carlo simulation study. *Educational and Psychological Measurement, 59*(1), 38-46.

Barnette, J. J. (2000). Effects of stem and likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*(3), 361-370.

Barnette, J. J. (2001). Likert survey primacy effect in the absence or presence of negatively-worded items. *Research in the Schools, 8*(1), 77-82.

Bassili, J. N. (1993). Response latency versus certainty as indexes of the strength of voting intentions in a CATI survey. *Public Opinion Quarterly, 57*, 54-61.

Bassili, J. N., & Roy, J. (1998). On the representation of strong and weak attitudes about policy in memory. *Political Psychology, 19*, 669-681.

Bassili, J., & Scott, S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*, 390-399.

Beatty, P., Herrmann, D., Puskar, C., & Kerwin, J. (1998). "Don't know" responses in surveys: is what I know what you want to know and do I want you to know it? *Memory, 6*(4), 407-426.

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

Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research, 14*, 280-288.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality and Quantity, 38*, 17-33.

Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly, 55*, 50-79.

Burdsal, C. A., & Bardo, J. W. (1986). Measuring students' perceptions of teaching: Dimension of evaluation. *Educational and Psychological Measurement, 56*, 63-79.

Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology, 51*(5), 1032-1043.

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

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

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

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*(3), 472-517.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*, 151-174.

Couch, A., & Keniston, K. (1961). Agreeing response set and social desirability. *Journal of Abnormal and Social Psychology, 62*, 175-179.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*, 401-415.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354, 1960.

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

Csiksczentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. New York: Harper & Row.

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

Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology, 22*(1), 113-120.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.

Duscherer, K., Holender, D., & Molenaar, E. (2008). Revisiting the affective simon effect. Cognition & Emotion, *22*(2), 193-217.

Eccleston, C., McCracken, L. M., Jordan, A., & Sleed, M. (2007). Development and preliminary psychometric evaluation of the parent report version of the Bath Adolescent Pain Questionnaire (BAPQ-P): A multidimensional parent report instrument to assess the impact of chronic pain on adolescents. *Pain,* 131, 48–56, 2007.

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

Fandt, P. M., & Stevens, G. E. (1991). Evaluation bias in the business classroom: Evidence relating to the effects of previous experiences. *Journal of Psychology: Interdisciplinary and Applied, 125*(4), 469-477.

Field, A. (2005). *Discovering statistics using SPSS (2nd ed.)*. Thousand Oaks, CA, US: Sage Publications, Inc.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*, 107-119.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. Journal of Experimental Psychology: General, *131*(1), 73-95.

Gray, G., & Guppy, N. (1999). *Successful surveys: Research methods and practice.* Toronto, ON: Harcourt Brace.

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

Hearst, E. (1991). Psychology and nothing. *American Scientist*, *79*, 432-443.

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

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expending the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92*, 555–566.

Hulin, C., Netemeyer, R., & Cudeck, R. (2001). Methodological and statistical concerns of the experimental behavioral researcher. *Journal of Consumer Psychology*, *10*(1/2), 55-58.

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

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

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

Ing, P. G., & Jackson, D. L. (2008). *Instructions, Response Anchors, and Scale Length: How Do Variations Affect Student Evaluations of Teaching Effectiveness?* Unpublished master's thesis, University of Windsor, Windsor, Ontario, Canada.

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

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

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

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

Kellogg, R. T. (1986). Designing idea processors for document composition. *Behavior Research Methods, Instruments & Computers, 18*(2), 118-128.

Kilmann, R. H., & Thomas, K. W. (1975). Interpersonal conflict handling behavior as reflections of Jungian personality dimensions. *Psychological Reports, 37*, 971-980.

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

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

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

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

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

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology, 123*(3), 297-315.

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

Langbein, L. I. (1994). "The Validity of Student Evaluations of Teaching." PS Political Science & Politics, *27*(3), 545-53.

Leary, M. R., & Kowalshi, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin, 107*, 34-47.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73-79.

Lueck, T. L., Endres, K. L., & Caplan, R. E. (1993). The interaction effects of gender on teaching evaluations. *Journalism Education*, *48*, 46-54.

