2017

Assessing Acquiescence in Surveys Using Positively and Negatively Worded Questions

Amy C. Hutton
Virginia Commonwealth University

ASSESSING ACQUIESCENCE IN SURVEYS USING POSITIVELY AND NEGATIVELY WORDED QUESTIONS

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

by

Amy Christine Hutton
Master of Fine Arts, Virginia Commonwealth University, 2008
Bachelor of Musical Arts, DePauw University, 2004

Director: Dr. James H. McMillan, Interim Associate Dean for Academic Affairs, School of Education; Professor, Department of Foundations of Education

Virginia Commonwealth University
Richmond, Virginia
February 2017

Acknowledgement

Support is key to completing any dissertation and this one is no exception. Although these acknowledgements are insufficient to truly recognize their importance in this journey, I hope these acknowledgements will provide a glimpse into my gratitude towards the family, friends, and professors who helped make this dissertation possible. I want to thank my wonderful husband, Daniel, for being my rock throughout this journey. Thank you for all the sacrifices you made, for always encouraging me to persist, and for taking such great care of our children. I truly could not have accomplished this without your support. It is an honor to have you by my side. Secondly, I would like to thank my advisor and advocate, Dr. Jim McMillan. Thank you for your constant support, guidance, and teaching. Thank you for being such an incredible role model. Third, I would like to thank my committee for the time and care they spent in helping me develop this dissertation. Your feedback and astute questions made this a much stronger study. Finally, I would like to thank my family, friends, and colleagues for making me laugh when I needed stress relief, helping take care of my family, and for just being the incredible loving people you are.

TABLE OF CONTENTS

List of Tables

List of Figures

Abstract

ASSESSING ACQUIESCENCE IN SURVEYS USING POSITIVELY AND NEGATIVELY WORDED QUESTIONS

By Amy Christine Hutton, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2017

Major Director: Dr. James H. McMillan, Ph.D.
Interim Associate Dean for Academic Affairs, Professor, Foundations of Education
School of Education

The purpose of this study was to assess the impact of acquiescence on both positively and negatively worded questions, both when unidimensionality was assumed and when it was not. To accomplish this, undergraduate student responses to a previously validated survey of student engagement were used to compare several models of acquiescence, using *a priori* goodness-of-fit statistics as evidence for model fit, in order to develop a model that adequately accounted for acquiescence bias. Using a true experimental design, undergraduate students from a variety of classes at a large, urban university were randomly assigned to one of three versions of the same

survey of student engagement (all positively worded items, all negatively worded items, an equal balance of both positively and negatively worded items). Structural equation modeling was used to analyze the results. Although the presence of acquiescence was confirmed for both positively and negatively worded items, it was not consistent by content scale or item polarization. This suggests that there may be an interaction between item polarization and content that may cause acquiescence to be present or absent. The scales that did not show acquiescence on the balanced survey portrayed a split factor loading based upon item polarization. Further, the splitting of factor loadings by item polarization was not due to acquiescence, suggesting that something other than acquiescence is causing the loadings to split. Further research is needed to develop models and/or methods to better assess and control for acquiescence. Although demographic groups were compared by gender and race/ethnicity to assess if different groups acquiesced differently, using multi-group confirmatory factor analysis, many of the models did not converge. The findings of this study were limited by the nature of the sample size. Additional research is needed to determine if acquiescence differs by group membership.

I. INTRODUCTION

Background for the Study

Unidimensionality occurs when all items in a scale support a single attribute or construct, a latent factor, in a rational manner (Nunnally, 1967; Hattie, 1985). Further, unidimensionality is best achieved when using a multiple indicator model, where each construct is defined by at least two indicators and each indicator is related to only one construct (Anderson, Gerbing, & Hunter, 1987). Survey developers commonly use both positively and negatively worded questions to measure a single construct. Two assumptions are commonly made when deciding to use an equal number of positive and negative questions, 1) both sets of questions measure the same construct, and 2) including both sets of questions increases validity (Benson & Hocevar, 1985). Yet, many studies (e.g., Herche & Engelland, 1996) demonstrate that including negatively worded questions threatens unidimensionality due to acquiescence, which causes a lack of internal consistency. If unidimensionality is not attained, then validity evidence based upon internal structure cannot be provided, thus limiting the generalizability of a study.

The *Standards for Educational and Psychological Testing* (2014), which is sponsored by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, was created "to promote sound testing practices and to provide a basis for evaluating the quality of those practices" and "to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (p. 1). Although the stated purposes of these standards apply most directly to tests, the text

acknowledges that these standards can apply to any standardized measure, such as scales and inventories. As the unidimensionality of scales is the focus of this dissertation, the standards set forth by these associations provide the foundation for this study's conceptual framework.

Validity, as defined by the *Standards of Educational and Psychological Testing* (2014), is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). One listed way evidence can be provided is through evidence based on internal structure, which is the extent to which items support the construct on which interpretations will be based. These standards specifically identify that a theory of unidimensionality would require evidence of item homogeneity, possibly in the form of a factor analysis, in order to demonstrate validity evidence based on internal structure.

Research into method biases (also known as response sets), one of the main sources of measurement error that can prevent unidimensionality, began in the 1950's (e.g., Cronbach, 1950) and continues to be an area of interest today (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). One specific type of method bias is acquiescence, where a participant agrees with positively worded statements and disagrees with negatively worded statements, regardless of their content. Acquiescence confounds the construct in questions, causing measurement error (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), by positively biasing positively worded questions and negatively biasing negatively worded questions. Since the initial research into method biases, many scholars have sought to disentangle acquiescence from the construct in question. First, scholars recommended using a balanced scale, where equal portions of positively and negatively worded questions were included to cancel out any acquiescence bias (e.g., Nunnally, 1967). However, subsequent researchers found that simply balancing a scale did not effectively eliminate acquiescence bias (e.g., Weems & Onwuegbuzie, 2001; Weems,

Onwuegbuzie, & Collins, 2006). Instead, researchers found that balanced scales caused other issues, like split factor loadings (e.g, Herche & Engelland, 1996). Therefore, using more modern statistical modeling methods, specifically structural equation modeling's confirmatory factor analysis, scholars sought to separate acquiescence as a latent construct from the content in question, from Billiet and McClendon in 2000 to Weijters, Baumgartner, and Schillewaert in 2013. While these 21st century scholars build off of each other's models for assessing acquiescence, consensus on a model that accurately separates acquiescence from content has not been found. Instead, researchers continue to explore alternative models. Since acquiescence can affect all self-report items, regardless of their polarization, finding an effective means of separating acquiescence from content is imperative for effective survey development and the accurate assessment of latent constructs.

Overview of the Literature

Acquiescence

Acquiescence is one of the primary sources of method bias, which causes measurement error and threatens validity (Podsakoff et al., 2003). It occurs when a participant responds more positively to positively worded questions and more negatively to negatively worded questions, regardless of the content of the questions. This biasing of scores can cause Type I or Type II error, resulting in erroneous conclusions. To prevent acquiescence bias, many researchers use balanced scales, with equal numbers of positively and negatively worded questions, but there is conflicting evidence on whether or not this practice is actually successful. If balancing scales with equal numbers of positively and negatively worded questions does not prevent acquiescence bias, then other means of accounting for acquiescence are needed.

Advocates for Negatively Worded Questions

Two types of advocates for negatively worded questions exist: (1) those who recommend them without reservation, and (2) those who endorse their judicious use. Those who recommend negatively worded questions without reservation (e.g., Mirowsky & Ross, 1991; Bergstrom & Lunz, 1998; Yorke, 2009), support this assertion with research demonstrating that including negatively worded questions eliminated acquiescence bias, did not affect the reliability of the scale, and maintained response integrity. However, those who recommended the judicious use of negatively worded questions (e.g., Schriesheim, Eisenbach, & Hill, 1991; Baumgartner & Steenkamp, 2001; Weijters, Geuens, & Schillewaert, 2008), determined that negatively worded questions should be included only within certain circumstances. These circumstances included when a scale was perfect balanced with equal numbers of positively and negatively worded questions (Baumgartner & Steenkamp, 2001), when negatively worded questions were not adjacent (Weijters, Gueuens, & Schillewaert), and when only certain types of negations were used (Schriescheim, Eisenbach, & Hill, 1991). The lack of agreement about the circumstances under which negatively worded questions should be used further demonstrates that acquiescence may not be controlled through the question wording itself. Other methods are necessary.

Critics of Negatively Worded Questions

Critics of negatively worded questions abound, as these questions can cause problems with agreement, misresponse, and factor loadings, even after recoding. For proper agreement within a unidimensional scale, the means should not differ significantly across items. Yet, research shows that agreement does not occur when negatively and positively worded questions are included in the same scale (Falthzik & Jolson, 1974; Chang, 1995; Cohen, Forbes & Garraway, 1996; Weems & Onwuegbuzie, 2001, 2006). Misresponse occurred more frequently on

negatively worded questions than on positively worded questions (Swain, Weathers, & Niedrich, 2008). Further, misresponse of 5% or more can significantly change the means (Hughes, 2009). Finally, split factor loadings on a theoretically unidimensional scale were commonly reported, where positively worded questions loaded on one factor and negatively worded questions on a separate factor (Herche & Engelland, 1996; Ibrahim, 2001; Magazine, Williams, & Williams, 1996). Including negatively worded questions can cause serious issues with measurement, but it is unclear whether these issues with negatively worded questions are due to the polarization of the questions or due to content issues.

Assessing Acquiescence Versus Polarization

Given the conflicting evidence regarding the use of negatively worded questions, several sets of researchers worked to assess whether these issues were due to content or acquiescence. Using structural equation modeling techniques to assess model fit, these researchers attempted to separate method effects from content. A significant acquiescence method bias for negatively worded questions was found (Motl, Conroy, & Horan, 2000; Horan, DiStefano, & Motl, 2002, 2003; DiStefano & Motl, 2006) and that bias was stable over time (Horan, DiStefano, & Motl, 2003). Although a significant positive acquiescence method effect was also found (DiStefano & Motl, 2006), it was not further explored. All of these models assumed a unidimensional content factor and did not assess the impact of acquiescence when two content factors, one positive, one negative, were present. In addition, little research has assessed whether demographic differences affect the impact of acquiescence, although research in this area is recommended (e.g., Horan, DiStefano, and Motl, 2003).

Summary

Some researchers recommend using negatively worded questions to prevent acquiescence bias, while others recommend against their inclusion, as they cause other problems. A method of statistically accounting for acquiescence is necessary to determine if the problems with negatively worded questions can be prevented, thus allowing for the effective use of both positively and negatively worded questions. Researchers, as evidenced previously, have begun developing statistical models to account for acquiescence, with some success. However, their findings lack replication. Further, there are models for accounting for acquiescence, like the ones proposed in this study, that have not been tested. It is possible that these new models can better account for acquiescence that those previously researched. Effectively preventing bias in surveys is essential for reliable and valid surveys. Providing a way to statistically account for acquiescence would allow survey developers to continue to use both positively and negatively worded questions, which provides them with additional variation in the survey design to keep respondents engaged.

Purpose for the Study

Eliminating bias from surveys is critical for validity. Yet, how acquiescence bias based upon question wording impacts surveys is not yet clear. The purpose of this study was to assess the impact of acquiescence on both positively and negatively worded questions, both when unidimensionality is assumed and when it is not. To accomplish this, undergraduate student responses to a previously validated survey of student engagement were used to compare several models of acquiescence, using *a priori* goodness-of-fit statistics (Kline, 2016) as evidence for model fit, in order to develop a model that adequately accounted for acquiescence bias.

Research Questions

The findings from this study sought to answer the following questions.

1) When using a balanced survey, can statistical modeling of acquiescence allow for the unidimensional scaling of content?
2) Does demographic group assignment impact the modeling of acquiescence?
3) Does acquiescence differ by survey type (all positive, all negative, or balanced)?

Design and Methods

This research study employed a true experimental design, where students in each class were randomly assigned to one of the three survey types (all positive, all negative, or balanced). The Burch Engagement Survey for Students (BESS; Burch, Heller, Burch, Freed, & Steed, 2015) included four scales of six questions each: emotional engagement, physical engagement, cognitive engagement: in-class, and cognitive engagement: out-of-class. From the BESS survey, which includes only positively worded questions, two alternate versions were developed, one balanced, and one negative. Demographic questions assessing age, race/ethnicity, and gender were included on all three survey forms. The survey was administered to a convenience sample of undergraduate classes at a large public university.

Three different analyses were performed to assess acquiescence. The goal of the first analysis was to find a statistical model that successfully differentiated between acquiescence and content for the balanced survey design across all respondents who responded to that survey. Several different statistical models were compared using confirmatory factor analysis and *a priori* goodness-of-fit statistics to determine which model had the best fit. The second step was to split the respondents into demographic groupings (minority / majority; male / female). The

same models from the first analysis were compared across demographic groups to see if participants responded differently to acquiescence, based upon group status. Finally, acquiescence was compared across all three survey types (positive, negative, and balanced) to see how acquiescence differed by survey design.

Definition of Terms

Negatively worded item – A negatively worded item is one that is opposite of a positively worded item. For this study, negated regular (typically including the word "not") will be used.

Acquiescence – Acquiescence is a respondent's proclivity to respond positively to positively worded questions and negatively to negatively worded questions, regardless of their content (Weijters, Baumgartner, and Schillewaert (2013).

Unidimensionality – Unidimensionality occurs when all items in a scale support a single attribute or construct, a latent factor, in a rational manner (Nunnally, 1967; Hattie, 1985).

Latent factor – A latent factor is a variable that cannot be directly measured, but is assumed to be related to observed variables that can be measured (Field, 2009).

Content latent factor – The content latent factor is a variable that, for this study, is the latent factor for the student engagement scale in question.

Acquiescence latent factor – The acquiescence latent factor is a variable that, for this study, is the latent factor that represents acquiescence.

II. REVIEW OF LITERATURE

Overview

Developing a statistical model to examine acquiescence in positively and negatively worded questions required an extensive understanding of prior models of analysis. This prior research is organized in several sections. First, acquiescence biases are defined and common assumptions surrounding those biases are presented. Secondly, research that advocates for the inclusion of negatively worded questions is explored. Support for negatively worded questions is then followed by critics who oppose using negatively worded questions and the challenges that arise, including poor agreement, misresponse, and split factor loadings. Finally, a chronological exploration of prior models assessing acquiescence versus content in surveys provides the foundation for the methodology and proposed models for this study. Several methods were used to retrieve these prior studies, including Academic Search Complete, Business Source Complete, Education Research Complete, ERIC via ProQuest, Google Scholar, and VCU Libraries Database. Search terms included "acquiescence," "reverse coding," and "unidimensionality." The most fruitful source of prior studies was the references of pertinent articles and Google Scholar's feature showing later works that cited that article. This linkage of articles was especially useful in following the articles presented in Structural Equation Modeling: A Multidisciplinary Journal. It was this journal's history of articles that led to the development of the models used for this study.

