University of Alberta

Effects of Negatively Worded Items and the Provision of a Warning about the Inclusion of Negatively Worded Items in an Attitude Questionnaire

by

Alexander Riedel

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation and Cognition

Educational Psychology

Fall, 2012
Edmonton, Alberta

Abstract

It has been widely demonstrated that the inclusion of negatively worded items in attitude questionnaires can have adverse effects on the responses (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).The present study examined the effect of a varying number of negatively worded items and the provision of a warning about the inclusion of negatively worded items on the responses of 333 students in a 12 item attitude questionnaire. Five questionnaire forms were used. The results revealed that the varying number of negatively worded items did not have an effect on the responses. The results furthermore indicated that no wording effects associated with negatively worded items were found except for one negatively worded item that exhibited high intensity in its negative form. Another finding was that there is no need for the provision of a warning. The implications of the results for practice and directions for future research are discussed.

.

Acknowledgements

I want to use this opportunity to thank all the great people who have contributed to this work. Above all, I want to thank my supervisor Dr. Todd Rogers for the endless hours that we spent talking about my thesis work, his valuable advice, his enthusiasm, and his commitment to his students. Thank you! I would also like to thank Dr. Cheryl Poth for her great guidance during the last two years. She has challenged me in many ways, and made me a stronger thinker.

I also want to express my thanks to my fellow CRAME graduate students. I would like to thank Qi Guo for his friendship, the enormous feedback on my thesis, and for the great discussions that we had about…. anything. I would also like to thank other CRAME students, Simon Turner, Oksana Babenko, Amin Mousavi, for their valuable comments on my work.

Next, a very special thank you goes out to my beloved climbing friends, who have been like a family to me in Edmonton. Thank you, Anthony Zeberoff, Sheelah Griffith, Emily Moss, Zoe Avner, Devin Goodsman, Niall Fink, Ian Scriver, and Wallace Currie. I also want to express my gratitude to Chantell Rendell, who has not only been a great support during my final phase of my thesis, but who also taught me many valuable things about life in general.

Finally, I want to thank my family. I want to thank my mum, my dad, my brother, and my grandparents for always being encouraging me to go new ways, learn new things, and see the world.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background to the Study

Researchers agree that method bias (i.e., variance that is attributable to the measurement method and not to the construct to be measured) is a serious threat to valid interpretation of scores obtained from of a measuring instrument (e.g., attitude scales, personality scales, self-report surveys, achievement test) (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Method bias may originate from the specific method of data collection, the respondent, or the questionnaire (Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 1991). The recognition of method bias is nothing new, and discussions surrounding this topic go back as far as 60 years (Cronbach, 1950). Throughout the last six decades several potential sources for measurement bias have been identified, such as common rater effects, item characteristic effects, item context effects, and response sets such as acquiescence, response preference, and social desirability (Podsakoff et al., 2003). In response to these sources, several approaches for mitigating the effects of method bias have been proposed.

One of the approaches is the inclusion of negatively worded items in questionnaires designed to measure attitudes, interests, and other personality factors. It has been recommended to guard against acquiescence (i.e., tendency to agree or disagree regardless of the content) when Likert type response formats are used (Jackson, 1967; Nunnally, 1978). To try to overcome acquiescence, half of the items are negatively worded to produce what is called a balanced scale. The rationale for this recommendation stems from the fact that the influence of

acquiescent response behavior will be cancelled out by the equal number of items worded in the opposite direction (Ray, 1979).

A review of several resources on the construction of questionnaires revealed that although similarities in the procedure involving the inclusion of negatively worded items were present, differences also existed. Whereas some authors recommended half of the items to be negatively worded (Anastasi & Urbina, 1997; DeVellis, 2012; Nunnally & Bernstein, 1994), others did not mention the inclusion of negatively worded items at all (Dillman, Smyth & Melanie Christian, 2009; Fowler, 2002; Peterson, 2000). Rubin and Babbie (2011) suggested the inclusion of negatively worded items, but did not specify the ratio of positively and negatively worded items. In his *Guidelines for Superior Survey Design*, Morrel-Samuels (2002) suggested one third of the items should be negatively worded.

One possible reason for the inconsistent recommendation on the use of and the number of negatively worded items might be another emerging rationale for the inclusion of negatively worded items. Whereas in its origins the inclusion of negatively worded items was recommended to mitigate the effects of acquiescence, recent recommendations on including negatively worded items suggest that this practice would prevent careless responding (Chen, Rendina-Gobioff, & Dedrick, 2010; Podsakoff et al., 2003). In their highly acknowledged paper *Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies*, Podsakoff et al. (2003) referred to negatively worded items as "cognitive 'speed bumps'", with the function "to

prevent careless responding" (p. 884). In fact, several researchers who recently examined the performance of negatively worded items stated that the reason for the inclusion of negatively worded items is prevention of careless responding (Chen et al., 2010; Roszkowski & Soven, 2010; Weems, Onwuegbuzie, Schreiber, & Eggers, 2003). If careless responding is the primary reason for inclusion of negatively worded items, then balanced scales may not be necessary. Consequently the question arises concerning the optimal ratio of negatively and positively worded items.

So far two lines of reasoning for the inclusion of negatively worded items were presented: 1) to prevent acquiescent responding and 2) to prevent careless responding. Regardless of the intended purpose of the inclusion of negatively worded items various researchers questioned this technique by pointing out the unanticipated problems associated with negatively worded items (Barnette, 2000; DiStefano & Motl, 2006; Idaszak & Drasgow, 1987; Schriesheim & Hill, 1981). Among the problems highlighted by these researchers, one is related to significantly different means and standard deviations of positively and negatively worded items (Carmines & Zeller, 1979; Steward & Frye, 2004). What is interesting is that not all studies have found this problem associated with negatively worded items (Barnette, 2000; Bergstrom & Lunz, 1998). Barnette (2000), for example, administered a questionnaire that had four forms with the direction of the rating scale crossed with the presence/absence of negatively worded items. The results revealed no significant interaction or main effects.

A second problem was revealed by researchers employing factor analytic techniques. DiStefano and Motl (2006), Idaszak and Drasgow, (1987) and Marsh (1996) found spurious factors that were defined by negatively worded items (DiStefano & Motl, 2006; Idaszak & Drasgow, 1987; H. Marsh, 1996).

Several reasons have been suggested for the observed problems associated with the inclusion of negatively worded items. Marsh (1986) proposed that age could play a role; others suggested that negatively worded items might not be measuring the same construct as their positively worded counterparts (Pilotte & Gable, 1990).Yet others argued that inattentiveness to negatively worded items might be a reason (Schmitt & Stults, 1985; Woods, 2006). It is important to mention that inattentiveness to negatively worded items is referred to as careless responding by some authors (e.g., Barnette, 2000; Schmitt & Stults, 1985). Within this study inattentiveness to negatively worded items will be used to refer to a special case of careless responding when the respondents fail to recognize that the polarity of items can be on opposite directions.

Interestingly enough, the explanation that inattentiveness to negatively worded items might be a reason for adverse effects associated with the negative wording seems to be counterproductive to the recent recommendation for inclusion of negatively worded items as a means to prevent careless responding. On one hand one is advised to include negatively worded item which should result in less careless respondents, on the other hand the researchers suggest that negatively worded items cause problems due to inattentiveness to negatively worded items. To solve this dilemma researchers have recommended providing a

sentence in the instructions that would warn the respondents about the inclusion of negatively worded items (Roszkowski & Soven, 2010; Schmitt & Stults, 1985). However no study was found in which the effect of such a warning had been empirically investigated.

Purpose of the Study

Two major purposes were addressed in this study in an attempt to address the uncertainty regarding the number of negative items to include and the need for an instruction alerting the respondents that negative items were present. The first purpose was to examine the effect of the inclusion of a varying number of negatively worded items on the responses of undergraduate students in a short questionnaire with a fixed length that measured students' attitudes toward assessment. The second purpose of the study was to investigate the effect of including a warning about the inclusion of negatively worded items on the responses of the undergraduate students.

Definition of Terms

Assessment: Assessment is broadly defined in the *Classroom Assessment Standards* as the process of collecting and interpreting information that can be used to inform students, and, when applicable, parents/guardians, about students' progress in attaining the knowledge, skills, attitudes, and behaviors to be learned or acquired in school. Adapted from the *Principles for Fair Student Assessment Practices for Education in Canada* (1993, p. 1).

Acquiescence: Acquiescence is a respondent's tendency to agree to say "True", "Yes", and "Agree" when in doubt (Cronbach, 1950).

Attitude: Attitude is comprised of five characteristics: emotion, consistency, target, direction and intensity (Anderson, 1981). An attitude can be considered as consistent moderately intense emotion that can vary from negative to positive predispositions toward a particular object.

Balanced Scales: Balanced scales contain an equal number of both positively worded and negatively worded items (Nunnally, 1978).

Inattentive responding to negatively worded items: lack of attention to the negative polarity of negatively worded items.

Likert rating scale: Likert rating scales are bipolar scales and typically consist of five response options measuring the intensity and direction of respondent's attitudes towards statements. Moreover, Likert rating scales vary in the way that they can include a "forced choice" by not displaying the middle option, they can be positively or negatively packed, or have all rating points labeled or only the endpoints.

Method bias: Variance is explained by measurement method rather than by the construct that is being measured (Bagozzi & Yi, 1990).

Negatively worded items: Negatively worded items are phrased in the opposite semantic direction from positively worded items (Colston, 1999). Researchers are using alternative names for negatively worded items (e.g., reversed items, negatively keyed items, and items with negative mode.).Throughout this study the term negatively worded items will be used.

Response set: response set is the respondent's tendency to respond to test content in a particular way, which would be different when the same test content is presented in a different format (Cronbach, 1946).

Organization of Thesis

The thesis is organized in five chapters. Chapter one situated the study within the field of measurement bias and provided a rationale for the study. Chapter 2 contains a structured literature review beginning with a historical overview on the practice of the inclusion of negatively worded items for addressing measurement bias, followed first by a discussion addressing the critiques associated with negatively worded items and then suggested reasons for the problems identified in the critiques. The methods used to address the research purpose, including the instrument, participants, and statistical analyses, are described in Chapter 3. The results are reported in Chapter 4, which is organized in two sections: descriptive statistics followed by the results of the inferential statistics that were conducted. Chapter 5 is organized in seven sections: summary of purpose and procedures, summary of results, discussion of results, limitations of the study, and conclusions of the study formulated in light of the limitations, implications for practice, and recommendations for future research.

Chapter 2: Literature Review

The present study is concerned with potential method bias that may be introduced by including negatively worded items in attitude questionnaires. The literature relevant to this concern is presented in the current chapter in seven sections. First, a historical account related to the introduction of negatively worded items is provided. Second, the types of negatively worded items are introduced. Third, literature related to the balance of negative and positive items in a scale is discussed and an alternative recent recommendation for the inclusion of negatively worded items is presented. In the fourth section the cognitive processes behind careless responding are presented. In the fifth section the relationship among acquiescence, careless responding, and inattentiveness to negatively worded items is discussed. Research findings that point out the potential of introducing undesirable method bias when negatively worded items are included in a questionnaire are discussed in the sixth section. Potential reasons for the method bias introduced by negatively worded items are presented in the seventh section. The chapter concludes with a short summary of the research and a statement of the research questions.

Historical Account on the Inclusion of Negatively Worded Items

The research related to method bias responded to the evidence presented in Cronbach's widely acknowledged article *Response Sets and Test Validity* (Cronbach, 1946). In his article, Cronbach assembled evidence from various early studies that response sets are present in tests of ability, personality, attitude, and interest. He defined response sets as "any tendency causing a person consistently

to give different responses to test items than he would when the same content is presented in a different form" (Cronbach, 1946, p. 476). In a later paper, Cronbach (1950) wrote that acquiescence is the tendency to say "True", "Yes", and "Agree" when in doubt and evasiveness, which is tendency to choose neutral response option, were the two most widely found response sets (p. 3). He stated that response sets can seriously affect the validity of a test score interpretation, and argued that "recognition and control of such irrelevant factors are precisely the improvements needed to raise mental measurement from its present imperfect level" (Cronbach, 1950). Cronbach (1950) suggested several approaches about how to control response sets: 1) designing items that prevent a response set, 2) altering directions to reduce a response set, and 3) correcting for a response set. For example, when directions are altered in a way that students know that about half of the items in a test are false, then the students will reduce their tendency to select true when in doubt.