Matlin, M. W. (2002). *Cognition (5th ed.)*. Orlando, FL, US: Harcourt Brace College Publishers.

Marsh, H. W. (1982). Validity of Students' Evaluations of College Teaching: A Multitrait-Multimethod Analysis. *Journal of Educational Psychology*, *74*(2), 264-279.

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

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

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

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

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

Means, B., & Loftus, E. (1991). When personal history repeats itself: Decomposing memories for recurring events. *Applied Cognitive Psychology*, *5*, 297-318.

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

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (2nd Ed.). New York: McGraw-Hill, 1978.

Osberg, T. M., Haseley, E. N., & Kamas, M. M. (2008). The MMPI-2 Clinical Scales and Restructured Clinical (RC) Scales: Comparative psychometric properties and relative diagnostic efficiency in young adults. *Journal of Personality Assessment*, *90*, 81-92.

Perkins, D., Guerin, D., & Schleh, J. (1990). Effects of grading standards information, assigned grade, and grade discrepancies on student's evaluations. *Psychological Reports, 66*(2), 635-642.

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19*(3), 291-312.

Piolat, A., Olive, T., Roussey, J., Thunin, O., & Ziegler, J. C. (1999). SCRIPTKELL: A tool for measuring cognitive effort and time processing in writing and other complex cognitive activities. *Behavior Research Methods, Instruments & Computers, 31*(1), 113-121.

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

Puklavetz, S., Rodzon, K., & Howell, R. (2009, May). Impact of the Internet on survey measurements. Poster session presented at the annual meeting of the Association for Psychological Science, San Francisco, CA.

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

Rea, L. M., & Parker, R. A. (1997). *Designing and conducting survey research.* San Francisco, CA: Joessey-Bass Inc.

Reber (1996). Rating scale response formats: Does number of response options make a difference? [Abstract]. *Dissertation Abstracts International: Section B: The Sciences and Engineering. 57*(5-B), pp. 3450

Robinson J. P., Shaver P. R., & Wrightsman L. S. (1991). *Measures of personality and social psychological attitudes.* San Diego: Academic Press

Rossi, P. H., Wright, J. D., & Anderson, A. B. (1983). *Handbook of survey research.* New York: Academic Press.

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement, 51*(1), 67-78.

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

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

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition Special Issue: What is an Attitude?, 25*(5), 638-656.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology. Special Issue: Cognitive Psychology and Survey Methodology: Nurturing the Continuing Dialogue between Disciplines, 21*(2), 277-287.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E. & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570-582.

Sears, D. O. (1983). The person-positivity bias. *Journal of Personality & Social Psychology, 44*, 233-250.

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

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning & Verbal Behavior, 15*(2), 143-157

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

Simon, H. (1979). A behavioral model of rational choice. In H. Simon (Ed.), *Models of thought* (pp. 7-19). New Haven: Yale University Press.

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

Stevens, J. P. (2002). *Applied Multivariate Statistics for the Social Sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research, 15*(1), 3-7.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly,* 68, 368-93.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299-314.

Tourangeau, R., Rasinski, K. A., & D'Andrade, R. (1991). Attitude structure and belief accessibility. *Journal of Experimental Social Psychology, 27*(1), 48-75.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York, NY; Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Q, 60*, 275-304.

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

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 255-274

Wason, P. C., & Jones, S. (1963). Negatives: Denotation and connotation. *British Journal of Psychology, 54*(4), 299-307.

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

Williams, J. D. (1999). *The teacher's grammar book.* Mahwah, NJ; Erlbaum.

Williams, M.D., & Hollan, J.D. (1981). The process of retrieval from very long term memory. Cognitive Science, *5*(2), 87-119.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology, 22*(1), 51-68.

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

Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology, 68*(1), 52-60.

Appendix A

You are being asked to fill out a questionnaire about **two classes** that you took *last semester*: one that you **liked** and one that you **disliked**. You will be asked a series of questions relating to past university class experiences. These questions will be presented to you one at a time. Please answer to the best of your ability, by selecting the most accurate response and then clicking next. Please note that you will be unable to return to previous questions once you have moved on (please do not attempt to go back or refresh the screen).