Acquiescence Biases

Method bias occurs when variance in a measure comes from the measurement method, rather than the construct the measure represents. Method biases are one of the primary sources of measurement error which threaten the validity of conclusions (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). In their review of prior studies, Podsakoff et al. (2003) found that method bias not only varies in strength, but can inflate or deflate the observed relationships between constructs, causing Type I or Type II error. Type I error occurs when a researcher rejects the null hypothesis, when it is actually true. Type II error occurs when the research fails to reject the null hypothesis, when it should be rejected (McMillan, 2012).

Podsakoff et al. (2003) cite acquiescence as one of the sources of method bias, as it can cause spurious relationships between constructs and cause artificial variance. As defined by Weijters, Baumgartner, and Schillewaert (2013), acquiescence is a respondent's proclivity to respond positively to positively worded questions and negatively to negatively worded questions, regardless of their content. Acquiescence generalizes across items within a given scale (Heaven, 1983). Acquiescence can also disguise the real relationships between items by falsely increasing the correlations among items with the same polarization (Winkler, Kanouse, & Ware, Jr, 1982). This, in turn, can cause factor analyses to show separate method factors based on polarization, rather than content (Winkler, Kanouse, & Ware, Jr, 1982).

A common assumption is that using balanced scales, where half of the items are worded positively and half negatively cancels out any acquiescence bias (Ferrando, Lorenzo-Seva, & Chico, 2003), yet researchers are divided on whether or not that assumption is valid. In social research, some survey development and measurement texts recommend balancing the number of positively and negatively worded items to prevent acquiescence (e.g., Mitchell & Jolley, 2013;

de Vaus, 2014), whereas others recommend against the use of negatively worded items (e.g., Wright & Masters, 1982; Nardi, 2006). However, there is a stark division between researchers who advocate for the use of mixed scales and those who warn against the use of negatively worded questions.

Advocates for Negatively Worded Questions

One of the most heavily cited proponents of using negatively worded items is Nunnally (1967). Nunnally states that acquiescence can be eliminated by including a balance of positively and negatively worded items. Many survey developers cite him as a reason to use balanced surveys (A Google Scholar search on August 3rd, 2016 showed that 84,853 publications have cited his book). Although Nunnally recommends balanced scales, he does not provide empirical support for this assertion. Other researchers, however, provide evidence that further supports Nunnally's assertion.

Several studies, across a variety of content areas, determined that including negatively worded questions did not impact the unidimensionality of scales. Marsh (1986) found that negatively and positively worded subscales did not need to be separated into separate factors and that differential weighting produced little or no improvement in reliability or internal consistency. Mirowsky and Ross (1991) determined that using a balanced scale eliminated acquiescence bias. Therefore, they advocated for the use of a balanced scale as a way of canceling out positive versus negative bias.

In a study of job satisfaction, Bergstrom and Lunz (1998) found using item response theory that positively and negatively worded questions appeared to be measuring the same construct. Polarized questions correlated highly (.77) and had the same reliability (.87), when

scaled together versus separately. Bergstrom and Lunz determined that participants responded similarly on both types of questions.

A study by Yorke (2009) compared responses to three different surveys, one that was all positive, one that was mostly positive, and one that was mostly negative. The study demonstrated that response patterns did not differ when negatively worded questions were included. Further, response patterns did not change when negative statements were placed early in the survey, as compared to later. The lack of significant findings demonstrated that negatively worded questions could be included without compromising the integrity of the responses.

Based on this research, utilizing negatively worded questions was recommended, without reservation. However, there are many researchers who strongly advocate for the use of negatively worded questions, but only when those questions are used judiciously and under certain circumstances (e.g., Schriesheim, Eisenbach, & Hill, 1991; Baumgartner & Steenkamp, 2001; Weijters, Geuens, & Schillewaert, 2008). Baumgartner and Steenkamp (2001) studied the effect of negatively worded questions across eleven countries in the European Union and found that unbalanced scales with negatively worded questions can bias scores. However, they found that balancing positive and negative questions successfully counteracts acquiescence. The authors strongly advocated for using balanced scales in research.

Weijters, Geuens, and Schillewaert (2008) strongly recommend utilizing negatively worded items, assuming they are used wisely. Their study found that negatively worded items are a cue to respondents to retrieve new information, versus using information previously recalled for use on other survey items. If these reversed items were next to non-reversed items that measured the same construct, respondents would retrieve new information to answer the reversed item and, in turn, produce different results from the positively worded items in the

construct. If negatively worded items measuring the same construct were adjacent, the two items would show zero correlation, as the participants would retrieve new information to answer the second negatively worded item. However, if the items were not grouped together by content, but were instead dispersed throughout the survey (with 0 to 6 questions in-between items measuring the same construct), respondents were less likely to use previously retrieved information to answer the questions, resulting in stronger correlations between items measuring the same construct. Therefore, Weijters, Gueuns, and Schillewaert recommend using reversed items and dispersing them across the questionnaire. However, the researchers recognize that using these negatively worded questions may result in lower factor loadings and composite reliabilities for confirmatory factor analyses. The authors advocate for using Billiet and McClendon's (2000) response style factor model, which improves the model fit for negatively worded items. A major limitation of this study is that participants were from Belgium, so generalizability to other countries is unknown. The authors recommend further research, as other cultures may respond to negatively worded questions differently.

Weijters, Geuens, and Schillewaert (2008) strongly advocate for Billiet and McClendon's (2000) model for modeling acquiescence in a balanced scale, which will be discussed in more detail in a later section. Yet, Billiet and McClendon do not make a recommendation for or against using negatively worded questions. They simply provide a method for assessing the acquiescence for a balanced scale. A major limitation of Billiet and McClendon's study is that they claim to use a balanced scale, yet admit that their negatively worded items are not exact negations of the positively worded questions. Without exact negations, they allowed their respondents to potentially agree with both positively and negatively worded questions. This limitation in their study limits generalizability. However, their method of assessing acquiescence

provides some of the foundation for the methods of this study and will be addressed more in-depth in chapter 3.

Schriesheim, Eisenbach, and Hill (1991) suggest that all types of negatively worded questions (polar opposite, negated polar opposite, and negated regular) may not cause issues with reliability and validity. Examples of these types of questions are included in

Table 1, adapted from Dr. Seuss (1960).

Table 1. Examples of regular positively and negatively worded statements.

Question Type	Example
Regular (Positively worded)	I like green eggs and ham.
Negated regular	I do not like green eggs and ham.
Polar opposite	I loathe green eggs and ham.
Negated polar opposite	I do not loathe green eggs and ham.

Schriesheim, Eisenbach, and Hill found that polar opposite and negated polar opposite questions should be avoided; however, they implored that negated regular items did not cause serious enough problems with reliability and validity to keep them from being used in survey design, even when regular items proved to be the most reliable. This finding is in direct contrast to Schriesheim and Hill's (1981) previous study that used a significant amount of negated regular items and found that including negatively worded items resulted in less accurate responses and impaired validity.

Ray (1979; 1983) and Weijters, Baumgartner, and Schillewaert (2013) recommend including both positively and negatively worded questions, as bias in a bipolar scale can be identified and assessed. The researchers argue that acquiescence may still be present in scales that are worded entirely in one direction, but that any bias is confounded with the content variance.

The inclusion of negatively worded questions has strong support from these researchers. Whereas some researchers recommend them without reservation, others believe they should be used judiciously. Although many critics of negatively worded questions exist, the amount of proponents that continue to arise in recent scholarship demonstrates that further research is needed to determine whether or not negatively worded questions should be used in surveys and how those questions are impacted by acquiescence.

Critics of Negatively Worded Questions

Although some studies have found that negatively and positively worded items can successfully create a unidimensional scale, many other studies show that negatively worded questions do more harm than good. Although all of these studies demonstrate measurement bias due to the presence of negatively worded questions, the studies can be split into three categories: agreement, misresponse, and split factor loadings.

Agreement

Many studies demonstrate a lack of agreement between positively and negatively worded items that measure the same construct, where the means differed significantly between the positively and negatively worded items (after recoding; e.g., Falthzik & Jolson, 1974; Chang, 1995; and Cohen, Forbes, & Garraway, 1996). In a unidimensional scale, the means should not differ significantly across items with the same scale.

Falthzik and Jolson (1974) found that for 7 out of 12 survey questions, the means for positively worded questions were significantly higher than for the same question when worded as a straight negation (i.e., using the word "not" as the reversal mechanism). The other five questions did not show statistically significant differences between the positively and negatively

worded versions. The researchers posit that consumerists could make more convincing arguments one way or the other depending if the appropriate question polarization was used, as the item direction alone could bias the results.

Chang (1995) demonstrated that reversed items are not necessarily fully exchangeable with regular items and recommended eliminating reversed items. However, it is unclear what kind of negatively worded questions were used in this study. If the items were not true negations (taking the regular item and adding "not" or "do not" while retaining the rest of the item wording), it is possible that these results were confounded. Further, this study used a test-retest method with a one-week interval between the two administrations, so memory and/or boredom effects could also have confounded the results.

A study about healthcare satisfaction in Scotland by Cohen, Forbes, and Garraway (1996) revealed that negatively worded questions differed dramatically from their positively worded counterparts. For example, on one question, responses differed by over 18%, with only 5.6% of respondents agreeing with the negatively worded question, when 23.9% disagreed with the positively worded question. The authors of this study state that degree of healthcare satisfaction is sensitive to changes in worded, specifically the polarization of the questions.

Two studies by Weems and Onwuegbuzie (2001) showed that the means of positive items were significantly higher than those of negatively worded items that supposedly measured the same construct. The authors attributed this difference to positively and negatively worded items not necessarily measuring the same constructs and that using mixed stems may reduce score reliability. However, this study was unable to control for differences in content between the positively and negatively worded items, requiring that additional research be done.

Weems, Onwuegbuzie, and Collins (2006) found mixed format scales to be problematic for graduate students. Although they measured the same construct, the researchers showed that responses to positively worded items were typically higher than responses to negatively worded items. Further, the authors recommend that instrument developers disaggregate scale and subscale scores by polarization, so that any differences between the positively and negatively worded questions can be explained. The authors suggest that this type of analysis be published alongside normative and psychometric data in instrument manuals.

Misresponse

After completing an exploratory meta-analysis to confirm that misresponse was an issue, Swain, Weathers, and Niedrich (2008) studied whether misresponse to negatively worded questions was due to the respondent acquiescence, respondent inattention, or issues with the negatively worded questions themselves. Misresponse occurs when respondents answer an item differently from the rest of the items in a unidimensional scale. The first experiment of the study compared misresponse and response latencies for four categories: true affirmations, false affirmations, false negations, and true negations. The researchers found that respondent misresponse was the lowest, by a substantial margin, for true affirmative items (.81%), whereas false affirmations (5.65%) and false negations (8.40%) has much higher rates of misresponse. The highest rate of misresponse was for true negations at 19.83%. Further, they found that latencies increased along with misresponse, with higher misresponse increasing linearly with latency. This experiment showed that item verification difficulty is a problem with negatively worded statements.

Given that Swain, Weathers, and Niedrich's (2008) first experiment used items with different content to compare the misresponse of positive versus negatively worded items, they

did a second experiment using items with the same content. Respondents were randomly assigned to a survey where each scale had one negated item or to a survey with only positively worded items. As with the first experiment, misresponse was significantly more of an issue for negated items (19.37%) than for those items when stated affirmatively (4.34%). However, this experiment showed that misresponse was not due to item content. This second experiment also confirmed the pattern of misresponse, with true affirmations being the lowest (3.74%), followed by false affirmations (9.84%), false negations (26.47%), and true negations (39.02%). Response latencies also followed this pattern for increase. This experiment also showed that acquiescence and inattention were minimal, so misresponse can happen without acquiescence or inattention and is more likely to happen when using negatively worded questions or items to which the respondent should disagree.

In a final experiment by Swain, Weathers, and Niedrich (2008), the researchers assessed the impact of three different types of negation (particle, affixal, and implicit), as compared to an affirmation as the control. All respondents received the same survey with four scales that each contained four questions. Three of the questions in each scale were negative, containing one affixal, one particle, and one implicit negation. The fourth question for each scale was an affirmation. In this experiment, the researchers found that misresponse lowest for affirmations (8.12%), followed by affixal negations (20.70%), particle negations (20.90%), and implicit negations (27.71%). Further, they found that inattention and acquiescence were poor predictors of misresponse. These three experiments demonstrate that misresponse can be a major issue with negatively worded questions; however, this may not be due to acquiescence or inattention.

Hughes (2009) did a simulation study of the impact if negatively worded questions are misinterpreted. For instance, a respondent might not realize the question was negatively worded

and instead respond as if it were positively worded. Hughes looked at the impact of the percentage of negatively worded questions crossed with a percentage of incorrect responses to those questions. The results showed that for scales with more than one negatively worded item and incorrect response rates of 5% or more, the scale means can be significantly different. This study showed the effect size could be as large as 0.64 standard deviations. Given that such a small percent of incorrect responses could shift the mean so dramatically, a small amount of misinterpretation could result in Type I or Type II error. Further, an exploratory meta-analysis of previous research by Swain, Weathers, and Niedrich (2008) showed that misresponse to negatively worded questions averaged 17.50%, which is much greater than the 5% found to be an issue by Hughes (2009).

Wong, Rindfleisch, and Burroughs (2003), in a cross cultural study, found that scales containing both positively and negatively worded questions were interpreted differently by people from different cultures. The researchers had difficulty obtaining cross-cultural measurement equivalence without controlling for the question polarization. However, the authors state that this could be due to translation errors, acquiescence, or cultural differences.

Split Factor Loadings

Using principal components analysis with varimax rotation, a study by Herche and Engelland (1996) used four different previously validated surveys and found that for each of them, items within the same scale loaded separately on two separate orthogonal factors by polarity. The negatively worded items loaded on one factor, while the positively worded ones loaded on another, instead of having them all load onto a single factor, as designed. Confirmatory factor analysis further demonstrated for all of the scales in all of the surveys,

splitting the scales into two factors based upon polarity demonstrated significantly better fit than maintaining single factor structures.