Cronbach's proposed approaches for controlling response sets set the stage for more exploration of techniques for controlling response sets. The construction of balanced scales, which is one of the common strategies for controlling acquiescence, was introduced in the 1950s (e.g., Edwards, 1957), and has been recommended to the present day by several researchers (Anastasi & Urbina, 1997; DeVellis, 2012; Jackson, 1967; Nunnally & Bernstein, 1994). According to these authors, the number of positively worded items within a balanced scale should equal the number of negatively worded items. Jackson (1967), for example, argued that: "if scales are developed with half of the items keyed *true* and half

keyed *false* (or half *agree* and half *disagree*), the massive cumulative effects of acquiescence may be avoided" (p. 93). Three important assumptions underlying the use of balanced scales were identified by Schriesheim and Eisenbach (1995). First, acquiescence is a serious threat to the validity of score interpretation. Second, the negatively worded and positively worded items are bipolar statements within the same construct. Third, negatively worded items can be used without major adverse side-effects on the psychometric properties of the instrument.

In the light of these assumptions, Schriesheim and Eisenbach (1995) took a critical standpoint towards the inclusion of negatively worded items as a means to guard against acquiescence in questionnaires. First, they described the technique of inclusion of negatively worded items as a "conservative practice" (p. 1178). They referred to Nunnally's (1978) conclusion that "the overwhelming weight of evidence now points to the fact that the agreement tendency is of very little importance either as a measure of personality or a course of systematic validity in measures of personality and sentiments" (p. 669). Consequently, the first assumption cannot be made. Second and most importantly, Schriesheim and Eisenbach pointed to the large body of literature that showed that the inclusion of negatively worded items resulted in unanticipated side effects (e.g., method artifacts)[1], which violated the third assumption.

Although Schriesheim and Eisenbach claimed that acquiescence is not a serious threat to the validity of questionnaires, the results of several studies contradict their claim and reveal that acquiescence does exist (Narayan &

[1] An overview of the studies that examined the effect of the inclusion of negatively worded items is provided in the sixth section.

Krosnick, 1996; Ray, 1983; Schuman & Presser, 1996)**.** In their experimental study Schuman and Presser (1996) showed that response format had an effect on agreement. In the General Social Survey (GSS) two forms of statement were presented to the 1,417 respondents: Form 1) *Do you agree or disagree with this statement: Most men are better suited emotionally for politics than are most women.* Form 2) *Would you say that most men are better suited emotionally than are most women, that men and women are equally suited, or that women are better suited than men in this area?* The results showed that whereas in the first form 47.0% agreed that men are better suited, in the second form 33.1% agreed that men are better suited and 66.9% indicated that both are equally suited. Narayan and Krosnick (1996) conducted a meta-analysis of the 13 experiments that assessed response effects related to acquiescence. They found that response acquiescence was significantly higher for the respondents with lower education than for the respondents with higher education.

Types of Negatively Worded Items – Nomenclature

The second of the three assumptions stated above has not been discussed to this point. The second assumption states that negatively worded and positively worded items are bipolar statements within the same construct. But, negatively worded items can be created in more than one way. The question then arises whether the different types equally qualify as bipolar counterparts of the positively worded items.

Positively and negatively worded items can be classified into four categories, which have different names depending on the author (e.g., Colston,

1999; Schriesheim et al., 1991). Schriesheim et al.'s, (1991) four categories include: 1) regular items, 2) negated regular items, 3) polar opposite items, and 4) negated polar opposite items. When applied to the statements designed to determine a respondent's attitude towards ice-cream, the four categories will result in, respectively: 1) *I like ice cream*, 2) *I don't like ice-cream*, 3) *I dislike ice-cream*, and 4) *I don't dislike ice-cream*. Items written in the negated negative mode (negated polar opposite) style incorporate a double negative and can be confusing to respondents, and therefore the use of this form is discouraged (e.g., Krosnick & Presser, 2010; Peterson, 2000)**.**

As mentioned earlier, the two methods for creating negatively worded items lead to the question about which of the two types should be used. Rorer (1965), for example, noted that it can be quite difficult to construct polar opposite items with meanings that are the opposite to the meanings of the corresponding regular items. Rorer referred to an earlier study conducted by Jackson and Messick (1957). Jackson and Messick deliberately wrote extreme polar opposites using statements included the California F Scale (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950) to illustrate that it was possible to write polar opposites such that the negative form and the corresponding regular items can both be rejected:

> "Obedience and respect for authority are the most important virtues children should learn." (California F-Scale, Adorno et al., 1950)

> "A love of freedom and complete independence are the most important virtues children should learn" (Jackson & Messick, 1957).

In this example, love of freedom and complete independence are polar opposites of obedience and respect for authority.

Polar opposites are more difficult to construct than negated regularly items. One simply needs to include a negation such as *no* or *not* to create the negative form. While simpler, the use of negations of the original statements is not advised (Jackson, 1967). Jackson stated that negated regular items:

> fail to qualify as truly unique items when administered in the context of the original. For example, it is doubtful that the items, 'I would like to hunt lions in Africa' and 'I would not like to hunt lions in Africa', represent two distinct harm-avoidance items. (p. 94)

In an attempt to examine the equality of the four different types of items, Schriesheim and Eisenbach (1995) examined the internal consistency (Cronbach's alpha) of the revised 95-item Leader Behavior Description Questionnaire. The authors randomly assigned 124 business administration students to one of four forms, each of which contained five items that were either regular items (R), polar opposite items (P), negated polar opposites items (NP), or negated regular items (NR). The highest coefficients values of alpha were found for R (0.89), followed by NR (0.84), P (0.80), and NP (0.70). Schriesheim and Eisenbach concluded that whereas the internal consistencies for NR and P forms were not substantially

different from the internal consistency for the R form, the internal consistency of the NP form was substantially lower. It must be acknowledged though, that in light of the low number of respondents (n=31) for each form, the differences among the four internal consistency values are likely to be statistically nonsignificant (due to the lack of information this speculation could not be verified). Despite this limitation, the study provided some support for the equality of the two types of negatively worded items.

Balanced versus Unbalanced Questionnaires

In addition to debating whether negatively worded items should or should not be included in a questionnaire, the question of ratio of positively worded items to negatively worded items remains. Whereas several authors (e.g., Anastasi & Urbina, 1997; Jackson, 1967; Nunnally, 1978; Ray, 1990) agree that the number of negatively and positively worded items should be equal, others (e.g., Morrel-Samuels, 2002a; Rubin & Babbie, 2011) offer other suggestions. Krosnick and Presser (2010) recommend that questions with single or double negations should be avoided (p. 264). Other suggestions include other combinations or do not specify the ratio of negatively to positively worded items. For example, Morell-Samuels (2002) recommended changing the wording to negative wording in about one-third of the items without providing any evidence for this recommendation. In their chapter on constructing measurement instruments, Rubin and Babbie (2011) talked about the use of negatively worded items, but they did not make any recommendations on the relative number of negatively and positively worded items. It is therefore not surprising that many questionnaires will differ in the

number of positively and negatively worded items or will not include negatively worded items at all.

For example, Gomleksiz (2004) used 12 reversed items in a 24 item questionnaire exploring teachers' attitudes towards the use of technology; Brown (2006) included seven negatively worded items in the set of 25 items in the questionnaire about teachers' conceptions of assessment COA-III abridged; and Kim (2011) included four negatively worded items in the set of ten items in the Attitude Towards Science Test. Within their study examining the effect of negatively worded items, Roszkowski and Soven (2010) commented that given the mixture of different ratios of negatively to positively worded items, it seemed that the choice of the ratio is arbitrary. They furthermore stated that "it's been our personal experience that the sponsors of surveys are frequently reluctant to use negative stems, so that in scales with mixes of the two modes, the tendency is to have more positive than negative items" (p. 120).

One possible reason for the inconsistent recommendations for the ratio of the positively to negatively worded items might be grounded in the current confusion about the purpose of using negatively worded items. Whereas originally the use of negatively worded items was recommended as a remedy against acquiescence, a recent recommendation states that the purpose is to control careless responding. The recommendation to include negatively worded items in questionnaires to control careless responding can be found in a widely recognized paper on method biases by Podsakoff et al. (2003) who suggested that reversed-coded items functioned as "cognitive 'speed bumps'" that require the

respondents to engage in more controlled, as opposed to automatic, cognitive processing" (p. 884). In fact, recently published articles on the effects of negatively worded items see the inclusion of negatively worded items as a means to prevent careless responding (e.g., Chen et al., 2010; Roszkowski & Soven, 2010; Weems et al., 2003). Weems et al. (2003), for example, wrote that "the fundamental reason for mixing item direction is to discourage non-attending behaviours" (p. 589).

At present a question that consequently arises is: if acquiescence is not the main reason for including negatively worded items, and instead prevention of careless responding is the main reason, then what should be the ratio of negatively to positively worded items? Unfortunately, neither Podsakoff et al. (2003) nor other researchers (e.g., Weems et al., 2003) made any suggestions regarding this issue and the question remains about the ratio of negatively to positively worded items. One way to approach this issue is to understand the processes behind careless responding.

Cognitive Processes Involved in Questionnaire Responding

Numerous survey methodologists have argued that errors in surveys are largely caused by difficulties in the cognitive processes that are involved when respondents are answering the questions (Tourangeau & Bradburn, 2010). The introduction of cognitive processing as a way to think about errors in surveys was initiated in a series of conferences that brought together survey methodologists and cognitive psychologists to explore potential relations between cognitive processes and errors present in surveys (Tourangeau & Bradburn, 2010). As a

result, the researchers came up with a heuristic model called the Cognitive Aspects of Survey Methodology (CASM) that describes four steps in formulating responses to surveys (Tourangeau & Bradburn, 2010). When presented with questions, the respondents' optimal cognitive processes will include four steps, where every single step has its specific pitfalls. The first step includes the interpretation of the intent of the question (question comprehension); the next step includes retrieval of information related to the question from memory (information retrieval); the third step includes forming a judgment based on the retrieved information (judging and estimation); and the final step includes the translation of the judgment into a response (reporting). Each of these four steps requires the respondent to do some cognitive work. Depending on the level of the commitment to do this work, Krosnick (1991) distinguishes between respondents who can be said to be *optimizing* and respondents who can be said to be *satisficing*. Optimizing respondents are motivated to go through the four cognitive steps in order to respond to questions in a thorough and unbiased manner. It is often the case, though, that respondents are not motivated to respond to a questionnaire, especially in circumstances where participation in a survey becomes obligatory with little or no reward. In these situations the respondents are likely to adapt certain strategies that can be termed *satisficing*. These strategies are characterized through subtle or drastic shortcuts rather than elaborate cognitive processing when formulating a response. One of those shortcuts could be inattentiveness to negative wording of the items (i.e., not following the four steps when responding to negatively worded items). Indeed,

several authors have suggested that one of the possible reasons for unanticipated side-effects of including of negatively worded items (e.g., method artifacts) could be related to inattentive responding to negatively worded items. For example, Roszkowski and Soven (2010) examined the effect of the inclusion of negatively worded items on course evaluation questionnaires and found an unfavourable effect on the internal consistency of the questionnaires. They concluded that the effect came about possibly because:

> after encountering a few stems that are positive (favourable) in nature, an expectation is created that the next item will positive as well. Once the mindset is established that the stems are positive and that agreement with the statement denotes a favourable evaluation and disagreement an unfavourable evaluation, it is easy to gloss over the negative content of a statement when it occurs suddenly without a warning after a series of positive statements. (p. 129).

That is, the respondents are satisficing in their response strategies.

Roszkowski and Soven (2010) looked at the number of negatively worded items and the placement of the items within two questionnaires. They included two negatively worded items and placed them together in the middle of a 13 item questionnaire and an 18 item questionnaire related to undergraduate course evaluation. Roszkowski and Soven concluded that inattentiveness to negatively worded items had an unfavourable effect. Further, the authors referred to the results of the simulation study done by Schmitt and Stults (1985) that indicated

that when at least 10% of the respondents do not recognize negatively worded items, factor analysis yielded a clearly definable factor that was solely comprised of negatively worded items.

While Roszkowski's and Soven's (2010) explanation seems plausible, dispersion of the two negatively worded items within the questionnaire would probably result in different findings. It seems logical that if negatively worded items are intended to prevent satisficing response behavior, then at least one negatively worded item should be placed within two or three items at the beginning of the questionnaire to prevent a potential buildup of a careless responding and inattentiveness to negatively worded items. To summarize, the reasons for the adverse effects of negatively worded items are potentially due to satisficing behavior related to inattentiveness to negative wording of some items. This creates a "dilemma" for researchers and practitioners faced with the decision of including or not including negatively worded items. One possible solution to this dilemma was mentioned earlier, namely the dispersion of negatively worded items throughout the questionnaire.