Appendix B

The followings will be presented to participants in counterbalanced order:

Please answer the following questions for a class that you completed last semester that you **liked**.

Please answer the following questions for a class that you completed last semester that you **disliked**.

Appendix C.1

Positively Worded Items:

The following questions will be presented with either *five* or *eleven* scale points (presented using radio buttons), with either *unidirectional* or *bidirectional* response options and the type of class being evaluated will be presented in counterbalanced order (i.e., *Liked Class* will appear in half of the cases and second for the other half).

Questions

(Liked/Disliked) Class

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

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

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

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

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

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

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

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

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

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

11. Leave blank and click next.

12. I usually went to classes with eager anticipation.

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

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

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

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

17. Leave blank and click next.

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

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

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

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

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

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

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

Demographics – Class:

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

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

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

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

29. What level was this course?

0) 100-level 1) 200-level 2) 300-level 3) 400-level
4) other

(Disliked/Liked) Class

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

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

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

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

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

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

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

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

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

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

40. Leave blank and click next.

41. I usually went to classes with eager anticipation.

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

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

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

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

46. Leave blank and click next.

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

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

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

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

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

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

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

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

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

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

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

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

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

60. Your age is:

0) 17 1) 18 2) 19 3) 20 4) 21 5) 22
6) 23
7) 24 8) 25 9) 26 or older

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

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

63. Your program of study is part of which faculty?
0) Arts 1) Social Science 2) Engineering 3) Human Kinetics
4) Nursing 5) Business 6) Business 7) Education

Appendix C.2

Negatively Worded Items:

The following questions will be presented with either *five* or *eleven* scale points (presented using radio buttons), with either *unidirectional* or *bidirectional* response options and the type of class being evaluated will be presented in counterbalanced order (i.e., *Liked Class* will appear in half of the cases and second for the other half).

Questions

(Liked/Disliked) Class

1. With respect to your progress in the course, the instructor was not concerned and was not actively helpful.

2. In terms of what I gained (learned) from the course, the grade that I obtained was not an excellent reflection.

3. By not raising challenging questions or problems for discussion, the instructor did not stimulate students to think for themselves in nearly every class.

4. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were not judged to be conducive to learning.

5. Based on your experience, the instructor's attitude toward students as individuals was not respectful.

6. As reflected in the classroom and in the presentation of course material, the instructor was not very enthusiastic.

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

8. As a result of this course, my knowledge level in this area has not greatly increased.

9. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor did not freely permitted comments.

10. The number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were not sufficient to reflect achievement.

11. Leave blank and click next.

12. I did not eagerly anticipate going to class.

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

14. As a result of this course, my interest in pursuing additional knowledge in this area was not stimulated.
15. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was not very useful or worthwhile.

16. The examination questions, or other evaluative methods used by the instructor, did not seem to be very clear or fair.

17. Leave blank and click next.

18. The method of assigning grades was not clearly understood and consistent.

19. The instructor's ability was not evident, as the concepts of this course were not conveyed in a clear, meaningful, and appropriate way.

20. The instructor's classroom presentation was not well prepared at all times.

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

22. From my own experience, the instructor did not come across as a person as well as a teacher very well.

23. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was not appropriate.

24. The general objectives of the course were not clearly understood.

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

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

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

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

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

(Disliked/Liked) Class

30. With respect to your progress in the course, the instructor was not concerned and was not actively helpful.

31. In terms of what I gained (learned) from the course, the grade that I obtained was not an excellent reflection.

32. By not raising challenging questions or problems for discussion, the instructor did not stimulate students to think for themselves in nearly every class.

33. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were not judged to be conducive to learning.

34. Based on your experience, the instructor's attitude toward students as individuals was not respectful.

35. As reflected in the classroom and in the presentation of course material, the instructor was not very enthusiastic.

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

37. As a result of this course, my knowledge level in this area has not greatly increased.

38. With respect to students' freedom to express opinions and ask questions in the classroom, the instructor did not freely permitted comments.

39. The number and type of evaluations (i.e., exams, assignments, papers, etc) used in determining the final grade were not sufficient to reflect achievement.

40. Leave blank and click next.

41. I did not eagerly anticipate going to class.

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

43. As a result of this course, my interest in pursuing additional knowledge in this area was not stimulated.

44. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was not very useful or worthwhile.