Magazine, Williams, and Williams (1996) conducted a study examining two different organizational behavior scales that contained both negatively and positively worded questions. The first scale, the Meyer and Allen Affective Commitment Scale, contained four negatively worded questions and four positively worded questions. The second scale, the Meyer and Allen Continuance Commitment Scale contained two negatively and six positively worded questions. For both scales, confirmatory factor analyses demonstrated stronger fit when the negatively worded questions were loaded onto a separate factor from the positively worded questions.

A study of college students by Ibrahim (2001) that included 17 positively worded items and 1 negatively worded item, all purportedly measuring the same construct, found that the negative item loaded separately from the positively worded items in an exploratory factor analysis. Further, after reverse scoring the negative item, it correlated negatively with 16 out of the 17 positively worded items.

Many other studies further demonstrate issues with factor loadings and reliability. Merritt (2012) varied which items in a scale were negative and found that the presence of negative items consistently created a second factor for those items. A study by Barnette (2000) showed that phrasing all questions positively demonstrated greater reliability than using mixed stems. Barnette recommended varying the direction of the Likert scale responses, as it demonstrated much stronger internal consistency than including negatively worded questions. Many other studies demonstrate that negatively worded items have poor reliability (e.g., Melnick & Gable, 1990; Harasym, Price, Brant, Violato, & Lorscheider, 1992; Kunda & Fong, 1993; Lam, 1995; McPherson & Mohr, 2005).

Summary

As evidenced by this research, major challenges can arise when using negatively worded questions in surveys, including issues with agreement, misresponse, and factor loadings. Yet, the conflicting advice between these critics and the proponents discussed previously creates a conundrum for survey developers. Should negatively worded questions be included or avoided? To address this issue, researchers have started trying to separate acquiescence from content, to see if reliable, unidimensional scales using negatively worded questions can be attained.

Assessing Acquiescence Versus Polarization

Given the lack of unidimensionality in bipolar scales, several researchers sought to assess whether the lack of unidimensionality was due to content issues or acquiescence bias. An early method of assessing acquiescence was developed by Winkler, Kanouse, and Ware, Jr (1982), called the Aquiescence Response Set (ARS) score. After collecting survey results, the researchers ran an exploratory factor analysis and found that the negatively and positively worded items loaded on separate factors, regardless of content. To assess the impact of acquiescence, the researchers counted how many times a respondent agreed with contradictory matched statements (one positive, one negative). Without acquiescence, a respondent should agree with one item and disagree with the other. If acquiescence was present, the respondent would agree or disagree with both items. The total number of agreements/disagreements with contradictory statements served as the respondent's ARS score, with a possible range of 0 to 12, based on a survey with twelve matched pairs, 5% of respondents scored a four or higher. The researchers then created a zero-order correlation matrix that excluded the 5% who scored highest on the ARS and a first-order partial correlation matrix controlling for ARS score. They found

that using the 5% exclusion method did not greatly change the factor loading structure. However, controlling for ARS score through a partial correlation matrix caused the exploratory factor analysis to generate unidimensional factors that included both positively and negatively worded items.

In a study by Marsh (1996), he found using confirmatory factor analysis that a single latent factor existed that included both positively and negatively worded items. However, he found method effects associated primarily with the negatively worded items, making interpretation difficult. In Marsh's model, he correlated the errors of the negatively worded questions, but did not consider a separate latent method effect.

Another method of controlling for acquiescence was to consider it a latent factor using a structural equation model. Billiet and McClendon (2000) developed a method for assessing acquiescence using structural equation modeling, where they compared three hypothetical models using a balanced scale with equal numbers of positively and negatively worded questions. As mentioned previously, their negatively worded questions were not negations of the positively worded questions, which was a limitation of their study that allowed for participants to potentially agree with both positively and negatively worded questions. Although this issue limited the generalizability of their results, their method of assessing acquiescence had merit and was utilized in this study.

The first of Billiet and McClendon's (2000) three models assessed a latent factor for each construct, including both positively and negatively worded questions in each latent factor, as seen in Figure 1. These factors were allowed to covary freely. The second model, seen in Figure 2, added a latent factor for acquiescence, using all indicators for all factors from the first model with lambdas constrained to 1, as acquiescence was theorized to be consistent across all items.

The final model included separate acquiescence latent factors for each of the construct factors, as seen in Figure 3. Two versions of this model were included, one where the acquiescence factors were allowed to covary and one where they were not allowed to covary.

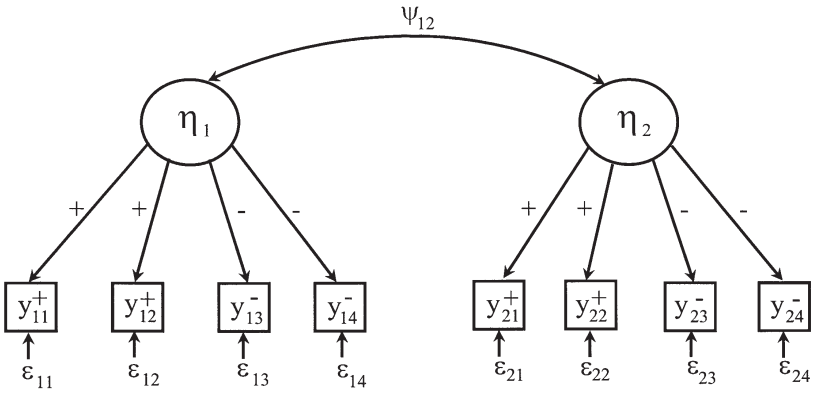


Figure 1. Billiet and McClendon's (2000) model without acquiescence (p. 613)

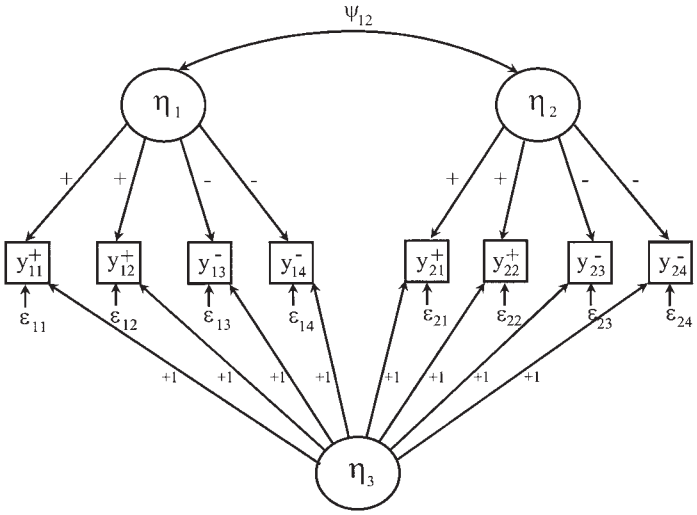


Figure 2. Billiet and McClendon's (2000, p. 613) single acquiescence factor model (p. 613)

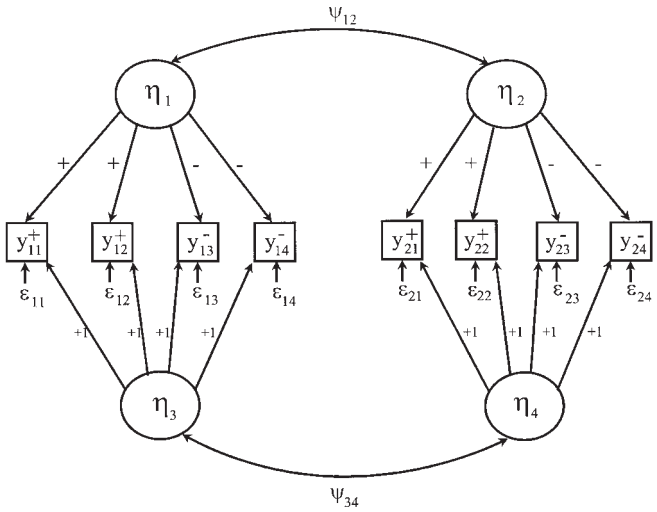


Figure 3. Billiet and McClendon's (2000) two acquiescence factor model (p. 614)

A final model by Billiet and McClendon (2000) added predictors to the model, including education and age, as indicators. This model found that respondents with more education were less susceptible to acquiescence. It also found that older respondents were more likely to acquiesce. Given that the sample for this study is homogenous, in terms of age and education, this model will not be included in the study.

A major assumption of Billiet and McClendon (2000) was that the positively and negatively worded questions for each construct loaded onto a single factor. The reason for this assumption was not addressed. Given that prior research demonstrates that negatively worded questions tend to load onto a separate factor from the positively worded questions that assess the same construct, this omission is surprising.

A subsequent study by Cambre, Welkenhuysen-Gybels, and Billiet (2002) took the Billiet and McClendon (2000) study a step further, by analyzing whether the positively and negatively worded questions of a construct fit better into a two factor model, with one factor for each

polarization, or a single factor model including all questions and a separate latent factor for acquiescence. These models can be seen in Figure 4.

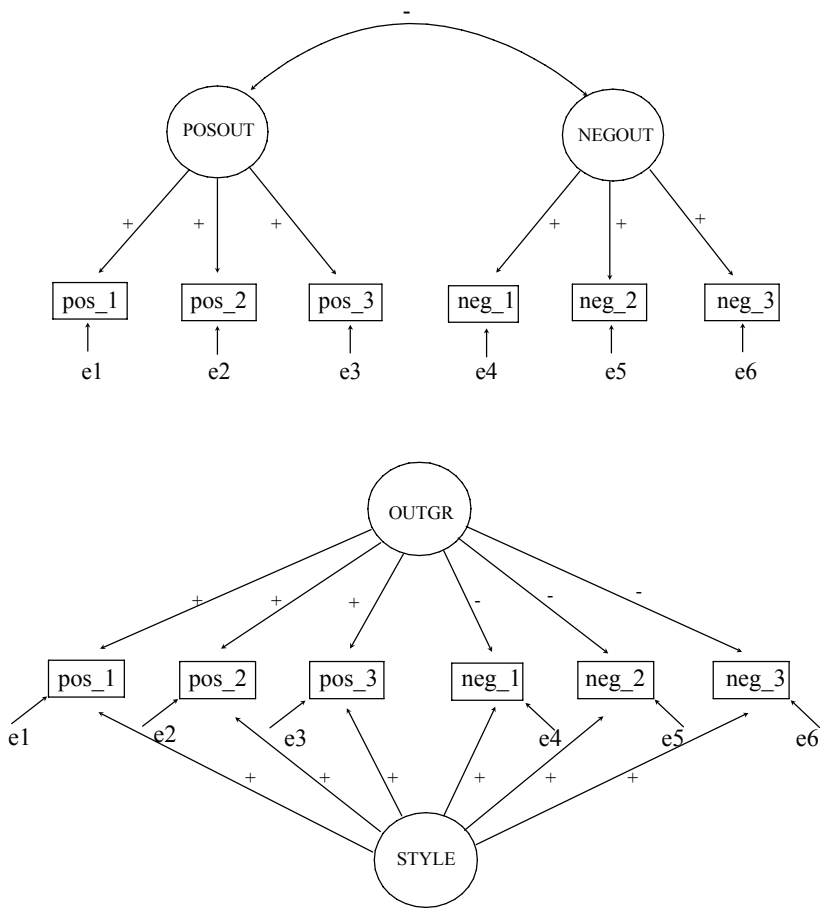


Figure 4. Cambre, Welkenhuysen-Gybels, and Billiet (2002) models (p. 3)

Unlike with the Billiet and McClendon (2000) study, the style latent variable was not considered for the two factor structure, fitting positively and negatively worded questions to separate factors. The style factor was only considered in a setting where the latent factor was unidimensional. Cambre, Welkenhuysen-Gybels, and Billiet (2002) compared these models across many different Western European countries and found that some countries fit the model with the style factor fit best, whereas other countries fit the two factor structure best. The United States was not included in this study, so it is unknown which model would best represent the

United States or if a third model, one that includes a style factor along with separate factors based upon polarization, would be a better fit.

A second set of researchers, led by Robert Motl, Christine DiStefano, and Patrick Horan, have also used confirmatory factor analyses in several studies to assess the role of acquiescence in balanced scales. In the Motl, Conroy, and Horan (2000) study, they compared nine different models, as seen in Figure 5.

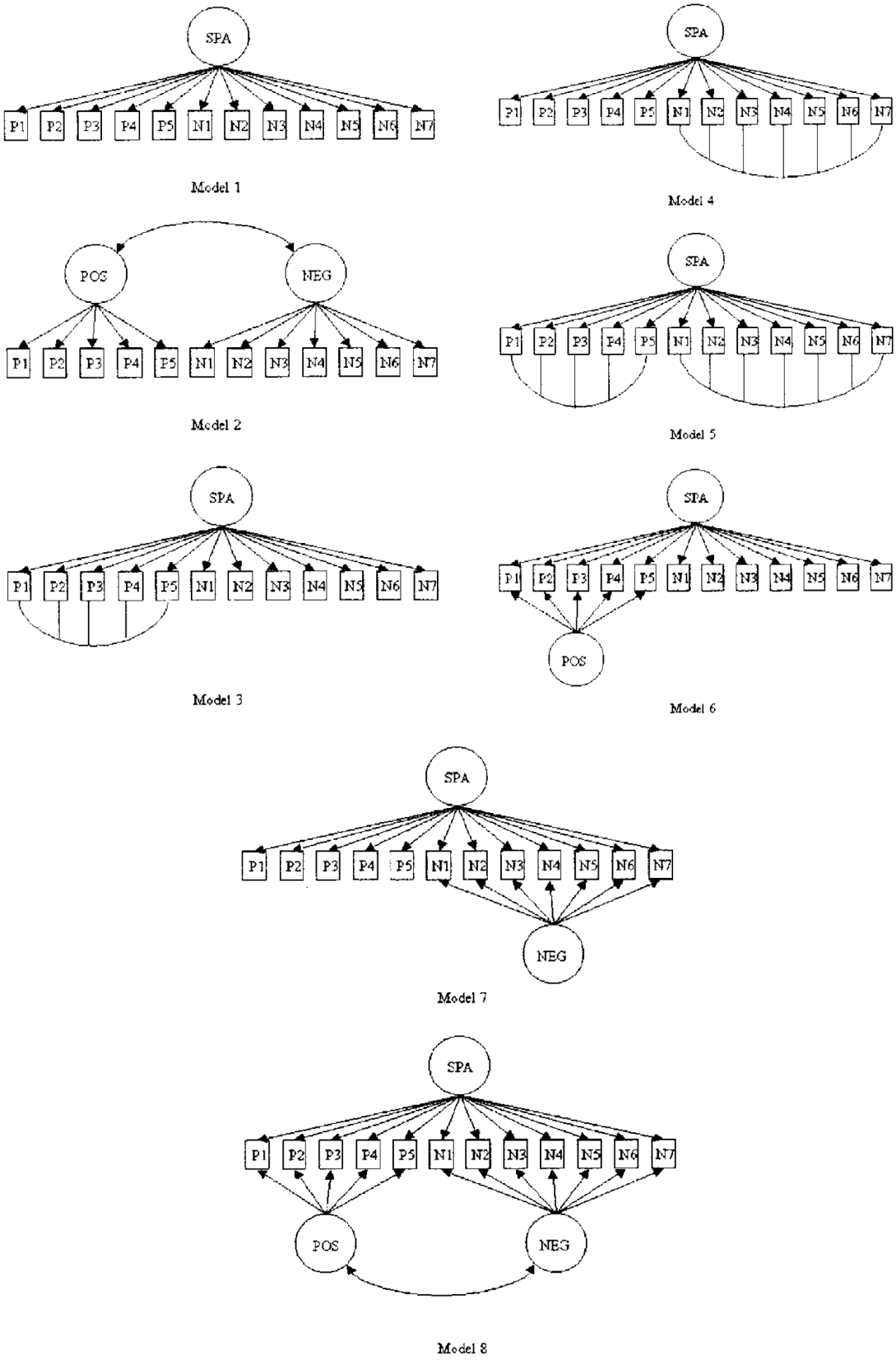


Figure 5. Motl, Conroy, and Horan (2000) study models (p. 337)

Their study revealed that models 4 and 7, which modeled method effects across negatively worded items, had the best fit. The model accounting for both positive and negative method effects did not fit much better than the models just accounting for the negative method effect. Therefore, in their subsequent studies, the model accounting for both positive and negative method effects was not utilized.