Another simple and plausible suggestion is to make it clear to respondents that some of the questions are worded in a negative mode by providing a warning (Roszkowski & Soven, 2010; Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1995). Schmitt and Stults (1985), for example, stated that one way to reduce potential problems with negatively worded items is to "include a warning to potential respondents that some questions will be negatively keyed and that they should attend to all items" (p. 371). Similarly, Roszkowski and Soven (2010)

stated that if negatively worded items are to be included the respondents need to be alerted, especially when the mix is unbalanced towards few negatively worded items. However no study was identified in which the effect of such a warning has been empirically investigated.

Relationship among Acquiescence, Careless Responding, and Inattentiveness to Negatively Worded Items

Historically the inclusion of negatively worded items was recommended to prevent bias due to acquiescence (Cronbach, 1950; Edwards, 1957). Recently the rationale for the inclusion of negatively worded items is to prevent careless responding (Podsakoff et al., 2003). At the same time, several authors (e.g., Roszkowski & Soven, 2010) point out that the inclusion of negatively worded items can lead to adverse side effects due to inattentiveness to negatively worded items.

A question that needs to be addressed is what is the relation among acquiescence, careless responding, and inattentiveness to negatively worded items? As was mentioned earlier, acquiescence is the respondents' tendency to say "True" or "Yes" rather "Wrong" or "No" regardless of the content (Cronbach, 1950). A review of the literature revealed that no standard definition exists in regards to what careless responding is. Meade and Craig (2012) report that careless responding can result in different data patterns. They write "for example some persons may randomly choose from all response options on a scale. Others may employ a nonrandom pattern, such as giving many consecutive items a response of "4," or repeating a pattern of "1, 2, 3, 4, 5, ..." (p. 2). Within the

context of the adverse effect related to the negative wording of the items Schmitt and Stults (1985) defined a careless respondent as someone who "is simply reading a few of the items in a measuring instrument, inferring what it is the items are asking of the respondent, and then reporting in a like manner to the remainder of the items in a questionnaire" (p. 367). It is worthwhile mentioning that Schmitt and Stults stressed the importance that this kind of careless responding is not responding randomly. In order to prevent confusion with the usage of terms, within this study careless responding due to the lack of attention to the negative polarity of negatively worded items is referred to as inattentiveness to negatively worded items.

Inattentiveness to negatively worded items can be seen as a special case of careless responding. Theoretically both inattentiveness to negatively worded items and acquiescence can lead to adverse side effects associated with the inclusion of negatively worded items. It is crucial to note that there is a substantial difference between the two concepts. Whereas an acquiescent person agrees or disagrees because of his/her tendency to agree or disagree regardless of the content when in doubt (Jackson, 1967), a respondent who fails to detect the different polarities within a questionnaire tends to answer in certain pattern (e.g., disagrees across all items). This pattern is determined by the polarity of the questionnaires items placed in the beginning of the questionnaire. Even though the two concepts are quite distinct, a novice to the area might easily confuse the two. For example, Krosnick (1991) wrote that one of the reasons for acquiescence is the respondents' inclination to satisfice rather than optimize in a questionnaire.

At the same time, inattentive responding to negatively worded items can be put under umbrella of satisficing responding, which seems to indicate that both acquiescence and inattentive responding to negatively worded items are related concepts. It is important to note though, that satisficing can take on different forms and is not a description of one single phenomenon. Selection of a midpoint in a scale, for example, would also fall under the umbrella of satisficing as well, because choosing the midpoint would allow the respondent to explain these answers with little difficulty ("keep things as they are") when pressed to do so (Krosnick, Judd, & Wittenbrink, 2005).

Inattentiveness to negatively worded items might be one reason for aberrant behavior of negatively worded items and acquiescence might be a second reason. Some researchers suggested other reasons, which will be discussed later following a review of studies that focused on effects of inclusion of negatively worded items.

Psychometric Quality of Negatively Worded Items

Numerous studies have examined the effects of negatively worded items on the psychometric properties of the scale in which the items are included. Some studies employed descriptive and inferential statistics (e.g., ANOVA) or internal consistency measures (i.e., Cronbach's alpha) to show the effects (Barnette, 2000; Guyatt et al., 1999; Roszkowski & Soven, 2010; Schriesheim & Hill, 1981; Steward & Frye, 2004). Other studies used factor analysis to examine the factor structure of questionnaires that included negatively worded items (DiStefano & Motl, 2006; Idaszak & Drasgow, 1987; Magazine, Williams, & Williams, 1996;

Marsh, 1996; Pilotte & Gable, 1990; Roszkowski & Soven, 2010; Schmitt & Stults, 1985).

Schriesheim and Hill (1981) administered the 10-item *Initiating Structure* questionnaire (responses to written managerial description) with three forms (all items positive, mixed, all negative) to 150 business undergraduate students who were randomly assigned to one of the three forms. Their results indicated that the total scores across the different forms differed practically and significantly, with mixed and all negative forms displaying higher total scores than the all positive form. In a more recent study, Steward and Frye (2004) investigated the effect of negatively worded items on the responses of 1,571 first year medical students to a medical education attitude survey. The survey included 55 items with seven unidimensional scales, which all included negatively worded items (number of negatively stated items was not specified). Their results indicate that across all seven scales the total mean score for the positively worded items was greater than the total mean score for the negatively worded items.

Several factor analytic studies have revealed that spurious factors can emerge that are solely defined by the negatively worded items. Carmines and Zeller (1979) examined the latent structure of the 10-items Rosenberg Self Esteem (RSE) scale (Rosenberg, 1965). Exploratory factor analysis yielded a two factor structure with negatively worded items loading on one factor and positively worded items on the other factor. Using a sample of 1,672 professionals, managers and workers, Idaszak and Drasgow (1987) investigated the factor structure of a 15-item Job Diagnostic Survey (JDS) and found that whereas five of

the six factors corresponded to the expected dimensions, the sixth factor was artifact factor comprised of solely the negatively worded items (n=5). In a subsequent study, Idaszak and Draskow (1987) reversed the negatively worded items and administered the JDS to 134 employees of a printing plant; the expected five factor solution was obtained with no artifact factor. Cordery and Sevastos (1993) administered the revised JDS (n=20 items) to 3,044 public sector workers in various jobs and varying educational levels. They confirmed earlier studies that the JDS comprised of only positively worded items fit the five factor structure, but when the negatively worded items (n=5, same as in the original JDS) were included, the five factor model showed poor fit, which was the case for both low and high educational subsamples. An important finding of the study was that sample characteristics, such as varying educational levels, were not responsible for the commonly observed dimensionality problems with the JDS.

Pilotte and Gable (1990) employed confirmatory factor analysis to evaluate the internal structure of three forms of a 9 item *Computer Anxiety Scale*. Form A consisted of all positively worded items, form B had all items worded negatively, and form C had four negatively worded items. The forms were administered to 270 students in Grades 9 to12; each student responded to one of the forms. The results showed: a) a single factor for Form A (all items were indicative of computer anxiety); b) a single factor for Form B (all items were indicative of a lack of computer anxiety); and c) two correlated ($r = 0.24$) factors for form C, with the first factor displaying loading of positively worded items and the second factor displaying loading of negatively worded items.

A rather comprehensive study was conducted by DiStefano and Motl (2006) who used data collected from 757 students to evaluate the separation of content and wording effects in two self-report scales (Rosenberg Self-Esteem (RSE) scale (Rosenberg, 1989) and Social Physique Anxiety Scale (SPAS; Heart, Leary, & Rejeski, 1989) by utilizing different structural equation models (SEM). The RSE is a 10 item questionnaire with five negatively worded items and the SPAS is a 12 item scale with five negatively worded items. Results indicated that wording effects related to negatively worded created a distinct latent variable for both the RSE and SPAS.

In the next step, DiStefano and Motl used an SEM model to examine whether wording effects across different substantive areas (RSE and SPAS) were potentially correlated. The results confirmed a relation between negative wording factors across the two scales. In the final step they examined the potential relation of four personality traits with method effects found among negatively worded items. Personality traits were assessed with four scales: 1) Short version of Marlowe-Crowne Social Desirability scale (Greenwald & Satow, 1970); 2) BIS/BAS scale (Carver & White, 1994) to measure anxiety and impulsivity; 3) Fear of Evaluation Scale (FNE; Leary, 1983); 4) Measure of self-consciousness (SC; Fenigstein et al., 1975). The scales were selected based on possible explanations offered in literature for response styles related to negatively worded items (e.g., Leary, 1983). Negatively worded items were excluded from the four personality scales to prevent potential confounding effects. The results showed that whereas two scales (SC and FNE) had a

significant, negative relation with the method factor associated with negatively worded items, three other scales (BIS/BAS, Social Desirability) were not significantly related with the method factor. The path analysis suggested that people with greater fear of negative evaluation were less likely to demonstrate the presence of a method effect associated with negatively worded items. Similarly, individuals with higher self-consciousness scores were less likely to demonstrate presence of method effects. It is important to acknowledge that the two personality scales accounted for a rather low percentage of the variation in the method factor ($R^2 = 0.08$).

As mentioned earlier, Roszkowski and Soven (2010) examined the inclusion of two negatively worded items in two course evaluation questionnaires with 14 items to 18 items, respectively. Their results of the exploratory factor analyses showed that a method factor defined solely by the two items would disappear once the items were stated in a positive mode. Further, as mentioned before, Schmitt and Stults (1985) found in their simulation study that when at least 10% of the respondents are careless about recognizing negatively worded items, factor analysis revealed a clearly definable factor solely comprised of all negatively worded items.

Not all research suggests the aberrant psychometric behavior of negatively worded items. Bergstrom and Lunz (1998) used Item Response Theory, specifically the Andrich Rating Scale Model, to analyze the effect of 19 negatively worded items in a 36 item job satisfaction questionnaire. They found that both the positively and negatively worded items appeared to be measuring the

same construct. Barnette (2000) administered a 20 item questionnaire that assessed the attitude toward year-round schooling to 605 high school students, university students and in-service teachers. The questionnaire contained four forms with the direction of the rating scales (strongly disagree (SD) to strongly agree (SA) and SA to SD) crossed with no negatively worded stems and negatively worded stems. Analysis of variance revealed no significant interaction between the two factors and no significant main effects. While the latter mentioned studies suggest that negatively worded items do not show aberrant behavior, the majority of the studies reported in this section resulted in opposite findings. Thus, more work is needed to better determine how the psychometric properties of an attitude scale, interest inventory, or personality scales are influenced by the presence of negatively worded items.

Reasons for Aberrant Behavior of Negatively Worded Items

Although some researchers did not discuss the reasons for method biases introduced by negatively worded items that were found in their studies (e.g., Barnette, 2000; Schriesheim & Hill, 1981), others suggested potential reasons (e.g., Pilotte & Gable, 1990; Schmitt & Stults, 1985; Steward & Frye, 2004), and yet others (e.g., DiStefano & Motl, 2006; Marsh, 1986) examined the relationship between variables that were thought to be potentially associated (e.g., age) with the method bias introduced by negatively worded items.

For example, after finding two distinct factors (one factor comprised of positively worded items and one factor comprised of negatively worded items) in a questionnaire that included both negatively and positively worded items, which

in the all positive and all negative forms resulted in one factor structure, Pilotte and Gable (1990) concluded that: “the mixing of positive and negative item stems on an affective instrument should be viewed with caution since it appears that the two sets of items do not define a single construct” (p. 609). Similarly, Steward and Frye (2004) raised the question whether the different version of an item (i.e., negatively and positively worded) equally measured the same construct.

As was mentioned before, some authors (e.g., Schmitt and Stults, 1985) suggest that inattentiveness to negatively worded items is potentially responsible for the aberrant behavior of negatively worded items. Marsh (1986) examined the relationship between age and the use of positively and negatively worded items. He used the multifactor *Self-Description Questionnaire* (SDQ; Shavelson, Hubner, & Stanton, 1976) for preadolescent children, which included 12 negatively worded items in a 36 items questionnaire. The questionnaire was administered to 658 children in Grades 2 to 5. The results indicated that no correlation (0.02) was found between positively and negatively worded items for Grade 2. The correlations decreased substantially for Grade 3 (-.42) to Grade 4 (-0.60) and Grade 5 (-0.59).These results indicated that younger children more often responded more truthfully to negatively worded items than older children, indicating a low self-concept, even if their responses to similar positively worded items indicated a high self-concept. In another study, which was presented in the previous section, Cordery and Sevastos (1993) found no relationship between educational level among adults and the emergence of a factor comprised of negatively worded items in a Job Diagnostic Survey.