45. The examination questions, or other evaluative methods used by the instructor, did not seem to be very clear or fair.

46. Leave blank and click next.

47. The method of assigning grades was not clearly understood and consistent.

48. The instructor's ability was not evident, as the concepts of this course were not conveyed in a clear, meaningful, and appropriate way.

49. The instructor's classroom presentation was not well prepared at all times.

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

51. From my own experience, the instructor did not come across as a person as well as a teacher very well.

52. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was not appropriate.

53. The general objectives of the course were not clearly understood.

Demographics – Class:

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

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

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

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

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

Demographics – Student:

59. Your gender is:
0 Female 1) Male

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

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

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

63. Your program of study is part of which faculty?
0) Arts 1) Social Science 2) Engineering 3) Human Kinetics
4) Nursing 5) Business 6) Business 7) Education

Appendix C.3

Mixed Worded Items:

The following questions will be presented with either *five* or *eleven* scale points (presented using radio buttons), with either *unidirectional* or *bidirectional* response options and the type of class being evaluated will be presented in counterbalanced order (i.e., *Liked Class* will appear in half of the cases and second for the other half).

Questions

(Liked/Disliked) Class

1. With respect to your progress in the course, the instructor was not concerned and was not actively helpful.

2. In terms of what I gained (learned) from the course, the grade that I obtained was not an excellent reflection.

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

4. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were not judged to be conducive to learning.

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

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

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

8. As a result of this course, my knowledge level in this area has not greatly increased.

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

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

11. Leave blank and click next.

12. I usually went to classes with eager anticipation.

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

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

15. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was not very useful or worthwhile.

16. The examination questions, or other evaluative methods used by the instructor, did not seem to be very clear or fair.

17. Leave blank and click next.

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

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

20. The instructor's classroom presentation was not well prepared at all times.

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

22. From my own experience, the instructor did not come across as a person as well as a teacher very well.

23. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was not appropriate.

24. The general objectives of the course were not clearly understood.

Demographics – Class:

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

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

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

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

29. What level was this course?

0) 100-level 1) 200-level 2) 300-level 3) 400-level
4) other

(Disliked/Liked) Class

30. With respect to your progress in the course, the instructor was not concerned and was not actively helpful.

31. In terms of what I gained (learned) from the course, the grade that I obtained was not an excellent reflection.

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

33. In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were not judged to be conducive to learning.

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

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

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

37. As a result of this course, my knowledge level in this area has not greatly increased.

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

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

40. Leave blank and click next.

41. I usually went to classes with eager anticipation.

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

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

44. In one way or another (whether in relationship to my major, other courses, or just life in general) this course was not very useful or worthwhile.

45. The examination questions, or other evaluative methods used by the instructor, did not seem to be very clear or fair.

46. Leave blank and click next.

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

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

49. The instructor's classroom presentation was not well prepared at all times.

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

51. From my own experience, the instructor did not come across as a person as well as a teacher very well.

52. Considering the nature of the course in terms of subject and class size, the method of presentation of the material (i.e., lecture, discussion groups, etc.) was not appropriate.

53. The general objectives of the course were not clearly understood.

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

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

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

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

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

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

60. Your age is:

0) 17 1) 18 2) 19 3) 20 4) 21 5) 22 6) 23
7) 24 8) 25 9) 26 or older

61. Your year in university is:

0) First B) Second 1) Third 2) Fourth 3) Other

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

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

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

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

Appendix D

Factor Loadings:

Rapport with Students

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

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

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

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

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

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

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

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

Course Value

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

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

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

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

15. & 44 In one way or another (whether in relationship to my major, other courses, or just life in general) this course was very useful and worthwhile.

Course Organization and Design

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

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

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

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

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

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

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

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

Fairness of Grading

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

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

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

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

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

VITA AUCTORIS

Marc Frey was born in 1985 in Windsor, Ontario. He graduated from St. Anne's High School in 2003. From there he went on to the University of Windsor where he obtained a B.A. in Psychology in 2007. He is currently a candidate for the Master's degree in Psychology at the University of Windsor and hopes to graduate in Fall 2009.