Similarly, Horan, DiStefano, and Motl (2002, 2003) evaluated different models to evaluate the method effects of positively and negatively worded items. Motl and DiStefano (2002) asked participants to take the same self-esteem survey three times over two years. They found that method effects associated with negatively worded items demonstrated longitudinal invariance and were, therefore, stable over time. In Horan, DiStefano, and Motl's 2003 study, they used a single content factor for self-esteem, a two-factor model separating the positive and negative items into separate factors, then several models exploring latent wording effects and correlated errors by item polarity. They found that models, seen in Figure 6, that had a single content factor and showed latent wording effects (1c) or correlated errors (1e) for negatively worded items had the best fit.

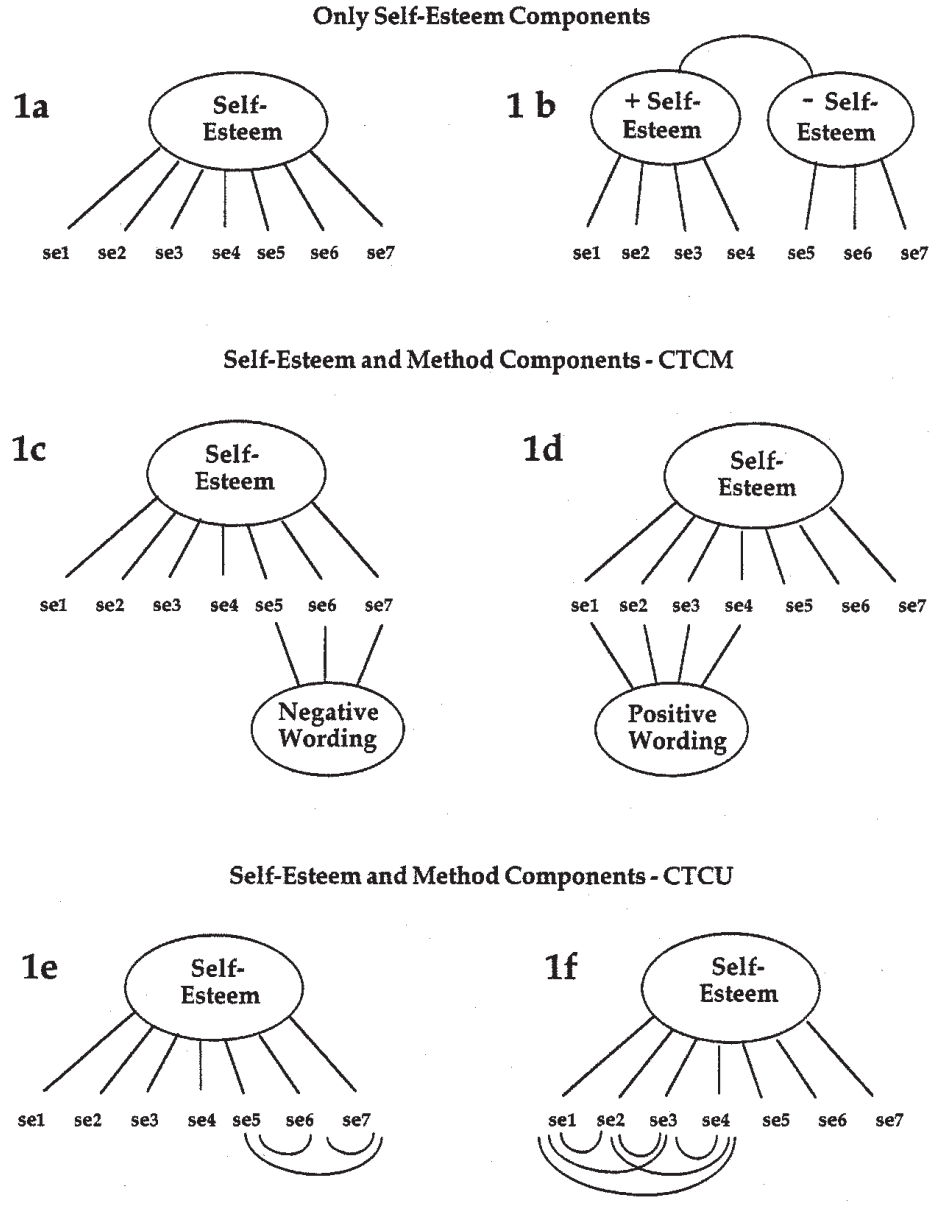


Figure 6. Horan, DiStefano, and Motl (2003) models (p. 442)

Horan, DiStefano, and Motl (2003) also assessed the presence of negative wording effects in other scales. They found that negative wording effects were consistent across the other scales they analyzed. Finally, the researchers assessed whether the negative wording effects were correlated across content areas or if they were independent. Three models were compared: one with three scales and no wording effects, one with three scales and negative wording effects for

each scale, and one with three scales and a single negative wording factor associated with the negative items for all three scales. The second model, seen in Figure 7, with individual negative wording effects that were correlated, had the best fit.

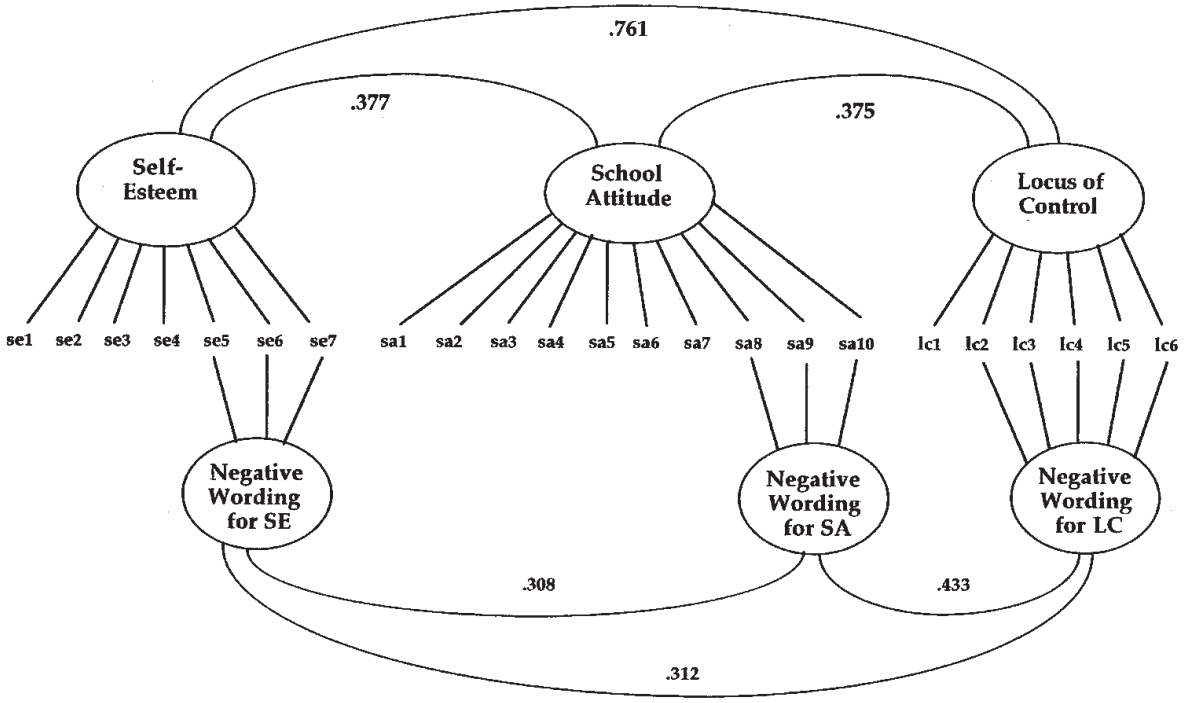


Figure 7. Horan, DiStefano, and Motl (2003) model of three independent scales with correlated negative wording effects (p. 448)

Finally, Horan, DiStefano, and Motl (2003) assessed whether wording effects were stable over time, using the model seen in Figure 8. Through a longitudinal analysis over two years, they found that negative wording effects were stable over time.

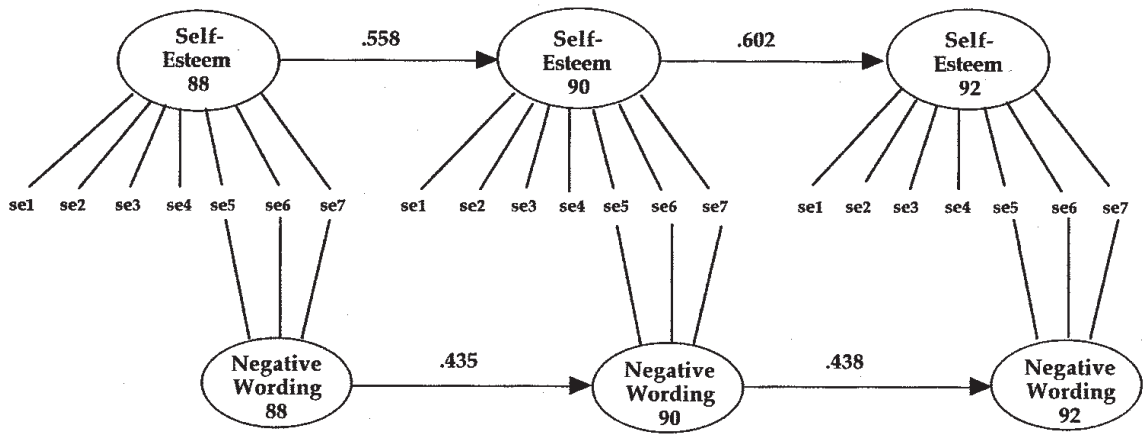


Figure 8. Horan, DiStefano, and Motl (2003) model of negative wording effect stability over time (p. 450)

In a subsequent study, DiStefano and Motl (2006) attempted to distinguish between content and wording effects in self-report scales. They compared many different models, including a single latent content factor model, a two latent factor content model with negative and positive questions loaded to separate factors, a single content factor with a method effect for the negatively worded items, and a single content factor with a method effect for the positively worded items, as seen in Figure 9 and Figure 10.

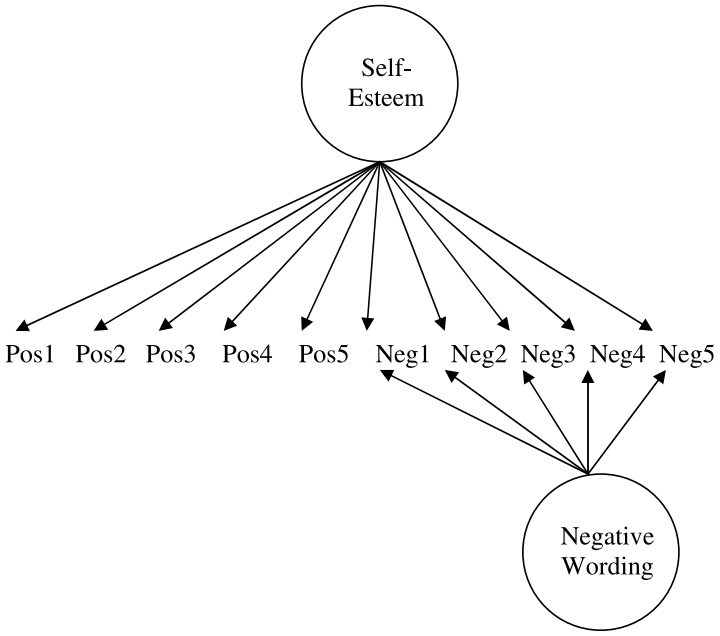


Figure 9. DiStefano and Motl (2006; 2009) single latent construct and negative latent method effect model (p. 449 & 137, respectively)

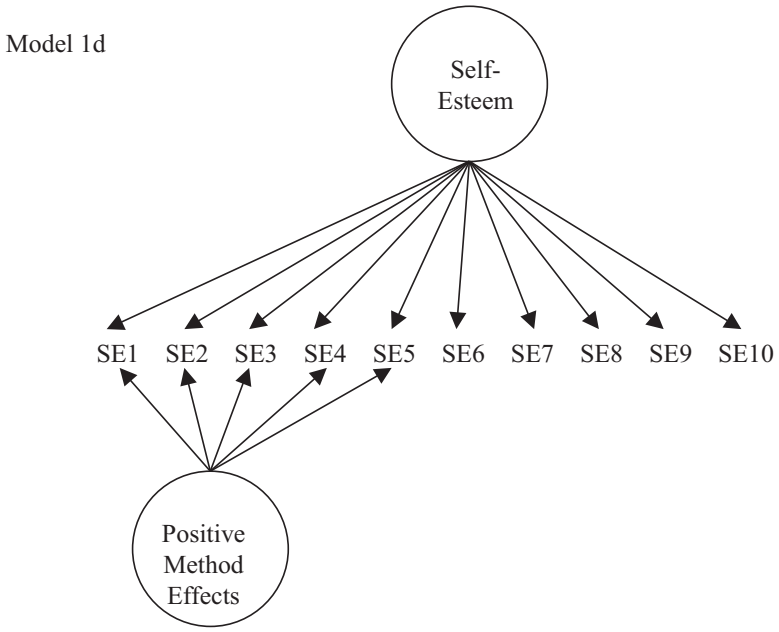


Figure 10. DiStefano and Motl (2006) single latent construct and positive latent method effect model (p. 450)

As with other researchers, DiStefano and Motl showed that models accounting for wording effects improved the fit over models without wording effects. They found that the model accounting for the negative method effects had the best fit of all the models, as the method effect remaining consistent for the negatively worded items, even when evaluating multiple latent content constructs. However, they did find a positive method effect model to also have good fit. Although DiStefano and Motl evaluated the fit with a positive method effect and a negative method effect, they did not combine these method effects into a single model to see if both positive and negative method effects existed simultaneously.