Probably, the most influential recent work that attempted to identify the variables that influence responses to negatively worded items was done by DiStefano and Motl and their colleagues (DiStefano & Motl, 2006; Horan, DiStefano, & Motl, 2003; Motl & DiStefano, 2009). DiStefano and Motl (2006) conducted a series of path analyses to examine a potential relationship between personality traits that influence the responses to negatively worded items. They found that participants who were more concerned with negative evaluations by others or had greater values in self-consciousness were less likely to show the presence of a method effect related to the negative phrasing of the items across two scales with different content (RSE and SPAS). Based on their results DiStefano and Motl argued that adverse effects associated with negative wording "are a type of response style, rather than a substantively irrelevant artifact" (p. 460). It is interesting to note that the authors do not use the word acquiescence for the found response style, but instead refer to personality traits (e.g., concern with negative evaluation) related to method effects associated with negatively worded items. In a later study, Motl and DiStefano (2009) examined the multi-group invariance (male and female students) in relation to the method effects due to the negatively worded items of the Rosenberg Self Esteem Scale (RSE). The authors found no differences in the method effect factor by gender. Within both studies the authors discussed several other factors that might be responsible for the method effects associated with negatively worded items. For example, DiStefano and Motl (2006) stated that: "There are other factors, such as the substantive content of the study, as well as personality factors and demographic

characteristics of the respondents that might exert an influence on item responses" (p. 461). They furthermore discussed the possibility of confounding variables such as characteristics of the scaling method (e.g., number of rating points). If one takes into consideration that the discussed factors might be interacting with each other, research that seeks to delineate the relationship between the factors and method effects associated with negatively worded items becomes even more complicated.

Summary and Research Questions

Historically the inclusion of negatively worded items was introduced as a technique to control acquiescence. It was recommended that an equal number of negatively worded items be included in scales to offset the acquiescent effect (Edwards, 1957). Subsequently, researchers suggested and found that the inclusion of negative items could address issues associated with the lack of carefulness in responding to questionnaires. The examination of the relationship between cognitive processes and errors in questionnaires then became of interest (Alwin, 2010). As a result, the inclusion of negatively worded items has been recently recommended to prevent careless responding by engaging the respondents in a more elaborate cognitive process (Podsakoff et al., 2003).

If careless responding is the main reason for the inclusion of negatively worded items, the question arises whether balanced scales are still necessary. If balanced scales are not necessary, the next question concerns the ratio of negatively to positively worded items. A review of several resources on the construction of questionnaires revealed that at this time there is no consensus on

the ratio of negatively to positively worded items. In fact, no studies have been found that examined the effect of varying number of negatively worded on the questionnaire responses. Therefore, studies that examine the effect of varying number of negatively worded items are needed. For example, the inclusion of relatively few negatively worded items might not elicit the anticipated effect of students being less careless because they might not spot the negatively worded items. Indeed, several researchers pointed out (e.g., Schmitt & Stults, 1985) that while the purpose of including negatively worded items is to prevent careless responding, clearly identifiable method effects associated with negative wording can emerge when at least 10% of the respondent fail to identify negatively worded items. To overcome this adverse effect several researchers (e.g., Roszkowski & Soven, 2010) have recommended providing a sentence in the questionnaire instructions that informs the respondents about the inclusion of negatively worded items. Yet no studies were found that empirically examined the effect of warning about the inclusion of negatively worded items.

As stated in Chapter 1, the purposes of the present study were to begin to address the issues related to the lack of studies that have empirically investigated the effects of varying number of negatively worded items on the questionnaire responses, and including a warning about the inclusion of negatively worded items. More specifically, the research questions addressed in the present study were:

1. Does the inclusion of a varying number of negatively worded items (three and six) and the provision/no provision of a warning about the

inclusion of negatively worded items in a 12 items attitude questionnaire have an effect on the means of the questionnaire items?

2. Is there an interaction effect between the number of negatively worded items (three and six) and the provision/no provision of a warning about the inclusion of negatively worded items?

The corresponding statistical hypotheses were:

1. Research Question

$$H_0: \sum_{j=1}^{4} \alpha_j^2 = 0$$

$$H_1: \sum_{j=1}^{4} \alpha_j^2 \neq 0,$$

where α is the effect of treatment j and is defined as $\alpha_j = \mu_j - \mu$.

2. Research Question

$$H_{0_{AB}}: \sum_{j=1}^{2} \sum_{k=1}^{2} \alpha\beta_{jk}^2 = 0$$

$$H_{1_{AB}}: \sum_{j=1}^{2} \sum_{k=1}^{2} \alpha\beta_{jk}^2 > 0,$$

where there are two levels of Factor A and two levels of Factor B, and

$\alpha\beta$ is the interaction effect between Factor A and B.

Chapter 3: Method

The methods used to address the purposes of the study provided in Chapter 1 and re-stated together with the research questions at the end of Chapter 2 are described in the present chapter. The instrument and its different forms are described in the first section. The participant selection procedure is provided in the second section. The third section includes the information on the ethics approval for the study. The fourth section describes the data entry, followed by the statistical procedures used to analyze the data.

Instrument

The instrument consisted of five variations of a 12 item questionnaire adapted from the self-report inventory *Conceptions of Assessment* (*COA-III Abridged*) designed to obtain teachers' conceptions of assessment (Brown, 2006). Brown used a 6-point positively packed agreement rating scale with two points with negative valence and four with positive valence. His argument was that the teachers' attitudes toward assessment would generally be positive and therefore a positively-packed scale would generate more variance. Using a confirmatory approach, Brown (2006) used Structural Equation Modeling to evaluate the model fit for 27 items. The model fit statistic revealed good fit displaying a multilevel, multifactorial model of nine first order factors with seven first order factors loading on one of two second order factors. The two second order factors were: Assessment improves education and Assessment is irrelevant. The two first order factors that did not load on the second order factors referred to accountability: Assessment makes school accountable and Assessment makes students

accountable. The two first order factors that did not load on either of the second order factors and the two second order factors were intercorrelated displaying low to moderate correlations (Brown, 2006).

As part of a larger study in being conducted at the University of Alberta to determine the students' conceptions and experiences related to assessment (Poth, Riedel, & Luth, 2011), the 27 items in the COA-III Abridged were changed to reflect assessment in the postsecondary context. For example, the item *Assessment is a good way to evaluate a school* was changed to *Assessment is a good way to evaluate the university.* Another change was made to the rating scale. Brown's assumption that a positively packed rating scale would generate more variance could not be made (Brown, 2004). Therefore a 5-point Likert rating scale was used and only the anchor points were labeled (*strongly disagree* and *strongly agree*). According to Lam and Klockars (1982) labeling only the endpoints produces results similar to results where each response option is labeled. They further indicated labeling only the endpoints leads to an interval scale.

Using the responses of 269 university students enrolled in undergraduate courses who completed the questionnaire, an exploratory factor analysis was conducted to determine the factor structure of the data. The number of factors was identified using the Kaiser-Guttmann (K-G) rule and Cattell's scree test applied to the eigenvalues yielded by a principal components extraction. The K-G suggested six factors and the scree test suggested three factors or six factors.

A principal axis factoring followed by an oblique transformation (Direct Oblimin; $\delta = 0$) yielded an acceptable interpretable pattern matrix with six

factors. Four of the items displayed complex loadings. The six factors included: 1) Assessment improves education, 2) Assessment is irrelevant, 3) Assessment is inaccurate, 4) Assessment is valid, 5) Assessment describes abilities, and 6) Assessment makes universities accountable. The six factors accounted for 47% of the total variance. The six factors were low to moderately correlated with each other in both negative and positive directions (Poth et al., 2011).

Due to the restricted amount of time (10 min) to administer the instrument in the present study, only 12 items were chosen from three of the six factors. Five items were selected from the first factor *Assessment improves education (*e.g., *Assessment helps students improve their learning*), four items were selected from the second factor *Assessment is irrelevant* (e.g., *Assessment results are filed away and ignored.*), and three items were chosen from the sixth factor *Assessment makes universities accountable* (e.g., *Assessment is a good way to evaluate the university*). The following criteria were used to select the items:

1. High loadings on the corresponding factor.
2. Amenability to rewriting so that the direction of the revised stem was in the opposite direction.

For example, the four items on the second factor *Assessment is irrelevant* were initially worded negatively, these items had to be reversed to positively worded items (e.g., *Assessment is unfair to students* into *Assessment is fair to students*) to create the all positively worded form (Form p), which constituted the baseline form in the present study.

The same 12 items were used in all five forms, and the same item order was maintained. As mentioned above, all 12 items within Form p were worded in the positive mode. Next, the four alternative forms were constructed by modifying either three (two forms) or six items (two forms) in the baseline form (Form p), so that they were negatively worded. These items were either negated (e.g., *Assessment in not integrated with instruction*) or a polar opposite was formulated (e.g., *Assessment is unfair to students*).

The full set of 12 items as initially stated in the initial COA-III Abridged form are listed in the left column of Table 1. The 12 items included in Form p are listed in the middle column of Table 1. The number at the end of each item indicates which of the three factors the item belonged to. For example, the first item belonged to the first factor .The direction of the four items in italics was changed from Brown's initial form to create the all positively worded Form p. The sets of items in the four alternative forms are listed in the right column. The set of items presented in bold are the three negatively worded items within the two forms with three negatively worded items; the set of items presented in combination of bold and italics are the three additional negatively worded items within the two forms with six negatively worded items. The two forms with three negatively worded items differed in that one form contained a warning about the inclusion of negatively worded items and the other form did not; the same is true for the two forms with the six negatively worded items. The five forms are referred to form p (all positive; control form); form n3 (no warning); form n3w (warning); form n6 (no warning); and form n6w (warning).

Table 1

Initial COA-III Abridged Item Wording and Modified Positively Worded and Negatively Worded Items in the Five Forms

Initial COA-III Abridged items	Modified items Form p	Items presented in the Form n3, n3w, n6, 6w
Assessment helps students improve their learning.	Assessment helps students improve their learning.(1)	Assessment helps students improve their learning.
Assessment interferes with teaching.	*Assessment supports teaching*.(2)	Assessment supports teaching.
Assessment is an accurate indicator of the school's quality.	Assessment results are an accurate indicator of the university's quality.(3)	**Assessment results are an <u>inaccurate</u> indicator of the university's quality[a].**
Assessment information modifies ongoing teaching of students.	Assessment results modify the ongoing instruction of students.(1)	***Assessment results <u>do not modify</u> the ongoing instruction of students[b].***
Assessment feeds back to students their learning needs.	Assessment provides information to students about their strengths and areas that need to be addressed.(1)	Assessment provides information to students about their strengths and areas that need to be addressed.
Assessment is an imprecise process.	*Assessment is a precise process*.(2)	***Assessment is an <u>imprecise</u> process.***
Assessment results are filed away and ignored.	*Assessment results are used and acknowledged by instructors*.(2)	Assessment results are used and acknowledged by instructors.
Assessment is integrated with teaching practice.	Assessment is integrated with instruction.(1)	**Assessment <u>is not integrated</u> with instruction.**
Assessment is a good way to evaluate a school.	Assessment is a good way to evaluate the university.(3)	***Assessment is a <u>bad</u> way to evaluate the university.***
Assessment allows different students to get different instruction.	Assessment informs instruction to meet specific learning needs.(1)	Assessment informs instruction to meet specific learning needs.
Assessment provides information on how well schools are doing.	Assessment provides information on how well the university is doing.(3)	Assessment provides information on how well the university is doing.
Assessment is unfair to students.	*Assessment is fair to students*.(2)	**Assessment is <u>unfair</u> to students.**

[a] *Items presented in bold are the negatively worded items within the Forms n3, n3w, n6 and n6w.*

[b] *Items presented in bold and italics are additional negatively worded items within the Forms n6 and n6w.*

Within the Forms n3 and n3w each negatively worded item belonged to a different factor. To reduce common method variance due to the clustering of items that belong to the same domain, the items belonging to the three factors were intermixed (Kline, Sulsky, & Rever-Moriyama, 2000). A copy of each of the five forms is provided in the Appendix I - V.

Participants

The participants in the present study were undergraduate students enrolled in five upper level undergraduate Educational Psychology courses with three sections at the 300-level (same course, three different sections) and two sections at 400-level (two different courses). It was expected that some students would be enrolled in 300-level as well as 400-level classes. Therefore, before the administration of the questionnaire the students were told to indicate a D on the questionnaire in case they had already responded to the questionnaire.

The data collection took place in three consecutive days. The questionnaire was administered at the end of the classes in the 300-level courses and at the beginning of the classes in the 400-level courses. The five forms were spiraled and distributed to the students within each class in the following order: 1, 2, 3, 4, and 5 so on as to achieve effectively random samples for each questionnaire. The expected time for the completion of the questionnaire was ten minutes. All but few students made use of this time to complete the questionnaire.