In their 2009 follow-up study, DiStefano and Motl continued to assess whether negatively worded items contained a latent method effect across varying genders, using the same negative method effect model from their 2006 study, seen in Figure 9. The researchers did not state why a method effect for positively worded questions was not considered, given the strength of the literature surrounding acquiescence on both positively and negatively worded questions.

Weijters, Baumgartner, and Schillewaert (2013) attempted to assess the impact of acquiescence on negatively worded items, but did not use negations. The authors suggest that future research be done to address the impact of acquiescence on negations.

Using structural equation modeling, these researchers have begun to separate acquiescence from content in order to allow for unidimensional scales that include negatively and positively worded questions. Their research shows that negatively worded questions included a distinct method effect; however, this method effect has only been assessed when assuming a unidimensional scale. And, the method effect of positively worded questions is left out of most models. More research is needed to compare the successful models of these

researchers to ones that include both positive and negative method effects, without assuming content unidimensionality.

Demographic Differences in Acquiescence

Researchers have mixed views on whether acquiescence differs by demographic group. Falthzik and Jolson (1974) did not find any relationship between acquiescence response bias and the demographic variables of gender, age, and race. DiStefano and Motl (2009) investigated whether a method effect of acquiescence on negatively worded items differed by gender, but did not find a difference. However, they suggest the need for future research testing invariance by racial characteristics associated with acquiescence on negatively worded items.

A study by Bachman and O'Malley (1984) found that blacks were more likely than whites to acquiescence; however, the authors did not focus on this. Instead, they focused on how blacks were more likely to use extreme responses, as there was a greater difference between blacks and whites on that measure than on the measure of acquiescence. The presence of this difference between blacks and whites on acquiescence warrants further research and discussion than these authors provided.

Alessandri et al. (2010), in their twin study, argue that acquiescence to positively worded items is a personality characteristic that might be inherited. The researchers posit that acquiescence may be a stable characteristic, versus something specifically related to item phrasing. The authors recommend that, based upon their preliminary findings of potential heritability, research be done assessing method effects in different populations and cultures. Given the potential heritability of acquiescence, it is plausible that acquiescence may differ based upon different demographic characteristics.

Research assessing the impact of demographics on acquiescence is limited. Horan, DiStefano, and Motl (2003) assert that further research is necessary to understand if different populations respond differently to positively and negatively worded questions.

Implications

There is conflicting advice amongst past and recent researchers as to whether or not including negatively worded questions prevents acquiescence or simply creates more problems. Further, models that attempt to disentangle acquiescence from content are in the early stages of development. Although these models clearly identify a negative method effect that can be modeled latently, they assume content unidimensionality and overlook positive method effects. Finally, little research assesses the impact of demographics on acquiescence. More research is necessary to see if (1) modeling positive and negative method effects simultaneously improves model fit, (2) removing the assumption of content unidimensionality changes model fit, and (3) demographic group assignment dictates model fit. It is these areas that this study seeks to address.

III. METHODOLOGY

Design

This study utilized a true experimental design with three intervention levels. Students within each class were randomly assigned to one of the three surveys. This randomization controlled for within-class group effects, as it is likely that engagement differed by class and instructor. Students did not see the surveys other students took, as the surveys were collected as soon as the students completed them. The students also did not have an opportunity to discuss the questions with their peers, as silence was maintained during administration. These precautions were intended to prevent contamination of the results.

Population

The population of interest for this study was undergraduate students. The sample consisted of undergraduate students at a large, public, mid-Atlantic university. A convenience sample of undergraduate classes was selected for conducting the survey, based upon the willingness of the course professors to participate. Professors who taught classes with enrollments above 30 students and with whom the researcher had prior relationships were contacted and asked to participate. All professors, except one, agreed to participate. Further, department chairs who oversaw core classes with large enrollments were contacted. Their assistance was requested in opening communication with the professors of the large classes. All

of the department chairs contacted agreed to assist. Each of the professors identified by the department chair agreed to allow the researcher to use their classes to recruit participants.

All students present in these classes were asked to participate, but were not required to do so. All participants in each class were randomly assigned to one of the three surveys. A power analysis (Preacher & Coffman, 2006) revealed that 227 respondents would be necessary for the balanced survey analyses. To determine this number, the power analysis was conducted using a significance level of less than .05, a power level of greater than .80, and an RMSEA of less than .08 (Kline, 2016).

19 classes at a large mid-Atlantic public university participated in the study. Class sizes ranged from 18 to 87 students, with an average class size of 51.6 students. The median class size was 48. These class sizes reflect the actual number of students present in the class the day the researcher visited, not the official number of students enrolled. The classes covered a broad range of topics, including education, English, marketing, statistics, and theatre. These courses also represented a variety of levels, from large, lower level core classes that were part of the general education requirements to smaller, upper level, major-specific classes. Out of a possible 930 students, 881 participated in the survey, for a response rate of 94.7%. Since the survey was given in person by pencil and paper, 12 out of the 19 classes had a 100% response rate. Of the seven remaining classes, six of them had response rates greater than 95%. One class had a response rate of 45% due to the professor requesting that the researcher come at the end of class isntead of the beginning and then the professor not leaving sufficient time for the survey to be completed prior to the end of class time.

Of the 881 who participated in the survey, 87 of them were not within the required age range of 18 to 23; therefore, their responses were removed prior to any analyses. The age of 18

to 23 was used in order to match the prior validation study of the instrument (Burch, Heller, Burch, Freed, & Steed, 2015). The total amount of usable responses was 794, with 403 responding to the balanced survey, 202 to the positively worded survey, and 189 to the negatively worded survey. These 794 respondents represent 3.5% of the total undergraduate population of the university. A demographic breakdown of the surveys can be seen in Table 2. The uneven number of survey respondents across survey type was due to nonresponse.

Table 2. Respondents by gender and race/ethnicity

	Balanced Survey		Positive Survey		Negative Survey		Total	
	n	%	n	%	n	%	n	%
Gender								
Female	281	69.7	144	71.3	129	68.3	554	69.8
Male	114	28.3	56	27.7	57	30.2	227	28.6
Other	7	1.7	2	1.0	3	1.6	12	1.5
Not Reported	1	0.2	0	0.0	0	0.0	1	0.1
Race/ethnicity								
White	196	48.6	117	57.9	90	47.6	403	50.8
Non-White	201	49.9	85	42.1	99	52.4	385	48.5
Not Reported	6	1.5	0	0.0	0	0.0	6	0.8
Total	403		202		189		794	

The university, as a whole, was 57.4% female and 50.8% white (Identifying Reference, 2016). In terms of race/ethnicity, the study mimicked the overall university population. However, women were overrepresented in the study population, which is not surprising, given that 18.4% of respondents were from classes in education.

Instrumentation

In order to assess acquiescence using structural equation models that extend the acquiescence literature, a previously validated survey with a confirmed factor structure was necessary. The Burch Engagement Survey for Students (BESS; Burch, Heller, Burch, Freed, & Steed, 2015) provided the foundation for this study. The BESS survey was chosen, because it was previously validated utilizing a population similar to the population used for this study. The pilot for the original BESS survey study included undergraduate students in the fall semester who averaged 21.7 years, were 53% women, and were 27.6% minorities. The confirmation study included undergraduate students who averaged 20.1 years, were 56% women, and were 27.8% minorities. Secondly, the survey contained four scales with an even number of questions in each scale. All questions on the original survey were phrased positively. The even number of positively worded questions made it easy to create negations of the questions and implement balanced and unbalanced combinations of positively and negatively worded questions. Finally, since this survey assessed student engagement at the classroom level, the sample did not need to be representative of the university as a whole, only of the classes in which it was administered.

The BESS survey was originally administered to 214 undergraduate students for development using exploratory factor analysis. All six items for each of the scales loaded well, without significant cross-loading. Coefficient alphas for reliability were very good: emotional engagement, .91; physical engagement, .93; cognitive engagement in class, .96, and cognitive engagement out of class, .96. Most of the variance was explained by the emotional engagement scale at 21.4%. The other three scales explained slightly less variance: physical engagement, 20.8%, cognitive engagement in class, 20.0% and cognitive engagement out of class, 17.4%. These four factors explained nearly 80% of the total variance.

After the exploratory analysis, the survey was then administered to 354 undergraduate students for confirmation of the factor structure using confirmatory factor analysis. Three models were considered with one, three, and four factors. Several goodness of fit measures were utilized, including comparative fit index (CFI), incremental fit index (IFI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and ratio of chi square relative to the degrees of freedom. The one factor model was unacceptable, with fit indices outside the a priori limits set for all five goodness of fit tests. Although the fit was better for the three factor model with the CFI, IFI, and SRMR being within the recommended range, the RMSEA was still well above .07 and the chi square over degrees of freedom was well above five. The four factor model had the best fit with all five fit statistics being within the recommended range. Therefore, the four-factor structure was confirmed, as seen in Table 3.

Table 3. Goodness of fit statistics for confirmatory factor analysis with unacceptable statistics highlighted in gray

Factors	CFI	IFI	RMSEA	SRMR	Chi Square
One	.84	.84	.23	.10	19.9
Three	>.9	>.9	.18	<.08	11.8
Four	.99	.99	.07	.04	2.6

The BESS survey focused on student engagement at the classroom level, whereas other engagement surveys, such as the National Survey of Student Engagement (Indiana University School of Education, 2016), focused on engagement at the university level. The BESS survey contained four scales, which each contained six positively worded questions. These questions are included in Table 4 Participants answered the questions using a five point Likert scale, ranging from “strongly disagree” to “strongly agree,” which was the same response scale as the original study and the same style (Likert) as the prior research.

Table 4. Burch Engagement Survey for Students questions

Scale	**Questions**
Emotional Engagement	I am enthusiastic about this class.
	I feel energetic when I am in this class.
	I am interested in material I learn in this class.
	I am proud of assignments I complete in this class.
	I feel positive about the assignment I complete in this class.
	I am excited about coming to this class.
Physical Engagement	I work with intensity on assignments for this class.
	I exert my full efforts towards this class.
	I devote a lot of energy toward this class.
	I try my hardest to perform well for this class.
	I strive as hard as I can to complete assignments for this class.
	I exert a lot of energy for this class.
Cognitive Engagement: In Class	When I am in the classroom for this class, my mind is focused on class discussion and activities.
	When I am in the classroom for this class, I pay a lot of attention to class discussion and activities.
	When I am in the classroom for this class, I focus a great deal of attention on class discussion and activities.
	When I am in the classroom for this class, I am absorbed by class discussion and activities.
	When I am in the classroom for this class, I concentrate on class discussion and activities.
	When I am in the classroom for this class, I devote a lot of attention to class discussion and activities.
Cognitive Engagement: Out of Class	When I am reading or studying material related to this class, my mind is focused on class discussion and activities.
	When I am reading or studying material related to this class, I pay a lot of attention to class discussion and activities.
	When I am reading or studying material related to this class, I focus a great deal of attention on class discussion and activities.
	When I am reading or studying material related to this class, I am absorbed by class discussion and activities.
	When I am reading or studying material related to this class, I concentrate on class discussion and activities.
	When I am reading or studying material related to this class, I devote a lot of attention to class discussion and activities.

Since the BESS survey was previously validated utilizing only positively worded questions, it provided an opportunity to incorporate negatively worded questions and gauge their impact. In addition to the pre-existing BESS survey, two new versions of the survey were created, one with

all negatively worded questions and one with an equal balance of positively and negatively worded questions. To accomplish this, negations of the positively worded questions were created. Swain, Weathers, and Niedrich's (2008) analysis of Bearden and Netemeyer's (1999) *Handbook of Marketing Scales* revealed that 81% of reversed items were negations. A negation is defined as the denial of an assertion (Horn, 1989). In the words of Dr. Seuss (1960), an assertion could be, "I like green eggs and ham." A negation would be, "I do not like green eggs and ham." Given the high percentage of reversed items in the literature that are formed as negations, this study utilized negation to create the negatively worded questions. The negative versions of the questions, created for this dissertation study, are included in Table 5.

Table 5. Burch Engagement Survey for Students with negatively worded questions

Scale	**Questions**
Emotional Engagement	I am NOT enthusiastic about this class.
	I do NOT feel energetic when I am in this class.
	I am NOT interested in material I learn in this class.
	I am NOT proud of assignments I complete in this class.
	I do NOT feel positive about the assignment I complete in this class.
	I am NOT excited about coming to this class.
Physical Engagement	I do NOT work with intensity on assignments for this class.
	I do NOT exert my full efforts towards this class.
	I do NOT devote a lot of energy toward this class.
	I do NOT try my hardest to perform well for this class.
	I do NOT strive as hard as I can to complete assignments for this class.
	I do NOT exert a lot of energy for this class.
Cognitive Engagement: In Class	When I am in the classroom for this class, my mind is NOT focused on class discussion and activities.
	When I am in the classroom for this class, I do NOT pay a lot of attention to class discussion and activities.
	When I am in the classroom for this class, I do NOT focus a great deal of attention on class discussion and activities.
	When I am in the classroom for this class, I am NOT absorbed by class discussion and activities.
	When I am in the classroom for this class, I do NOT concentrate on class discussion and activities.
	When I am in the classroom for this class, I do NOT devote a lot of attention to class discussion and activities.
Cognitive Engagement: Out of Class	When I am reading or studying material related to this class, my mind is NOT focused on class discussion and activities.
	When I am reading or studying material related to this class, I do NOT pay a lot of attention to class discussion and activities.
	When I am reading or studying material related to this class, I do NOT focus a great deal of attention on class discussion and activities.
	When I am reading or studying material related to this class, I am NOT absorbed by class discussion and activities.
	When I am reading or studying material related to this class, I do NOT concentrate on class discussion and activities.
	When I am reading or studying material related to this class, I do NOT devote a lot of attention to class discussion and activities.