Ethics

Ethical approval for this study was granted by Faculties of Education, Extension and Augustana Research Ethics Board (EEA REB) at the University of

Alberta on January 25, 2011. The approval was forwarded to the supervisor, Dr. W. Todd Rogers, and approved on February 22, 2011.

Data Entry

Prior to data entry, nine forms with a D marked on the form were removed. The student responses for the remaining forms were entered manually into SPSS with 100% verification. Two students were removed due to the selection of more than one point on the rating scale. Analysis of missing data revealed that across the five forms 16 participants (4.6%) omitted at least one item. The responses of these participants were deleted to yield a data set with complete data. The total number of students that completed their forms was 333, of which two-thirds were enrolled in the three 300 level courses and one-third were enrolled in the two 400 level courses. Of the 333 students, 68 completed Form p, 69 completed Form n3, 66 completed Form n3w, 66 completed Form n6, and 64 completed Form n6w.

Statistical Analyses

Prior to conducting the statistical analyses, the polarity of negatively worded items in Forms n3, n3w, n6 and n6w was reversed so that the sum of the item scores could be validly compared. The statistical analyses involved two steps.

First, descriptive statistics were computed to describe each of the 12 items across the five forms. The descriptive statistics included the item means, item standard deviations, and the distribution of responses across the five scale points for each item.

Second, inferential statistics at the item level were computed to test the two pairs of statistical hypotheses to answer the corresponding research questions. For the first set of statistical hypotheses, Levene's test of homogeneity of variance was conducted for each item to see if it was possible to pool the variances of the five forms. As will be shown, there was homogeneity of variance. Thus the variances could be pooled to get the Mean Square which was used in the denominator for Dunnett's *t*-statistic (Glass & Hopkins, 1996). The formula for Dunnett's t-statistic is given by

$$t_{D,df_{res}} = \frac{\bar{Y}_{.j} - \bar{Y}_{.p}}{\sqrt{\frac{MS_w}{n} \sum_{j=1}^{j} c_j^2}},$$

where $\bar{Y}_{.j}$ is the mean of the j_{th} group, j = 2, 3, 4, 5.

$\bar{Y}_{.p}$ is the mean for the control group,

MS_w is the pooled residual mean of squares, and

c is the number of groups to be compared with the control group.

Dunnett's *t*-tests were conducted to test the significance of the difference between the mean of each of the experimental forms and the mean of the form p, which was treated as the control condition, for each item.

Third, to examine the second research question a two-way fixed effects ANOVA was conducted to examine the effect of the interaction between the number of negatively worded items (three and six) and the provision/no provision of a warning that negative items were included in the set of items as shown in Figure 1. Given the exploratory nature of this study and the fact that a Type II error is more costly, the 0.05 level of significance was used.

	All positively worded items	Three (3) negatively worded items	Six (6) negatively worded items
No warning	*Form p*	*Form n3*	*Form n6*
Warning		*Form n3w*	*Form n6w*

Figure 1. Research Design

The difference between two means that were significantly different was evaluated with the effect size $\hat{\Delta}$ provided by Cohen (1992)**.** In Cohen's specifications, a small effect corresponds to $\hat{\Delta}$ = 0.20, a medium effect size corresponds to $\hat{\Delta}$ = 0.50, and a large effect size corresponds to $\hat{\Delta}$ = 0.80. Beyond these guidelines, Cohen provided no other criteria such as the range for small, medium, and large effects. For the present study, the following ranges were established: $\hat{\Delta}$ = 0.00 to 0.35 corresponds to small effect size, $\hat{\Delta}$ > 0.35 to 0.65 corresponds to a medium effect size, and $\hat{\Delta}$ > 0.65 corresponds to a large effect size.

Chapter 4: Results

The results for the descriptive and inferential statistical analyses described in the previous chapter are provided in Chapter 4. Interpretation of the results and implications are discussed in Chapter 5. Within Chapter 4 the item level descriptive statistics for each of the five questionnaire forms are reported first. The results of Dunnett's *t*-test and the results for the 2 x 2 fixed effects ANOVA for each item are then presented. The chapter concludes with a summary of the results of the descriptive analysis and the two inferential tests.

Descriptive Results

The frequency distributions, means and standard deviations for each item are reported in Table 2. The ranges of the means of the five forms for each item are similar for the first 11 items and greater for the twelfth item. Compared with the second item, *Assessment supports teaching*, the range of the means for the twelfth item, *Assessment is unfair to students*, is considerably greater (0.10 vs. 0.63). An interesting observation is that whereas the second item was positively worded across all five forms, the twelfth item was worded negatively across all forms other than baseline Form p. The examination of the mean ranges of the remaining items, which revealed that six negatively worded items tended to have higher mean ranges than the six positively worded items. This finding suggests the possibility of wording effects associated with the negatively worded items.

Table 2

Distribution of responses (frequencies in %), means, standard deviations (SD) for each item across the five forms

Item	n*	Form	Scale points (frequencies in %)					Mean	SD
			1	2	3	4	5		
1. Assessment helps students improve their learning.	68	p	1.4	7.2	20.3	42.0	29.0	3.90	0.96
	69	n3	2.8	5.6	16.9	53.5	21.1	3.85	0.92
	66	n3w	0.0	7.2	21.7	50.7	20.3	3.80	0.83
	66	n6	1.4	6.8	21.9	45.2	24.7	3.82	0.96
	64	n6w	1.5	3.0	26.9	38.8	29.9	3.95	0.92
2. Assessment supports teaching.	68	p	0.0	10.1	21.7	33.3	34.8	3.93	0.99
	69	n3	1.4	5.6	23.9	47.9	21.1	3.83	0.89
	66	n3w	0.0	4.3	24.6	50.7	20.3	3.85	0.77
	66	n6	0.0	4.1	24.7	46.6	24.7	3.87	0.81
	64	n6w	1.5	1.5	23.9	49.3	23.9	3.92	0.82
3. Assessment results are an inaccurate indicator of the university's quality.	68	p	10.1	24.6	40.6	20.3	4.3	2.82	1.01
	69	n3	15.5	29.6	33.8	21.1	0.0	2.57	0.98
	66	n3w	4.3	21.7	44.9	29.0	0.0	2.97	0.82
	66	n6	11.1	27.8	41.7	18.1	1.4	2.71	0.97
	64	n6w	10.4	23.9	40.3	23.9	1.5	2.78	0.97
4. *Assessment results do not modify the ongoing instruction of students.*	68	p	8.8	14.7	22.1	38.2	16.2	3.38	1.18
	69	n3	1.4	18.6	32.9	35.7	11.4	3.36	0.97
	66	n3w	4.3	18.8	26.1	40.6	10.1	3.36	1.03
	66	n6	9.7	23.6	26.4	26.4	13.9	3.15	1.20
	64	n6w	12.1	16.7	27.3	27.3	16.7	3.19	1.25

Table 2 - Continued

Item	n	Form	Scale points (frequencies in %)					Mean	SD
			1	2	3	4	5		
5. Assessment provides information to students about their strengths and areas that need to be addressed.	68	p	7.2	8.7	17.4	37.7	29.0	3.72	1.19
	69	n3	8.5	16.9	18.3	40.8	15.5	3.41	1.17
	66	n3w	4.3	20.3	17.4	42.0	15.9	3.48	1.07
	66	n6	5.5	23.3	9.6	39.7	21.9	3.41	1.25
	64	n6w	6.1	15.2	18.2	37.9	22.7	3.56	1.17
6. Assessment is an imprecise process.	68	p	15.9	23.2	31.9	27.5	1.4	2.75	1.08
	69	n3	16.9	22.5	40.8	12.7	7.0	2.68	1.12
	66	n3w	20.3	29.0	30.4	20.3	0.0	2.53	1.01
	66	n6	12.9	24.3	40.0	18.6	4.3	2.78	1.03
	64	n6w	13.4	20.9	40.3	20.9	4.5	2.81	1.08
7. Assessment results are used and acknowledged by instructors.	68	p	4.3	10.1	33.3	43.5	8.7	3.41	0.95
	69	n3	1.4	20.0	31.4	30.0	17.1	3.39	1.03
	66	n3w	1.5	14.7	33.8	44.1	5.9	3.38	0.87
	66	n6	2.7	15.1	26.0	42.5	13.7	3.47	1.08
	64	n6w	4.5	13.4	28.4	47.8	6.0	3.36	0.97
8. Assessment is not integrated with instruction.	68	p	7.2	8.7	24.6	47.8	11.6	3.47	1.06
	69	n3	1.4	10.0	34.3	41.4	12.9	3.55	0.90
	66	n3w	0.0	21.7	36.2	39.1	2.9	3.20	0.83
	66	n6	2.8	18.1	25.0	36.1	18.1	3.47	1.08
	64	n6w	1.5	9.0	34.3	40.3	14.9	3.56	0.91

Table 2 - Continued

Item	n	Form	Scale points (frequencies in %)					Mean	SD
			1	2	3	4	5		
9. Assessment is a bad way to evaluate the university.	68	p	11.6	24.6	42.0	20.3	1.4	2.75	0.96
	69	n3	14.1	36.6	38.0	11.3	0.0	2.49	0.87
	66	n3w	13.0	30.4	34.8	18.8	2.9	2.71	0.99
	66	n6	2.8	18.1	25.0	36.1	18.1	2.77	0.99
	64	n6w	7.6	27.3	36.4	22.7	6.1	2.92	1.03
10. Assessment informs instruction to meet specific learning needs.	68	p	4.3	15.9	33.3	29.0	17.4	3.40	1.09
	69	n3	5.6	14.1	28.2	45.1	7.0	3.35	1.00
	66	n3w	6.0	19.4	31.3	40.3	3.0	3.12	0.95
	66	n6	2.8	20.8	29.2	34.7	12.5	3.33	1.07
	64	n6w	6.1	24.2	25.8	36.4	7.6	3.14	1.08
11. Assessment provides information on how well the university is doing.	68	p	8.7	29.0	43.5	15.9	2.9	2.75	0.94
	69	n3	8.5	38.0	31.0	21.1	1.4	2.68	0.95
	66	n3w	8.8	32.4	36.8	22.1	0.0	2.77	0.87
	66	n6	5.5	27.4	34.2	31.5	1.4	2.92	0.95
	64	n6w	7.5	32.8	38.8	17.9	3.0	2.83	0.95
12. Assessment is unfair to students.	68	p	15.9	15.9	43.5	20.3	4.3	2.82	1.08
	69	n3	5.6	21.1	36.6	28.2	8.5	3.13	1.03
	66	n3w	4.3	18.8	31.9	30.4	14.5	3.29	1.09
	66	n6	4.1	12.3	43.8	28.8	11.0	3.29	1.00
	64	n6w	3.0	14.9	32.8	31.3	17.9	3.45	1.04

Notes. Items presented in bold are negatively worded items within the Forms n3, n3w, n6 and n6w.
Items presented in bold and italics are additional negatively worded items within the Forms n6 and n6w.
The polarity of the negative items was reversed for the negative item.

Examination of ranges for standard deviations (SD) of the five forms across the 12 items revealed greater similarity across items when compared with the means (as reported above). The smallest range (0.08) was found for eleventh item, *Assessment provides information on how well the university is doing*; the largest range (0.25) was found for the seventh and eight items, *Assessment results are used and acknowledged by instructors* and *Assessment is not integrated with instruction.*

Examination of the frequency distributions of responses across the five scale points revealed two trends related to score distribution of groups of items. In particular, responses to the scale points to the seven items pertaining to the use of assessment for formative purposes (e.g., 10. *Assessment informs instruction to meet specific learning needs)* tended to cluster to the right side of the scale, indicating that on average students agreed with the statements (i.e., selected favorable rating points) (see Table 2). In contrast, the students' responses to the three items that were related to use of assessment results as an indicator of university's quality tended to cluster, but to a lesser degree than above, to the left side, indicating that students showed a tendency to disagree (i.e., select unfavorable rating points). Similar to the ratings related to the item group related to university's quality, the students' responses to the sixth item, *Assessment is an imprecise process*, also tended to cluster to the left side of the rating scale. Whereas the scale point distributions were similar across the five forms for the first eleven items, the students' responses to the twelfth item, Assessment *is unfair to students,* revealed that scale point distributions between the positive

form and the four forms with negative items differed considerably. For this item the scores tended to cluster on the left side of the rating scale for the positive form, and on the right side for the four forms that included negative items. Given similar results were found for the twelfth item across the four forms that included negatively worded items, the difference between these forms and the positive form might be an indicator that negative phrasing of this particular item was perceived differently by the students when it was negatively worded than when it was positively worded.