Given that the BESS survey had four constructs, it provided an opportunity to compare the impact of positively and negatively worded questions on (dis)acquiescence across multiple

constructs. From the sets of positively and negatively worded questions, three surveys were developed, the breakdown of which is shown in Table 6. One survey was completely balanced, with three positive and three negative questions included for each scale. The second survey was entirely positive. The third survey was entirely negative. Copies of all three surveys are included in the appendix (See page 105).

Table 6. Breakdown of survey question polarization

BESS Scale	Survey 1 (Balanced)	Survey 2 (Positive)	Survey 3 (Negative)
Emotional Engagement	3 Positive 3 Negative	6 Positive 0 Negative	0 Positive 6 Negative
Physical Engagement	3 Positive 3 Negative	6 Positive 0 Negative	0 Positive 6 Negative
Cognitive Engagement: In Class	3 Positive 3 Negative	6 Positive 0 Negative	0 Positive 6 Negative
Cognitive Engagement: Out of Class	3 Positive 3 Negative	6 Positive 0 Negative	0 Positive 6 Negative

In addition to the BESS survey items, all three surveys included identical demographic questions. These questions requested the student's ethnicity, race and gender. An additional question assessed the age of the participants to ensure they were in the same age range as the original survey (18 to 23 years).

Procedure

Following IRB approval, the survey was administered via paper survey in late October and early November of 2016. Administering the survey during this timeframe made sure that students had spent enough time in their classes to be able to adequately self-report their engagement.

Within each class, students were randomly assigned to one of the survey treatments (all positive, all negative, or balanced). To accomplish this randomization, prior to entering a class for survey administration, the surveys were sorted in the following order: balanced survey, negative survey, balanced survey, positive survey, then repeated until the number of participants in the class is reached. Wherever one class left off in the survey order is where the next class would pick up. Since the balanced survey required higher power for analysis, oversampling of that survey was included in the randomization. Upon entering the class, the researcher read the following standardized script to the students:

> Script:
> Hello! I am Amy Hutton. I am conducting a research study that examines good research practices using a survey of student engagement in the classroom. If you would like to participate, you will be asked to take a short survey that will take less than 10 minutes to complete. Participation is voluntary. You are not required to participate in this study, but I hope you will choose to be part of it. Please be sure to read the introductory information. This is not an evaluation of your instructor and will not affect your grade for this class. Responses will be anonymous. I will now hand out the surveys. Once you have read the introductory information, you may begin answering the survey questions. Once all students have completed the survey, you will be asked to pass them forward.

After the standardized script was given, the surveys were distributed in the pre-arranged order to participating students, beginning with the front left side of the room (when facing the students), and proceeding left to right and front to back. Each of the three surveys contained identical introductory information, in order to ensure validity. Only the survey questions themselves varied, as seen in the appendix. Further, the front side of the survey was identical across all three

surveys. The back, which contained the varied scale questions, was formatted to look identical, minus the minor word changes for polarization.

Data Analysis

Prior to any data analysis, negatively worded questions were recoded to match the scaling of the positively worded questions. If the race question had two or more races selected, the respondent was coded as "non-white." Several respondents wrote in their race as "Arab" or "Middle Eastern." These respondents were also coded as "non-white." If any respondent double-marked an answer to one of the scaled items, that item was coded as missing. Once all data were inputted and recoded, the data were then assessed for outliers. Full information maximum likelihood estimation was used for the analyses, as it allows all data points to be included, even when a respondent was missing an item or items (Allison, 2001). To assess homogeneity of the sample across the three surveys, chi-square tests were performed for each of the demographic characteristics (Falthzik & Jolson, 1974). There were no significant differences across survey type by gender (p = .983) or race (p = .166).

Using a structural equation modeling framework, several confirmatory factor analyses were run. Confirmatory factor analysis was the best method for this study, as it directly assessed the unidimensionality of scales and provided goodness of fit statistics allowing different models to be compared (Hattie, 1985; Anderson, Gerbing, & Hunter, 1987; Gerbing & Anderson, 1988). The factor structure was based upon the models developed by Billiet and McClendon (2000), Cambre, Welkenhuysen-Gybels, and Billiet (2002), and DiStefano and Motl (2006; 2009). Separate analyses were run depending on the survey type. These analyses are discussed more in-depth in the following sections.

Balanced Survey

The balanced survey contained three positive and three negative questions for each of the four scales. Four different models were compared to determine if acquiescence differed depending on the polarization of the questions. Given that in the original survey study, the emotional engagement scale accounted for the most variance (21.4%), it was used for the model comparison. The other three scales were used to confirm the findings of the emotional engagement scale.

Before comparing the four acquiescence models, the balanced survey was modeled as unidimensional or with two factors (one positive, one negative). These two models served as fit comparisons for the more complex models that included acquiescence. It was expected, based upon prior research (Billiet & McClendon, 2000; Cambre, Welkenhuysen-Gybels, & Billiet, 2002; DiStefano & Motl, 2006; DiStefano & Motl, 2009), that these two models would have poorer fit than those that account for acquiescence.

Acquiescence Model 1, as seen in Figure 11, assumed a unidimensional content factor, where both positively and negatively worded questions were loaded onto a single construct. A unidimensional acquiescence factor used all questions as predictors. The lambdas for the acquiescence factor were constrained to one, as acquiescence was theorized to be equal across all questions.

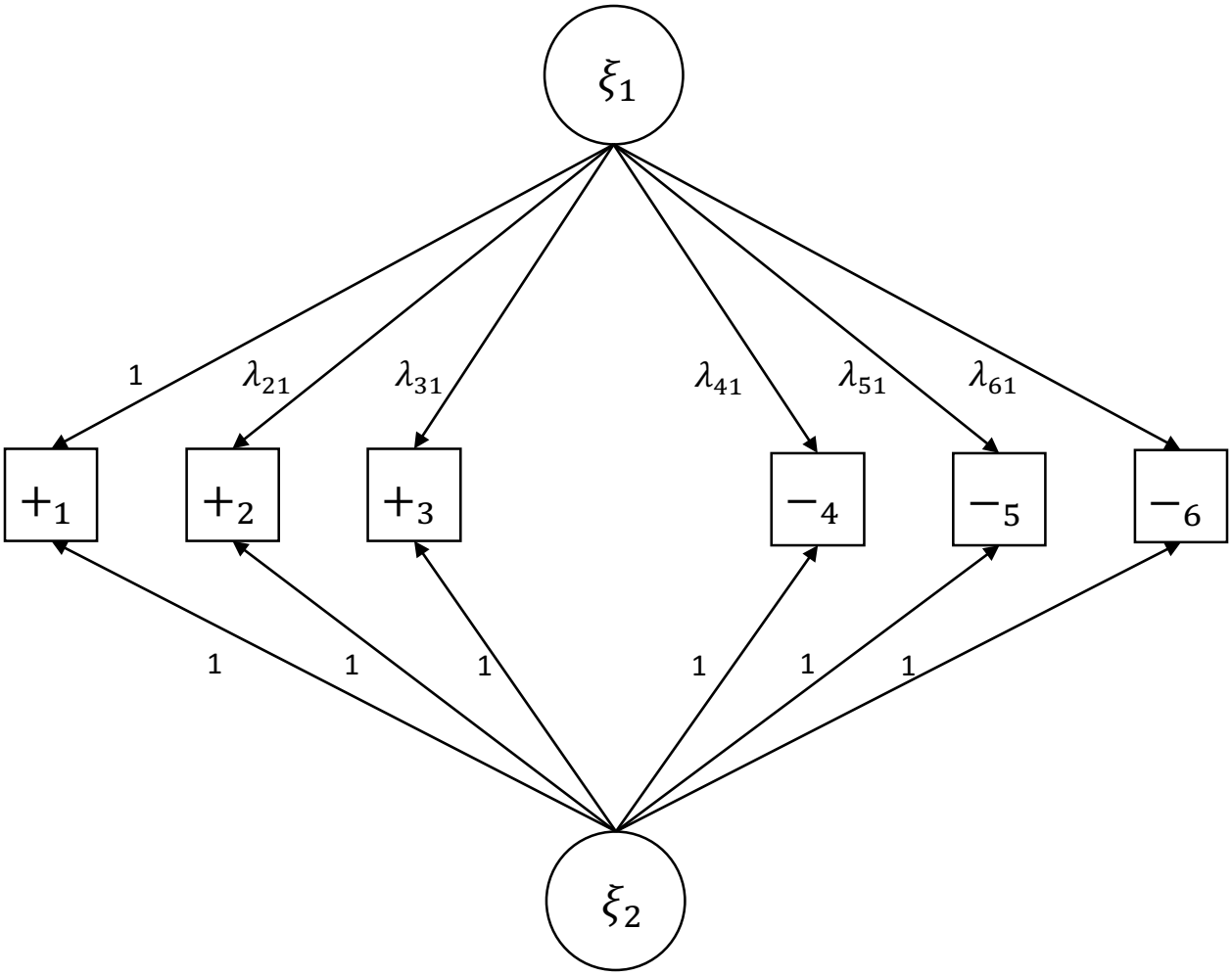


Figure 11. Model 1: One construct factor, one acquiescence factor for a balanced scale

The second model continued to assume a unidimensional construct latent factor, but split the acquiescence latent factor into a positive acquiescence factor and a negative acquiescence factor (disacquiescence). Like in the first model, the lambdas for each of the acquiescence factors were set to one, as each acquiescence factor was theorized to be equal across its specified questions, as seen in Figure 12. Two versions of Model 2 were considered, one where the acquiescence constructs were uncorrelated and one where they were correlated. This provided an opportunity to determine whether how a respondent acquiesced to positively worded questions was related to how they acquiesced slightly differently to negatively worded questions.

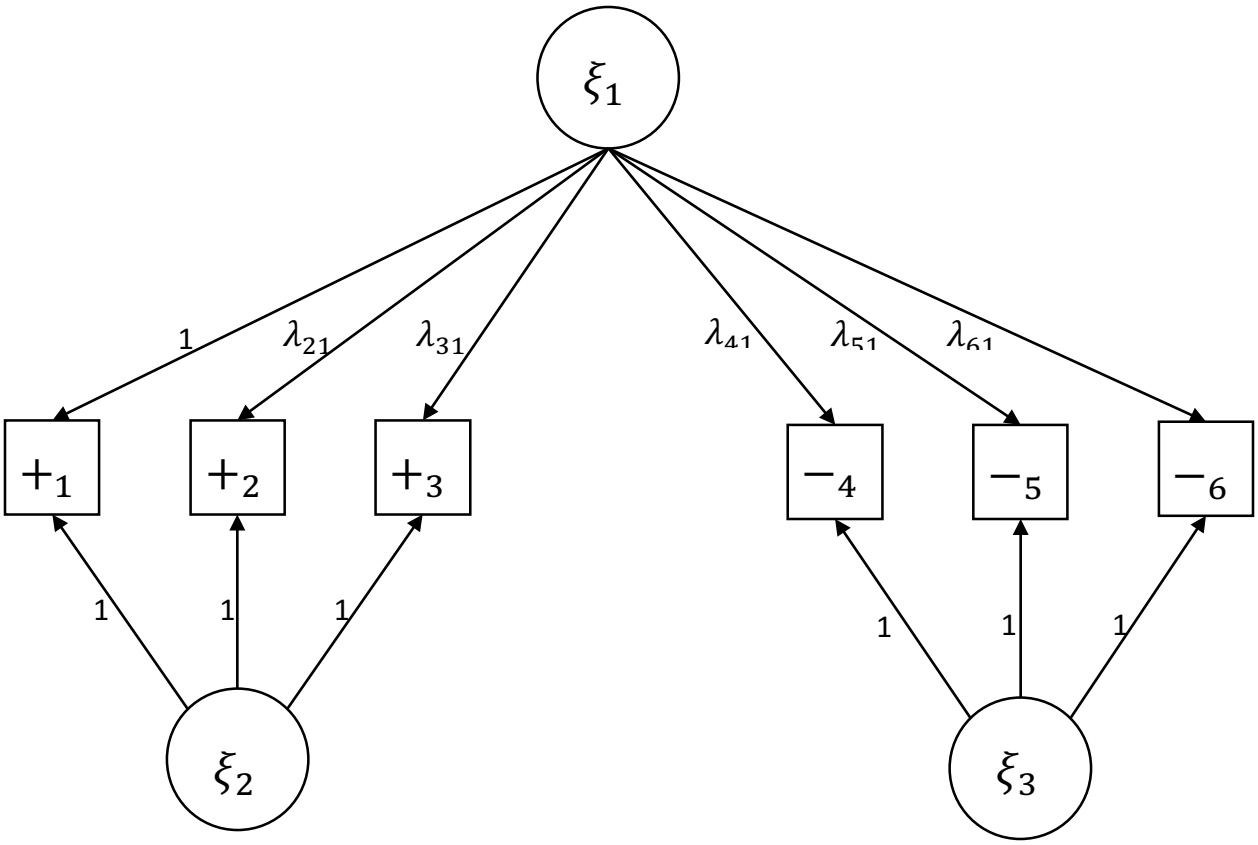


Figure 12. Model 2: One unidimensional content factor and two acquiescence factors, based upon wording polarization

The third model no longer assumed that the content factor was unidimensional, but split the content latent factor into a factor for the positively worded questions and a separate factor for the negatively worded questions. Given that these two latent content factors were theorized to be measuring similar, although slightly different, constructs, the two latent constructs were correlated. This third model assumed that acquiescence was unidimensional and did not vary based on item polarization; therefore, all lambdas were constrained to equal one, as shown in Figure 13.

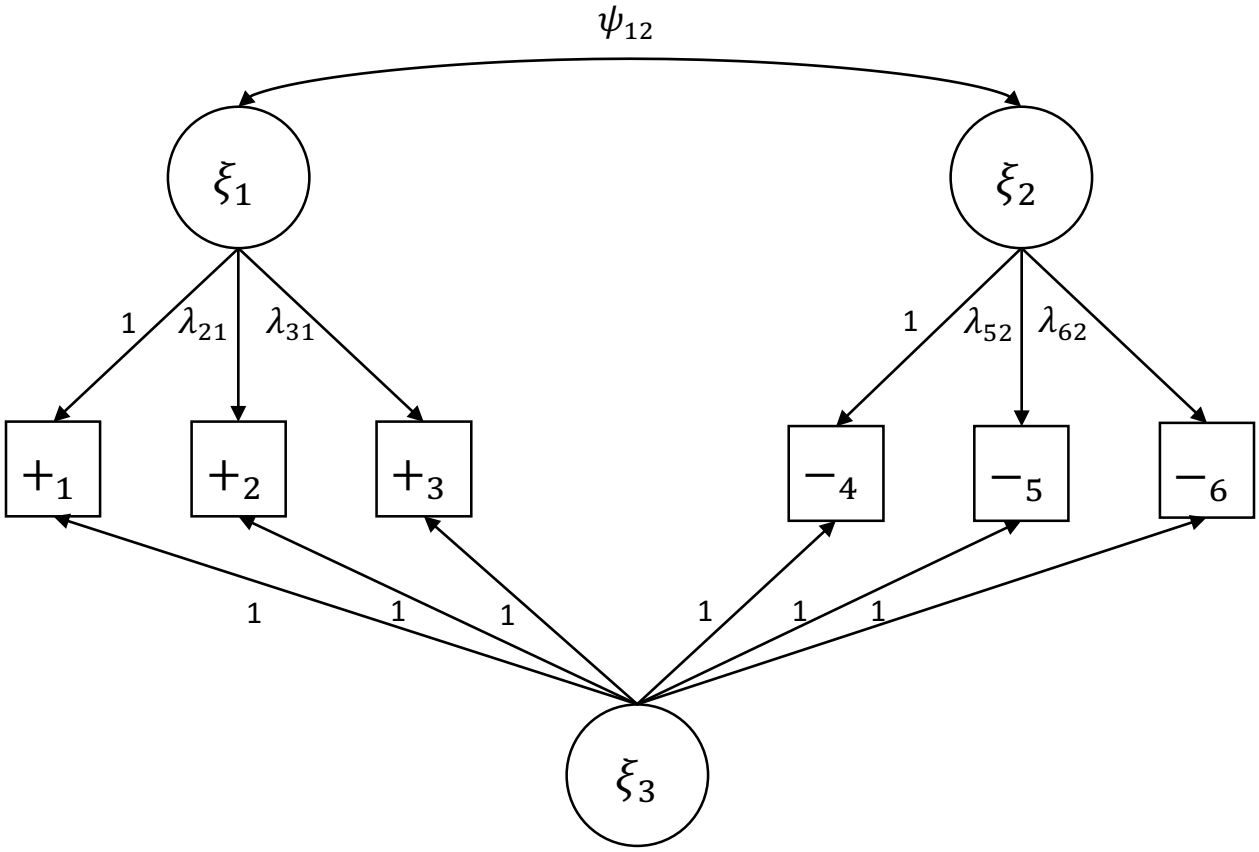


Figure 13. Model 3: Two construct factors, one acquiescence factor for a balanced scale

The final model, model four, no longer assumed that the content and acquiescence latent factors were unidimensional. Instead, both latent factors were split into separate factors, based upon the item polarization, as seen in Figure 14. The content latent factors were correlated, as they were in model 3. However, the acquiescence factors are uncorrelated, as it is theorized that acquiescence functions differently depending on the polarity of the questions. For comparison, the acquiescence factors were also considered correlated. For each of the acquiescence factors, the lambdas were constrained to equal one, as acquiescence was theorized to be equal across the items associated with that factor.

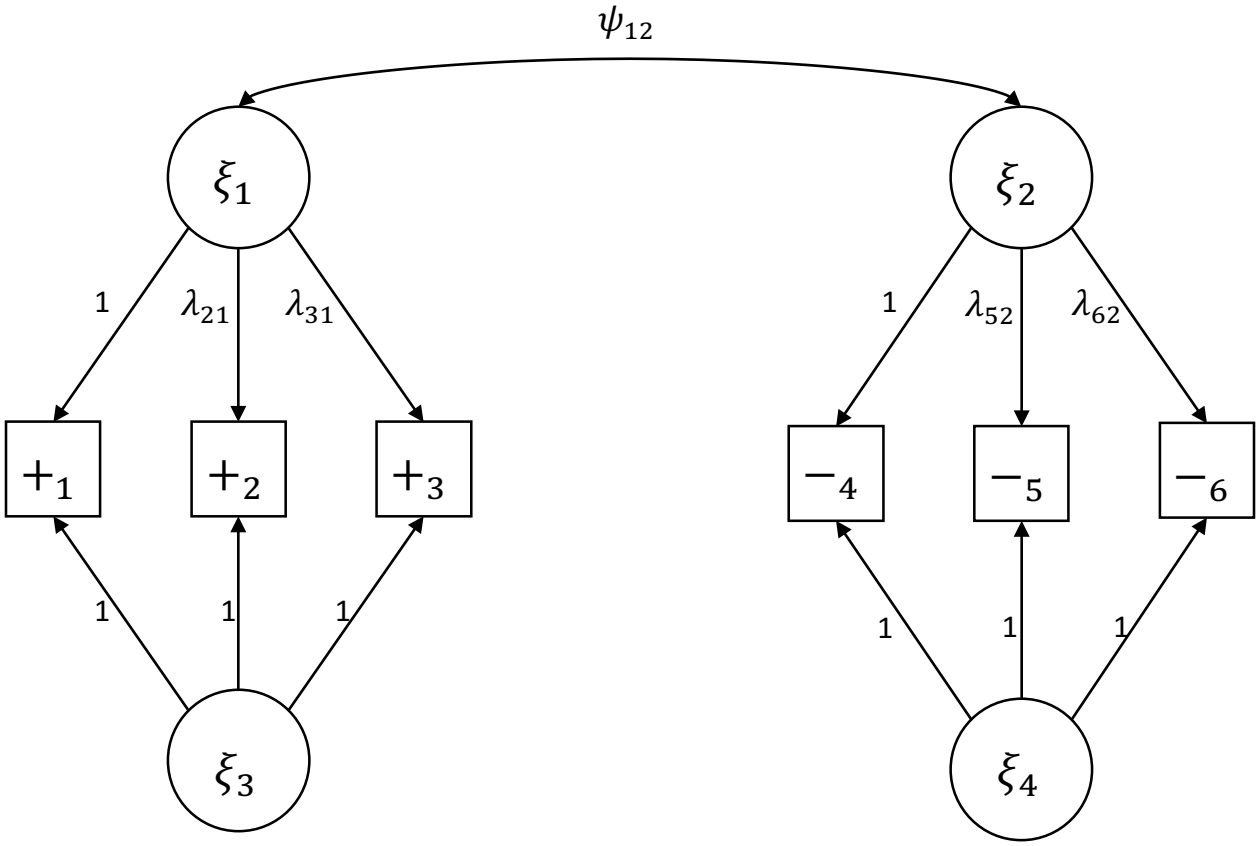


Figure 14. Model 4: Two construct factors, two acquiescence factors for a balanced scale

These four models were compared utilizing goodness of fit statistics with thresholds set a priori. The same thresholds utilized in the original survey (Burch, Heller, Burch, Freed, & Steed, 2015) were maintained for this study. These measures are the comparative fit index (CFI; above .90), root mean square error of approximation (RMSEA; values close to .05), standardized root mean square residual (SRMR; values less than .08), and the ratio of chi square relative to the degrees of freedom (χ^2/df; less than 5). The change in AIC was also observed across models. In addition to these thresholds being used in the original confirmatory factor analysis of the survey, they are shown to be appropriate in the structural equation modeling literature (Wheaton, Muthen, Alwin, & Summers, 1977; Bollen, 1989; Bentler, 1990; Hu & Bentler, 1999; Kline, 2016).

Demographics

The second step of the analysis was to conduct a multiple-samples confirmatory factor analysis to compare the four balanced survey models to determine if measurement invariance existed between the demographic groups. Respondents were compared by race (white vs. non-white) and gender (male vs. female). If at least strong invariance was determined across demographic groups, then any difference in the group means was directly related to the factor, rather than to bias (Kline, 2016). Kline (2016) states that strong invariance is the minimum requirement to interpret any differences in group means

As with the prior analysis, the fit of the groups was compared using the one and two factor base models, where acquiescence was not included. It was expected that strong invariance will not be attained, as these two models did not account for acquiescence bias. After assessing the base models, measurement invariance between the demographic groups was sought across all four acquiescence models. Given the potential heritability of acquiescence found by Alessandri et al. (2010) and the recommendation to see if different populations respond different to positively and negatively worded questions by Horan, DiStefano, and Motl (2003), it was possible that different demographic groups might fit different models better.

Comparison Across Surveys

The models proposed up to this point relied on a balanced survey. A slightly different approach was required in order to compare acquiescence across the three versions of the survey. To make this comparison, three versions of Model 1 were used, as seen in Figure 15, Figure 16, and Figure 17.