Inferential Results

As indicated in Chapter 3, Levene's homogeneity of variance test revealed that the five variances for each item were homogeneous ($p < 0.05$). Therefore, the variances for the five forms were pooled to obtain the residual mean square (MS_w) and degrees for freedom (df_w) needed for Dunnett's *t*-test. Given the five sample sizes were not equal but close in value, the harmonic mean of the sample sizes, 66.65, was used in Dunnett's *t*-test. The results of the four simple contrasts (Form p vs. each experimental form) carried out with Dunnett's *t*-statistic at the item level are reported in Table 3. The findings indicate that three statistically significant contrasts were found only for the twelfth item. Although the range of means for the negatively worded items tended to be greater than the range of means for the positively worded items, the differences between pairs of means were not statistically significant other than for item 12.

Table 3

t-values for Dunnett's test

Items	MS_w	t_{p-3n}	t_{p-n3i}	t_{p-n6}	t_{p-n6i}
1. Assessment helps students improve their learning.	0.85	0.31	0.63	0.50	-0.31
2. Assessment supports teaching.	0.75	0.67	0.53	0.40	0.07
3. Assessment results are an inaccurate indicator of the university's quality.	0.90	1.52	-0.91	0.67	0.24
4. Assessment results do not modify the ongoing instruction of students.	1.27	0.10	-0.10	1.18	0.97
5. Assessment provides information to students about their strengths and areas that need to be addressed.	1.37	1.53	1.18	1.53	0.79
6. Assessment is an imprecise process.	1.14	0.38	1.19	-0.16	-0.32
7. Assessment results are used and acknowledged by instructors.	0.94	0.12	0.18	-0.36	0.30
8. Assessment is not integrated with instruction.	0.92	-0.47	1.63	0.01	-0.53
9. Assessment is a bad way to evaluate the university.	0.94	1.55	0.24	-0.12	-1.01
10. Assessment informs instruction to meet specific learning needs.	1.08	0.28	1.55	0.39	1.44
11. Assessment provides information on how well the university is doing.	0.87	0.43	-0.12	-1.05	-0.50
12. Assessment is unfair to students.	1.10	-1.71	-2.59*	-2.59*	-3.47*

Note. Items presented in bold are the negatively worded items within the Forms n3, n3w, n6 and n6w.
Items presented in bold and italics are additional negatively worded items within the Forms n6 and n6w.
Note. The harmonic mean (ñ) for the five Forms was 66.55. It was rounded down to 66 when obtaining critical values for Dunnett's t-test.
* $p < .05$

For item 12, statistically significant differences were found between Form p and Form n3w: $t(5, 333) = 2.59$, $p < 0.05$; Form p and form 6n: $t(5, 333) = 2.59$, $p < 0.05$; and Form p and form n6w: $t(5, 333) = 3.47$, $p < 0.05$. In all three cases, the mean of the form containing negatively worded items was significantly greater than the mean of the Form p (after reversing the polarity) (3.28 vs. 2.82; 3.28 vs. 2.82; 3.45 vs. 2.82). The effect sizes were moderate: the values of Cohen's $\hat{\Delta}$ were 0.40, 0.43, and 0.49, respectively. This suggests that these three differences deserve practical consideration, although somewhat troubling is the lack of a significant difference between Form p and Form n3.

The results of the 2 x 2 (number of negative items-by-presence of warning) indicated that none of the interactions were statistically significant, meaning that the differences in means within the number of negatively worded items did not depend on the presence of a warning and vice versa. However, two statistically significant main effects were found for two items. Therefore, in the interest of space, the results of the factorial ANOVA for only these two items are present in Table 4. The full set of results for all 12 items is provided in Appendix VI. As shown in Table 4, a statistically significant main effect was found for the factor warning/no warning for the third item, *Assessment results are an inaccurate indicator of the university's quality*, $F(1, 261) = 4.22$, $p < 0.05$. The mean for the forms with warning ($M = 2.88$, $SD = 0.89$) was significantly greater than the mean score for the forms with no warning ($M = 2.63$, $SD = 0.97$). However the effect size, $\hat{\Delta} = 0.28$, was weak.

Table 4

Factorial ANOVAs for the factor varying number of items and provision/no provision of a warning

Items	Source	Df	MS	F
3. Assessment results are an inaccurate indicator of the university's quality.	number	1	0.03	0.03
	warning	1	3.71	4.22*
	n x w	1	1.86	2.19
	error	261	0.88	
9. Assessment is a bad way to evaluate the university.	number	1	3.97	4.23*
	warning	1	2.25	2.39
	n x w	1	0.08	0.09
	error	261	0.94	

Note. Items presented in bold are the negatively worded items within the Forms n3, n3w, n6 and n6w.
Items presented in bold and italics are additional negatively worded items within the Forms n6 and n6w.
* $p < .05$

The second statistically significant main effect was found for the factor number of negatively worded items for the ninth item, *Assessment is a bad way to evaluate the university*, F (1, 261) = 4.23, $p < 0.05$). The mean score for the forms including six negatively worded items (M = 2.85, SD = 1.01) was significantly greater than the mean score (M = 2.60, SD = 0.93) for forms including three negatively worded items. Similar to the first main effect, the effect size was weak, $\hat{\Delta}$ = 0.31, suggesting that this main effect is of little practical interest.

Summary

The descriptive results showed that overall the mean ranges tended to be greater for items that were worded negatively, suggesting the possibility for wording effect associated with negatively worded items. In contrast to the means, ranges for the standard deviations were more similar in value for both positively

and negatively worded items. Examination of the frequency distributions of responses across the five scale points showed that items related to the formative purpose of assessment tended to cluster to the right side of the scale, indicating that on average students tended to agree with the statements. In contrast, the students' responses items that were related to use of assessment results as an indicator of university's quality showed a weak tendency to cluster to the left side. For the twelfth item, *Assessment is unfair to students*; the score distributions differed considerably for positive and negative forms, suggesting that negatively worded form was perceived differently than the positively worded form of the same item. When the means differences were examined at the item level using inferential tests, the results revealed that significant differences for means were found solely for the twelfth item, *Assessment is unfair to students.*The means for forms n3w, n6, and n6w were significantly greater than the mean for the form p, and the corresponding effect sizes were moderate, respectively. These differences were all of moderate size and deserve practical consideration, which will be provided in the next chapter. No statistically significant interactions were found at the second step of the analysis, indicating that the differences in scores within the factor number of items did not depend on the level of the other factor presence of warning and vice versa. However, two statistically main effects were found, one for number of negatively worded items and the other for presence of warning. However the magnitudes of both these effects were weak and therefore these effects can be discounted.

Chapter 5: Summary, Discussion, and Conclusions

In Chapter 5 the research questions and a brief description of the methods are presented first. A summary of findings is presented next, followed by a discussion of the results. The fourth section includes the limitations of the study. Then, in the light of the limitations the conclusions are presented. Implications for practice are provided in the sixth section. The chapter concludes with recommendations for future research.

Summary of Research Questions and Methods

The purposes of this study were to investigate the effect of the inclusion of a varying number of negatively worded items on the responses of undergraduate students in a short questionnaire of fixed length and the effect of including a warning about the inclusion of negatively worded items on the responses of the undergraduate students. The specific research questions addressed included:

1. Does the inclusion of a varying number of negatively worded items (three and six) and the provision/no provision of a warning about the inclusion of negatively worded items in a 12 items attitude questionnaire have an effect on the means of questionnaire items?
2. Is there an interaction effect between the number of negatively worded items (three and six) and the provision/no provision of a warning about the inclusion of negatively worded items?

In order to examine these research questions, five variations of a 12 item questionnaire adapted from the *Conceptions of Assessment* self-report inventory (*COA-III Abridged*) (Brown, 2006) were used. All 12 items within the baseline

form (Form p) were worded in the positive mode. The four alternative forms were constructed by modifying either three (two forms) or six items (two forms) in the baseline form so that they were negatively worded. Two of these forms incorporated a warning (Form n3w and Form n6w) and remaining two did not (Form n3 and Form n6). The five variations were administered to 333 students enrolled in five senior undergraduate Educational Psychology courses. The five forms were administered in each class in a spiral fashion to achieve effectively five random samples.

To address the first research question, Levene's test was used to test for homogeneity of variance across the five forms at the item level. Given homogeneity was found for each item ($p < 0.05$), Dunnett's *t*-statistic was conducted to test the significance of the difference between the mean of the Form p (treated as the control condition) and each of the experimental forms at the item level. To address the second research question, a two-way fixed effects ANOVA was conducted to examine the effect of the interaction between the number of negatively worded items (three and six) and the provision/no provision of a warning about the inclusion of negatively worded items.

Summary of Findings

The descriptive statistics revealed that the range of the mean responses for the five forms for the individual items tended to be greater for items that were negatively worded than for items that were positively worded, thus suggesting the possibility of wording effects associated with negatively worded items. The standard deviations across the twelve items were comparable. The frequency

distributions of responses across the five scale points tended to be similar across the five forms except for the twelfth item, *Assessment is unfair to students*.

Dunnett's test revealed that except for the twelfth item, *Assessment is unfair to students*, all contrasts between the mean of the baseline form and each of the four experimental forms were statistically nonsignificant at the 0.05 level. In addition, the 2x2 ANOVAs revealed that there were no statistically significant interactions or meaningful main effects. Overall, these findings indicate that rewriting items into a negative form and the inclusion of a warning that there are negative items did not influence student responses regardless of the number of negative items with one exception, Item 12.

Discussion of Findings

The findings for the present study differ from the findings of Roszkowski's and Soven's (2010), who conducted a study with a similar population (undergraduate students), and who concluded that adverse effects associated with the inclusion of negatively worded items found in their study were most likely due to inattentive responding to negatively worded items. One explanation for the divergent findings is that whereas Roszkowski and Soven included two negatively worded items in the middle of a 13 and18 items course evaluation questionnaire, the present study embedded a negatively worded item at the beginning of the questionnaires, so that the students were given the opportunity to notice at a relatively early stage that the wording was reversed for some items. Therefore it is most likely that the students did not establish a mindset that the wording for all the items was positive. Another explanation for

the difference between the results of the two studies is that the proportion of negatively worded items included in the present study was considerably larger (25% and 50%) than the proportion of negatively worded items in Roszkowski's and Soven's study (15% and 11%, respectively).

As indicated above, the findings for the twelfth item, *Assessment is unfair to students,* differed from the findings for the other negatively worded items and remaining positively worded items; n3w, n6, and n6w greater than for form p. The mean responses for both of the forms with six negatively worded items (Form n6 and n6w) and the form with three items negatively worded and a warning (Form n3w) were significantly greater than the mean for the form with only positively worded items and the corresponding effect sizes were of moderate size. There is no apparent reason for why the contrast between the form with three negatively worded items and no warning (Form n3) and the mean of the positively worded form did not differ considerably. The corresponding effect size, while in the same direction as for the other three contrasts, was weak for this contrast.

Despite the finding that Form n3 did not differ from the mean of the Form p for item 12, it is noteworthy that significant differences were found only for this item. The question is why significant differences were found only for this item and not the other five items that were negatively worded items or any of the positively worded items? One possible explanation could be that the negative wording of the twelfth item was created using a polar opposite with what may be an inconspicuous prefix "un", as opposed to other negatively worded items that were created using either polar opposites that are more distinct, for example

"good" vs. "bad" for the ninth item, *Assessment is a bad way to evaluate university*, or using negated regular forms that include a distinct "not" used in the fourth and eighth items (e.g., *Assessment results do not modify the ongoing instruction of students*). Yet, this explanation is unlikely to be correct, since both third and sixth item were created using comparably inconspicuous prefixes "in" as in *inaccurate* (third item) and "im" as in *imprecise* (sixth item).

A second possible reason that might explain the significant effects of the negative wording for the twelfth item can be provided through the use of the CASM model which describes the four cognitive steps involved in answering questions (Tourangeau & Bradburn, 2010; see Chapter 2). Before the CAMS model can be applied to explain the adverse effects of negatively worded items, it is useful to note that attitudinal responses are mostly defined by the valence and the intensity of the item (Olsen, 1999). Valence is concerned with the positive or negative direction. Intensity specifies the strength with which the item elicits an attitude (Schuman & Presser, 1996).

According to CASM model when a respondent is presented with a statement such as *Assessment is unfair to students,* the first step would involve the comprehension of the question, which is assumed to be a rather easy cognitive task for the participants of the present study. During the second step the respondents retrieve different experiences related to assessment being *unfair* to them. Using the retrieved experiences, the respondents make a judgment about the intensity of the response at the third step. Lastly, at step 4 the judgments formed in step 3 are translated to the response option that best reflects the intensity. It is

assumed that the differences in responses between the positive wording of the twelfth item, *Assessment is fair to students,* and the negative wording of the item, *Assessment is unfair to students,* are due to differences in item intensity. A respondent who is retrieving information related to assessment being fair cues the memory for events where he or she experienced assessment as being fair. In contrast, the negative wording in the item, *Assessment is unfair to students,* triggers greater intensity, leading to retrieval of more negative experiences with assessment that may be stronger in nature. Given the strong emotional tone attached to negative experiences with assessment, the intensity of this item is increased when the valence of this item is negative. Therefore, when student are presented with the negative form, *Assessment is unfair to students*, on average they are more likely to agree that assessment is unfair.

Given that the differences in means between the positive forms and the negative form of the other five items that were worded negatively in the experimental forms were considerably smaller, a question is whether the processes within the four cognitive steps described by the CASM were somewhat different for the twelfth item than the other five items?

Examination of the content of the six negatively worded items showed that there is indeed a difference. Except for the twelfth item, none of the other negatively worded items was directly related to the students. Item three and item nine were both related to the function of assessment to evaluate university (3. *Assessments results are an inaccurate indicator of university's quality* and 4. *Assessment is a bad way to evaluate the university*). Items four and eight were

both related to the formative function of assessment to enhance instruction (4. *Assessment results do not modify the ongoing instruction of stud*ents and 8. *Assessment is not integrated with instruction).* Item six was related to the measurement aspect of assessment (6. *Assessment is an imprecise process*). Thus, there was no "personal" referent as there was in item 12. The respondents likely had less or no direct experiences with the referent in the other five items. For example, it is assumed that the participants had only a general idea whether assessment results modified the "ongoing instruction" of the students. A more informed opinion on this item is more likely to be obtained from university instructors. Therefore, it can be argued that the negative wording of these items did not necessarily trigger the retrieval of emotionally charged negative experiences that are likely to result in greater intensity. Moreover, it could be argued that if the twelfth item, *Assessment is unfair to students,* was changed to *In my experience assessment is unfair*, the intensity of this item would be even higher, because the item is even more directed toward the individual student.

However, as noted in Chapter 3 the vast majority of students required the full ten minutes of administration time to complete the 12 items. This finding suggests that the students employed the CASM model for all items.

Limitations of the Study

The study was limited in two ways. First, with only 12 items the questionnaire was rather short. When a longer questionnaire is used, the respondents are more likely to become fatigued or demotivated (Krosnick & Presser, 2010). Thus, the tendency for careless responding (satisficing) is more

likely to increase (Krosnick, 1991). One consequence might be that students are then less likely to be attentive to the negatively worded items.

Second, the majority of the items used in this study were not related to the respondent directly and were written in such a way that they were of low intensity. The only item that was related to the students and had high intensity was the twelfth item in the negative form, *Assessment is unfair to students.* In the context of assessment, the word "unfair" creates intensity within the respondents. Within the other 11 items only two had the student as the referent (1. *Assessment helps students improve their learning* and 5. *Assessment provides information to students about their strengths and areas that need to be addressed*). What remains to be examined is what would happen if the word "really" is inserted before helps in the first item and, perhaps, "meaningful" or "really useful" before information in the second item? Would the negative forms of these items possibly have higher intensities than their positive counterparts?

Conclusions

In the light of the limitations of the present study, two main conclusions were drawn from the results of this study:

1. Varying number of negatively worded items has no effects on the responses of the students.
2. There are no wording effects associated with negatively worded items unless the effect of negative wording changes the intensity of the item.

3. There is no need to provide a warning about the inclusion of the negatively worded items in an attitude scale given that a negatively worded item is placed within the first three items.

Implications for Practice

The results of the present study indicate that the inclusion of a warning is not necessary, regardless of the number of negatively worded items. At the same time its inclusion likely causes no harm. Hence, if inattentiveness to negatively worded items is thought to be possibly present, the questionnaire can include a warning.

The answer to the question whether negatively worded items should be included is a more complex. Based on the results of the present study, it appears that including negatively worded items to avoid the effects of careless responding is warranted. But this recommendation is moderated by the level of intensity of the items. If they have greater intensity than their positive counterpart, then there could be differences due to the presence of the negative wording. If items, when positively written and negatively written have the same intensity, then it is not likely that the scores for these items will differ substantially after the scores for the negatively worded items were reversed. In this case negatively worded items can be included. If, on the other hand, rewording positively worded items into negatively worded is likely to results in higher intensity of the items when negatively worded, then the inclusion of negatively worded items is not recommended.

Future Research

1. As just indicated, one topic in need of additional research is the effect of intensity of both positively and negatively worded items. In the present study significant differences between the positive form and the experimental forms were found for only one negatively worded item. This item referred directly to the respondents and a strong intensity word (*unfair*) was used to achieve the negative polarity. Unfairness to students (me) will arouse strong feelings. Strong emotions are more likely to be triggered by words that elicit intensity and that call for personal experiences that are emotionally charged rather than when more benign or neutral words are used and the item is not related to personal experiences of the respondent. For example, it would be interesting to see whether the changing of the wording of an item such as *Assessment results encourage my learning* to *Assessment results really encourage my learning* to *Assessment results really discourage my learning* would potentially lead to substantial effects related first to the increase in intensity and second to the use of negative wording. Could it be that the negative wording is more likely to cue the memory for instances where assessment discouraged the respondent's learning as opposed to the positive wording which is likely to results in respondent's memory scan with less instances of assessment being discouraging?
2. Think aloud interviews followed by protocol analysis of what the respondents said should be conducted to gain a better understanding of the

cognitive processes that are involved when answering a questionnaire item (Willis, 2005).

3. The relationship between inattentive responding to negatively worded items and questionnaire length needs to be investigated. Are student responses to the items in a longer questionnaire with a relatively small number of negative items influenced differently than in a longer questionnaire with similar content but a relatively large number of negative items? Again, the use of think aloud protocols could be used to clarify the thinking used by the students.
4. The generalizability of the findings to other student populations needs to be examined to clarify the relationship between the educational level of the respondents and their responses. For example, would students in lower undergraduate courses respond in the same way as students in upper level undergraduate courses?
5. During the last decades many studies have been administered utilizing computers to assess achievement. However, as was seen from the review of research no studies involved computer based assessment of attitudes, or the mode of assessment was not specified. Within the measurement of achievement the results are mixed (Dr. Todd Rogers, personal communication, September 26, 2012). Therefore, it is recommended to conduct studies examining the impact of negatively worded items in both administration modes.

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The Authoritarian Personality. Oxford, England: Harpers; England,* Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=psyc1&AN=1950-05796-000; http://resolver.library.ualberta.ca/resolver?sid=OVID:psycdb&id=pmid:&id=doi:&issn=&isbn=&volume=&issue=&spage=&pages=&date=1950&title=The+authoritarian+personality.&atitle=The+authoritarian+personality.&aulast=Adorno&pid=%3Cauthor%3EAdorno%2C+T.+W%2CFrenkel-Brunswik%2C+Else%2CLevinson%2C+Daniel+J%2CSanford%2C+R.+Nevitt%3C%2Fauthor%3E%3CAN%3E1950-05796-000%3C%2FAN%3E%3CDT%3E%3C%2FDT%3E

Alwin, F. A. (2010). How good is survey measurement? Assessing the reliability and validity of survey measures. In P. V. Marsden, & J. D. Wright (Eds.) *Handbook of survey research*. (2nd. ed., pp. 405-436). Howard House, UK: Emerald Group Publishing Limited.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). NJ: Prentice Hall, Inc.

Anderson, L. W. (1981). *Assessing affecive characteristics in schools*. Boston, Mass.: Allyn and Bacon.

Bagozzi, R., & Yi, Y. (1990). Assessing method variance in multitrait multimethod matrices - the case of self-reported affect and perceptions at work. *Journal of Applied Psychology, 75*(5), 547-560. doi:10.1037//0021-9010.75.5.547

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. doi:10.1177/00131640021970592

Bergstrom, B. A., & Lunz, M. E. (1998). Rating scale analysis: Gauging the impact of positively and negatively worded items. *Paper Presented at the Annual Meeting of American Educational Research Association, San Diego, California.*

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (1991). Preface. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement error in surveys*. New York: John Wiley & Sons, Inc.

Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an abridged version. *Psychological Reports, 99*(1), 166-170. doi:10.2466/PR0.99.1.166-170

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Carver, S. C., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology,* (67), 319-333.

Chen, Y., Rendina-Gobioff, G., & Dedrick, R. F. (2010). Factorial invariance of a Chinese self-esteem scale for third and sixth grade students: Evaluating method effects associated with positively and negatively worded items. *The International Journal of Educational and Psychological Assessment, 6*(1), 21-35.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Colston, H. L. (1999). "Not good" is "bad," but "not bad" is not "good": An analysis of three accounts of negation asymmetry. *Discourse Processes, 28*(3), 237-256. doi:http://dx.doi.org/10.1080/01638539909545083

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised job diagnostic survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78*(1), 141-143. doi:10.1037//0021-9010.78.1.141

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*(4), 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3-31. doi:10.1177/001316445001000101

DeVellis, R. F. (2012). *Scale development* (3rd ed.). Los Angeles: Sage.

Dillman, D. A., & Smyth, J.D., Melanie Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, New Jersey: John Wiley&Sons, Inc.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 440-464. doi:10.1207/s15328007sem1303_6

Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York, NY: Appleton - Centrury - Crofts.

Fowler, F. J. (2002). *Survey research methods*. Thousand Oaks, CA: Sage.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn and Bacon.

Gomleksiz, M. H. (2004). Use of educational technology in english classes. *The Turkish Online Journal of Educational Technology – TOJET, 3*(2), January 2, 2012.

Greenwald, H. J., & Satow, Y. (1970) A short social desirability scale. *Psychological Reports*, 27, 131-135.

Guyatt, G., Cook, D., King, D., Norman, G., Kane, S., & van Ineveld, C. (1999). Effect of the framing of questionnaire items regarding satisfaction with training on residents' responses. *Academic Medicine, 74*(2), 192-194. doi:10.1097/00001888-199902000-00018

Horan, P., DiStefano, C., & Motl, R. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, 10(3), 435-455. doi:10.1207/S15328007SEM1003_6

Idaszak, J. R., & Drasgow, F. (1987). A revision of the job diagnostic survey - elimination of a measurement artifact. *Journal of Applied Psychology, 72*(1), 69-74.

Jackson, D. N. (1967). Acquiescence response styles: Problems of identification and control. In I. Berg A. (Ed.), *Response set in personality assessment* (pp. 71-114). Chicago: Aldine Publishing Company.

Jackson, D. N., & Messick, S. J. (1957). A note on ethnocentrism and acquiescent response sets. *Journal of Abnormal and Social Psychology, 54*(1), 132-134. doi:10.1037/h0049281

Kim, H. (2011). Inquiry-based science and technology enrichment program: Green earth enhanced with inquiry and technology. *Journal of Science Education and Technology, 20*(6), 803-814. doi:10.1007/s10956-011-9334-z

Kline, T. J. B., Sulsky, L. M., & Rever-Moriyama, S. D. (2000). Common method variance and specification errors: A practical approach to detection. *Journal of Psychology, 134*(4), 401-421.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. doi:http://dx.doi.org/10.1002/acp.2350050305

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracin, B. T. Johnson & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 21-78). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden, & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263-313). Howard House, UK: Emerald Group Publishing Limited.

Lam, T., & Klockars, A. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*(4), 317-322. doi:10.1111/j.1745-3984.1982.tb00137.x

Leary, M. R. (1983). A brief version of the fear of negative evaluation scale. *Personality and Social Psychology Bulletin*, (9), 371-375.

Magazine, S. L., Williams, L. J., & Williams, M. L. (1996). A confirmatory factor analysis examination of reverse coding effects in Meyer and Allen's affective and continuance commitment scales. *Educational and Psychological Measurement, 56*(2), 241-250. doi:10.1177/0013164496056002005

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children - a cognitive developmental phenomenon. *Developmental Psychology, 22*(1), 37-49. doi:10.1037/0012-1649.22.1.37

Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? RID B-4555-2008. *Journal of Personality and Social Psychology, 70*(4), 810-819. doi:10.1037//0022-3514.70.4.810

Meade, A.W., & Craig, S.B. (2012, April 16). Identifying careless responses in survey data. *Psychological Methods*. Advance online publication. doi: 10.1037/a0028085.

Morrel-Samuels, P. (2002a). Getting the truth into workplace surveys. *Harvard Business Review, 80*(2), 111-118.

Motl, R. W., & DiStefano, C. (2009). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex.

Structural Equation Modeling: A Multidisciplinary Journal, 16(1), 134-146. doi:10.1080/10705510802565403

Narayan, S., & Krosnick, J. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*(1), 58-88. doi:10.1086/297739

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill, Inc.

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

Olsen, S. O. (1999). Strength and conflicting valence in the measurement of food attitudes and preferences. *Food Quality and Preference, 10,* 483-494.

Poth, C., Riedel, A., & Luth, R. (March, 2011). *Undergraduate experiences and attitudes towards assessment: The Canadian context*. Paper presented at the Global Learn Asia Pacific annual meeting on Learning Technology, Melbourne, Australia.

Peterson, R. A. (2000). *Constructing effective questionnaires*. Thousand Oaks: Sage Publications.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity on a computer anxiety scale. *Educational and Psychological Measurement, 50,* 603-610.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. doi:10.1037/0021-9101.88.5.879

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, Alberta: Joint Advisory Committee. (Mailing Address: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, 3-104 Education Building North, University of Alberta, Edmonton, Alberta, T6G 2G5).

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology, 121*(1), 81-96.

Ray, J. J. (1990). Acquiescence and problems with forced-choice scales. *Journal of Social Psychology, 130*(3), 397-399.

Ray, J. (1979). Is the acquiescent response style problem not so mythical after all - some results from a successful balanced F-scale. *Journal of Personality Assessment, 43*(6), 638-643. doi:10.1207/s15327752jpa4306_14

Rorer, L. (1965). The great response-style myth. *Psychological Bulletin, 63*(3), 129-156. doi:10.1037/h0021888

Rosenberg, M. (1965). Society and adolescent self-image. Princeton, N.J.: Princeton University Press.

Rosenberg, M. (1989). Society and the adolescent self-image (Rev. ed.). Middleton, CT: Wesleyan University Press.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*(1), 117-134. doi:10.1080/02602930802618344

Rubin, A., & Babbie, E. (2011). *Research methods for social work* (7th ed.). Belmont, CA: Brooks/Cole.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items - the result of careless respondents. *Applied Psychological Measurement, 9*(4), 367-373.

Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management, 21*(6), 1177-1193.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals - the effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114. doi:10.1177/001316448104100420

Schriesheim, C., Eisenbach, R., & Hill, K. (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. doi:10.1177/0013164491511005

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys*. Thousand Oaks, California: SAGE Publications, Inc.

Shavelson, R., Hubner, J., & Stanton, G. (1976). Self-concept - validation of construct interpretations. *Review of Educational Research,* 46(3), 407-441. doi:10.3102/00346543046003407

Steward, T. J., & Frye, A. W. (2004). Investigating the use of negatively phrased survey items in medical education setting: Common wisdom or common mistake? *Academic Medicine, 79*(10), 18-20.

Tourangeau, R., & Bradburn, N. M. (2010). The psychology of survey response. In P. V. Marsden, & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 347-404). Howard House, UK: Emerald Group Publishing Limited.

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education, 28*(6), 587-607. doi:10.1080/0260293023000120234

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 189-194. doi:10.1007/s10862-005-9004-7

Appendix I: Questionnaire form p

Instructions

I am interested in knowing about your attitudes and beliefs towards post-secondary assessment. Please use a five-point scale (1 - Strongly Disagree …… 5 - Strongly Agree) to indicate to what extent you agree or disagree with each statement.

The term assessment means to measure and evaluate what students have learned.

	Strongly Disagree				Strongly Agree
Assessment…..	1	2	3	4	5
helps students improve their learning.	□	□	□	□	□
supports teaching.	□	□	□	□	□
results are an accurate indicator of the university's quality.	□	□	□	□	□
results modify the ongoing instruction of students.	□	□	□	□	□
provides information to students about their strengths and areas that need to be addressed.	□	□	□	□	□
is a precise process.	□	□	□	□	□
results are used and acknowledged by instructors.	□	□	□	□	□
is integrated with instruction.	□	□	□	□	□
is a good way to evaluate the university.	□	□	□	□	□
informs instruction to meet specific learning needs.	□	□	□	□	□
provides information on how well the university is doing.	□	□	□	□	□
is fair to students.	□	□	□	□	□

If you would like to participate in the draw, please provide your e-mail address. The draw includes one of three Tim Hortons cards worth $15 each.

If you are interested in a summary of the results, please provide your e-mail address.

Thank you for your participation!

Appendix II: Questionnaire form n3

Instructions

I am interested in knowing about your attitudes and beliefs towards post-secondary assessment. Please use a five-point scale (1 - Strongly Disagree …… 5 - Strongly Agree) to indicate to what extent you agree or disagree with each statement.

The term assessment means to measure and evaluate what students have learned.

	Strongly Disagree				Strongly Agree
Assessment…..	1	2	3	4	5
helps students improve their learning.	□	□	□	□	□
supports teaching.	□	□	□	□	□
results are an inaccurate indicator of the university's quality.	□	□	□	□	□
results modify the ongoing instruction of students.	□	□	□	□	□
provides information to students about their strengths and areas that need to be addressed.	□	□	□	□	□
is a precise process.	□	□	□	□	□
results are used and acknowledged by instructors.	□	□	□	□	□
is not integrated with instruction.	□	□	□	□	□
is a good way to evaluate the university.	□	□	□	□	□
informs instruction to meet specific learning needs.	□	□	□	□	□
provides information on how well the university is doing.	□	□	□	□	□
is unfair to students.	□	□	□	□	□

If you would like to participate in the draw, please provide your e-mail address. The draw includes one of three Tim Hortons cards worth $15 each.

If you are interested in a summary of the results, please provide your e-mail address.

Thank you for your participation!

Appendix III: Questionnaire form n3w

Instructions

I am interested in knowing about your attitudes and beliefs towards post-secondary assessment. Please use a five-point scale (1 - Strongly Disagree …… 5 - Strongly Agree) to indicate to what extent you agree or disagree with each statement.

The term assessment means to measure and evaluate what students have learned.

Please pay attention to the wording of the items when you respond; some items are worded in a negative mode (e.g., *assessment results are inconsistent*.).

	Strongly Disagree				Strongly Agree
Assessment…..	1	2	3	4	5
helps students improve their learning.	□	□	□	□	□
supports teaching.	□	□	□	□	□
results are an inaccurate indicator of the university's quality.	□	□	□	□	□
results modify the ongoing instruction of students.	□	□	□	□	□
provides information to students about their strengths and areas that need to be addressed.	□	□	□	□	□
is a precise process.	□	□	□	□	□
results are used and acknowledged by instructors.	□	□	□	□	□
is not integrated with instruction.	□	□	□	□	□
is a good way to evaluate the university.	□	□	□	□	□
informs instruction to meet specific learning needs.	□	□	□	□	□
provides information on how well the university is doing.	□	□	□	□	□
is unfair to students.	□	□	□	□	□

If you would like to participate in the draw, please provide your e-mail address. The draw includes one of three Tim Hortons cards worth $15 each.

If you are interested in a summary of the results, please provide your e-mail address.

Thank you for your participation!

Appendix IV: Questionnaire form n6

Instructions

I am interested in knowing about your attitudes and beliefs towards post-secondary assessment. Please use a five-point scale (1 - Strongly Disagree…… 5 - Strongly Agree) to indicate to what extent you agree or disagree with each statement.

The term assessment means to measure and evaluate what students have learned.

	Strongly Disagree				Strongly Agree
Assessment…..	1	2	3	4	5
helps students improve their learning.	☐	☐	☐	☐	☐
supports teaching.	☐	☐	☐	☐	☐
results are an inaccurate indicator of the university's quality.	☐	☐	☐	☐	☐
results do not modify the ongoing instruction of students.	☐	☐	☐	☐	☐
provides information to students about their strengths and areas that need to be addressed.	☐	☐	☐	☐	☐
is an imprecise process.	☐	☐	☐	☐	☐
results are used and acknowledged by instructors.	☐	☐	☐	☐	☐
is not integrated with instruction.	☐	☐	☐	☐	☐
is a bad way to evaluate the university.	☐	☐	☐	☐	☐
informs instruction to meet specific learning needs.	☐	☐	☐	☐	☐
provides information on how well the university is doing.	☐	☐	☐	☐	☐
is unfair to students.	☐	☐	☐	☐	☐

If you would like to participate in the draw, please provide your e-mail address. The draw includes one of three Tim Hortons cards worth $15 each.

If you are interested in a summary of the results, please provide your e-mail address.

Thank you for your participation!

Appendix V: Questionnaire form n6w

Instructions

I am interested in knowing about your attitudes and beliefs towards post-secondary assessment. Please use a five-point scale (1 - Strongly Disagree…… 5 - Strongly Agree) to indicate to what extent you agree or disagree with each statement.

The term assessment means to measure and evaluate what students have learned.

Please pay attention to the wording of the items when you respond; some items are worded in a negative mode (e.g., *assessment results are inconsistent.*).

	Strongly Disagree				Strongly Agree
Assessment…..	1	2	3	4	5
helps students improve their learning.	☐	☐	☐	☐	☐
supports teaching.	☐	☐	☐	☐	☐
results are an inaccurate indicator of the university's quality.	☐	☐	☐	☐	☐
results do not modify the ongoing instruction of students.	☐	☐	☐	☐	☐
provides information to students about their strengths and areas that need to be addressed.	☐	☐	☐	☐	☐
is an imprecise process.	☐	☐	☐	☐	☐
results are used and acknowledged by instructors.	☐	☐	☐	☐	☐
is not integrated with instruction.	☐	☐	☐	☐	☐
is a bad way to evaluate the university.	☐	☐	☐	☐	☐
informs instruction to meet specific learning needs.	☐	☐	☐	☐	☐
provides information on how well the university is doing.	☐	☐	☐	☐	☐
is unfair to students.	☐	☐	☐	☐	☐

If you would like to participate in the draw, please provide your e-mail address. The draw includes one of three Tim Hortons cards worth $15 each.

If you are interested in a summary of the results, please provide your e-mail address.

Thank you for your participation!

Appendix VI: Factorial ANOVAs for the Factor Varying Number of Items and Provision/No Provision of a Warning.

Factorial ANOVAs for the factor varying number of items and provision/no provision of a warning.

Items	Source	Df	MS	F
1. Assessment helps students improve their learning.	number	1	0.21	0.26
	warning	1	0.11	0.14
	n x w	1	0.58	0.70
	error	261	0.83	
2. Assessment supports teaching.	number	1	0.63	0.39
	warning	1	0.70	0.10
	n x w	1	0.01	0.01
	error	261	0.68	
3. Assessment results are an inaccurate indicator of the university's quality.	number	1	0.03	0.03
	warning	1	3.71	4.22*
	n x w	1	1.86	2.19
	error	261	0.88	
4. Assessment results do not modify the ongoing Warning of students.	number	1	2.48	2.00
	warning	1	0.02	0.02
	n x w	1	0.02	0.02
	error	261	1.24	
5. Assessment provides information to students about their strengths and areas that need to be addressed.	number	1	0.11	0.08
	warning	1	0.89	0.66
	n x w	1	0.09	0.07
	error	261	1.36	
6. Assessment is an imprecise process.	number	1	2.50	2.22
	warning	1	0.26	0.23
	n x w	1	0.51	0.45
	error	261	1.13	
7. Assessment results are used and acknowledged by instructors.	number	1	0.06	0.06
	warning	1	0.25	0.26
	n x w	1	0.16	0.17
	error	261	0.95	

(continued)

Continued

8. Assessment is not integrated with instruction.	number	1	1.34	1.54
	warning	1	1.13	1.29
	n x w	1	3.30	3.78
	error	261	0.87	
9. Assessment is a bad way to evaluate the university.	number	1	3.97	4.23*
	warning	1	2.25	2.39
	n x w	1	0.08	0.09
	error	261	0.94	
10. Assessment informs instruction to meet specific learning needs.	number	1	0.01	0.01
	warning	1	2.50	2.38
	n x w	1	0.07	0.06
	error	261	1.05	
11. Assessment provides information on how well the university is doing.	number	1	0.69	0.80
	warning	1	0.16	0.18
	n x w	1	1.31	1.52
	error	261	0.86	
12. Assessment is unfair to students.	number	1	1.72	1.59
	warning	1	1.72	1.59
	n x w	1	0.00	0.00
	error	261	1.08	

Note. Items presented in bold are the negatively worded items within the second, third, fourth and fifth form. Items presented in bold and italics are additional negatively worded items within the fourth and fifth form.
** $p < .05$*