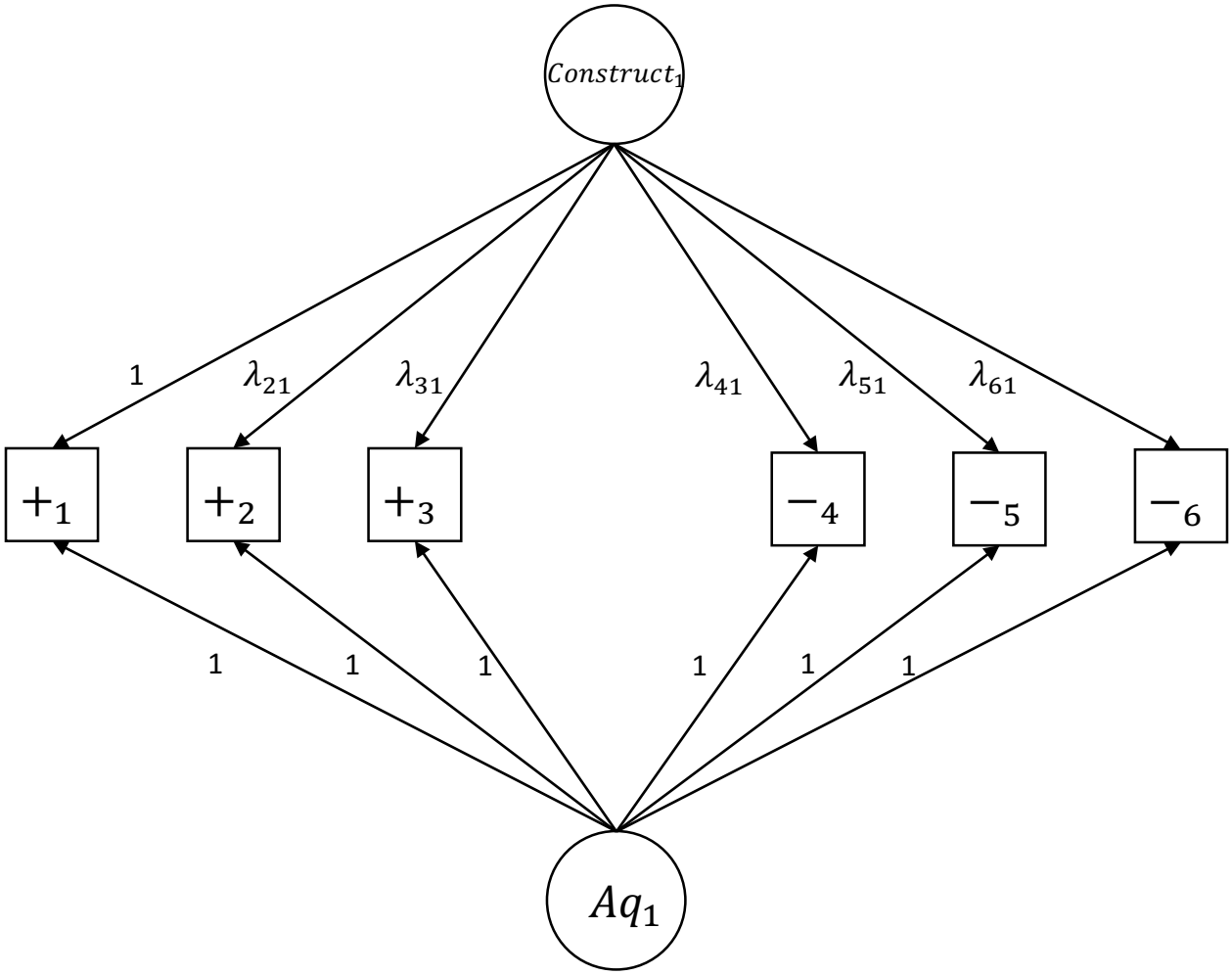


Figure 15. Model 1a for balanced survey

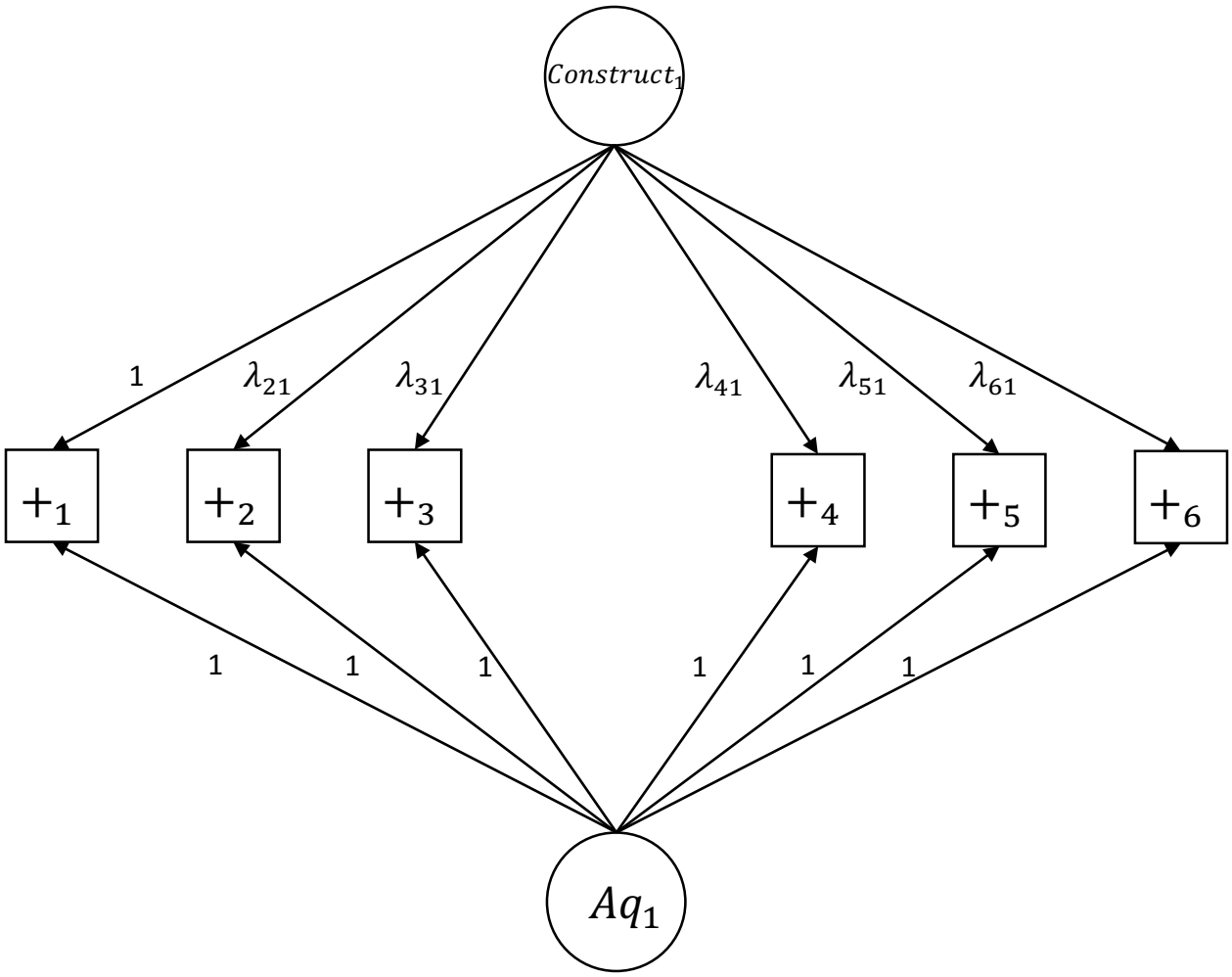


Figure 16. Model 1b for entirely positive survey

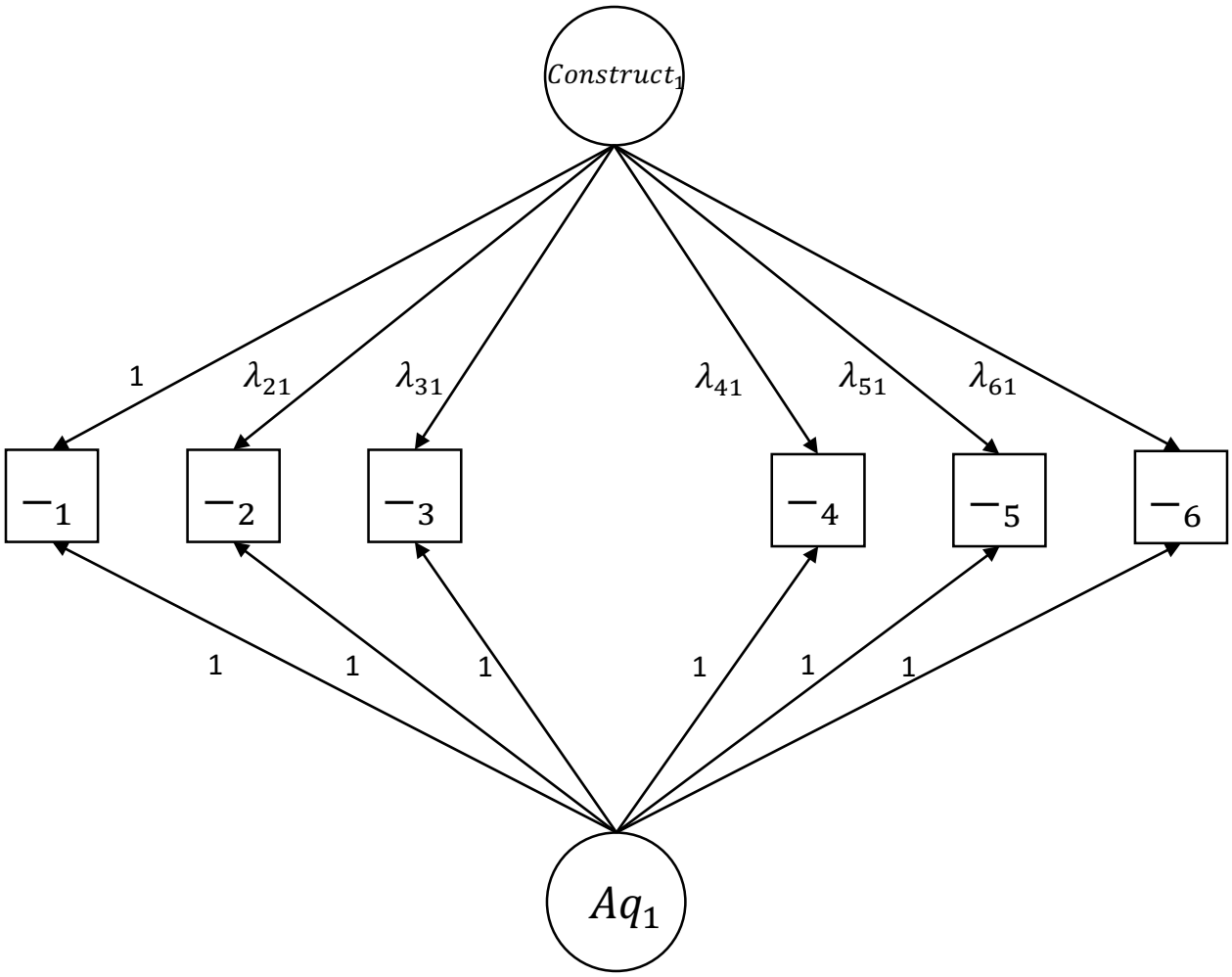


Figure 17. Model 1c for entirely negative survey

All three models assumed unidimensionality for both the construct latent factor and the acquiescence latent factor. The three models were compared to gauge the impact of acquiescence across the three survey interventions.

IV. RESULTS

For ease of understanding the progression of the study, the results are presented in the order of the three research questions. For the first research question, descriptive information specific to the balanced survey is presented, followed by the results of the model evaluation. The descriptions of the composition of the demographic groups for the second research question are provided first, followed by the invariance testing. Descriptive information and the results of the exploratory factor analyses are presented for the final research question before the results of the confirmatory factor analyses.

Research Question #1

The first research question asked, "When using a balanced survey, can statistical modeling of acquiescence allow for the unidimensional scaling of content?"

Descriptive Information

Across the entire balanced survey sample, the means differed for the four scales, based upon the polarity of the questions. For three out of four scales, as seen in Table 7, the mean for the recoded negatively worded question was higher than the positively worded mean, suggesting that participants could have been acquiescing more strongly to negatively worded questions, as supported by the literature (e.g., Motl, Conroy, and Horan, 2000). However, for the emotional engagement scale, the mean for the recoded negatively worded questions was actually lower than

the mean for the positively worded questions. This was surprising, as it was expected that all negatively worded means, once recoded, would be higher than the positively worded means. The lack of difference between the positively and negatively worded items for the cognitive engagement: in-class scale suggests that acquiescence may not be present, whereas the mean differences in the other scales suggests that something is causing respondents to differ between the positively and negatively worded items.

Table 7. Sample sizes, means, and standard deviations by scale

	Cognitive Engagement: In Class	Emotional Engagement	Physical Engagement	Cognitive Engagement: Out of Class
Positively Worded Questions				
n	378	384	386	390
Mean	3.71	3.51	3.37	3.50
SD	.863	.779	.881	.860
Negatively Worded Questions[a]				
n	393	393	398	391
Mean	3.76	3.32	3.58	3.65
SD	.872	1.025	1.000	.911
All Questions				
n	371	377	383	384
Mean	3.74	3.42	3.48	3.58
SD	.825	.830	.877	.852
Difference in Positive and Negative Means	.05	.19	.21	.15

[a]Negatively worded questions recoded to match positively worded questions

Prescreening for assumptions

For any given variable, less than 5% of the responses were missing. For the emotional engagement (p=.493), physical engagement (p=.778), and cognitive engagement: in-class (p = .17) scales, Little's MCAR test was non-significant; therefore the null hypothesis that the data was missing completely at random failed to be rejected (Garson, 2015). However, Little's MCAR test for the cognitive engagement: out-of-class scale was significant ($p < .001$), demonstrating that the data might not be missing completely at random. Therefore, this scale was not used in any further analyses. Given that the cognitive engagement: out-of-class scale

was the last of the four scales to be presented on the survey, it is not surprising that if participants skipped questions, they would be in the last portion of the survey.

Outliers were not evaluated for two reasons. First, all responses were within the range of the scale. Second, the purpose of this study was to account for bias that could be a cause of outliers. For these reasons, all responses were used in the analyses.

Models

For the emotional engagement scale, the base unidimensional model did not have good fit, as all fit indices were outside of recommended ranges (Kline, 2016). Splitting the emotional engagement content factor into two factors, one positive and one negative, decreased fit across most of the indices. These results seemed to confirm the unidimensional structure for this scale in the original study (Burch, Heller, Burch, Freed, & Steed, 2015).

Adding an acquiescence factor to the unidimensional content factor improved several of the fit indices over the unidimensional base model. The AIC and SRMR decreased and the CFI increased; however, the RMSEA stayed the same and the chi-square ratio increased slightly. Thus, model 1 was not a significant improvement over the unidimensional base model. It was in model 2, where the acquiescence factor was split into the positive and negative acquiescence factors while keeping a unidimensional content factor that fit increased significantly over the unidimensional base model. It was with this model that all of the fit indices were within the recommended ranges (Kline, 2016). Interestingly, adding a covariance between the acquiescence factors did not significantly improve fit. While the CFI increased and the SRMR decreased, demonstrating better fit, the RMSEA increased and the AIC increased, demonstrating weaker fit. This lack of improved fit demonstrates that how someone acquiesces to negatively worded questions may not be related to how he/she acquiesces to positively worded questions.

Models 3 and 4 did not converge; therefore, goodness of fit could not be assessed. This lack of convergence could be due to how well model 2 fit. Adding polarity-based acquiescence factors allowed the content factor to be unidimensional. Therefore, it is likely that models 3 and 4, which split the content factor by polarity, would either have poor fit or not converge. All fit indices for the emotional engagement scale can be found in Table 8.

Table 8. Comparison of models for emotional engagement scale

Fit Statistic	Model A	Model B	Model 1	Model 2 No cov	Model 2 w/ cov
CFI	0.846	0.834	0.863	0.994	0.995
RMSEA	.227	.250	.227	0.049	0.051
X^2/df	21.7	26.2	21.8	2.0	2.1
SRMR	0.087	0.107	0.080	0.048	0.039
# free parameters	18	19	19	20	21
AIC	6038.048	6054.235	6019.038	5860.318	5860.807

To confirm the results of the emotional engagement scale, the physical engagement and cognitive engagement: in class scales were used. The physical engagement scale provided very different results than the emotional engagement scale. Given that the questions for the physical engagement and emotional engagement scale were mixed together, it was expected that participants would acquiesce similarly on the two scales. Therefore, the fact that these scales behaved differently is important. For the physical engagement scale, the base models fit the data well. For Model A, the CFI and SRMR were in range, but the RMSEA and Chi-square ratio were out of range. Unlike the emotional engagement scale, splitting the content factor in the base model for the physical engagement scale had better fit than the unidimensional model. All fit indices, except for the RMSEA were in range. The RMSEA was close and significantly better

than for Base Model A. Further, the AIC decreased from Model A to Model B, as seen in Table 9. Where the physical engagement differed dramatically from the emotional engagement scale was that none of the models including acquiescence would converge. This demonstrated that any variance within the physical engagement scale was not due to acquiescence, but to some other factor. It is possible that the polarization of the question changed how respondents interpreted the questions, thus creating two content factors, when the original study only had one. This could be because the original study only used positively worded questions. How the entirely positively worded survey behaved in this study is discussed in more detail in the results of research question 3.

Table 9. Comparison of models for physical engagement scale

Fit Statistic	Model A	Model B
CFI	0.939	0.978
RMSEA	0.149	0.095
X^2/df	10.0	4.6
SRMR	0.041	0.026
# free parameters	18	19
AIC	5893.247	5842.504

The third scale behaved similarly to the physical engagement scale. Like the physical engagement scale, base model B had extremely good fit and none of the acquiescence models would converge. All fit indices were within range for base model B. The AIC also decreased from Model A to Model B. All fit indices for the base models are included in Table 10.

Table 10. Comparison of models for cognitive engagement: in class scale

Fit Statistic	Model A	Model B
CFI	0.966	0.997
RMSEA	0.122	0.040
X^2/df	7.0	1.6
SRMR	0.031	0.013
# free parameters	18	19
AIC	5088.765	5040.964

The excellent fit of base model B for this scale is likely due to a lack of acquiescence in this scale. The mean for the positively worded questions ($M = 3.71$) was not very different from the recoded mean for the negatively worded questions ($M = 3.76$). Although the recoded negatively worded mean is slightly higher, supporting the hypothesis that people acquiesce more strongly to negatively worded questions, there was not much difference between the two means. The individual item means can be found in Table 11.

Table 11. Item means for balanced cognitive engagement: in class scale

Item	Mean[a]
My mind is focused on class discussion and activities.	3.72
I DO NOT pay a lot of attention to class discussion and activities.	3.93
I focus a great deal of attention on class discussion and activities.	3.61
I am NOT absorbed by class discussion and activities.	3.57
I concentrate on class discussion and activities.	3.80
I DO NOT devote a lot of attention to class discussion and activities.	3.78

[a]Means for negatively worded questions were recoded for comparison.

Significance

Although not all of the proposed models converged, the most significant finding is that within the same sample, acquiescence differs depending on the content of the scale. Although participants acquiesced to the emotional engagement scale, they did not acquiesce to the physical or cognitive: in class engagement scales. Given that the physical and emotional engagement items were mixed together, acquiescence being present in only one of the scales demonstrates that this likely was not due to a method bias, but due to actual differences in how respondents acquiesce. This suggests that more research into how acquiescence is related to content is necessary before a generalizable statistical model accounting for acquiescence can be attained, assuming that a generalizable model is even possible.

Further, the physical and cognitive: in-class engagement scales did converge with an acquiescence factor at all, but still had a split factor structure, based upon polarization. This research shows that these split factor loadings are not due to acquiescence. Further research is required to ascertain what causes these split factor loadings.

Research Question 2

Research question 2 asked, "Does demographic group assignment impact the modeling of acquiescence?" Given the lack of convergence in the previous models, only the models that converged were considered for this analysis.

Descriptive Information

In comparing the means by group, as shown in Table 12, the differences between males and females were much larger than the differences between whites and non-whites, except on the cognitive engagement: in-class scale. *t*-tests revealed significant differences by gender on the

emotional and physical engagement scales. No other scales showed statistically significant differences by gender. None of the scales had statistically significant differences based upon race / ethnicity. Therefore, it appears that respondents did not differ much by race / ethnicity, but might acquiesce differently by gender.

Table 12. Sample size, means, and standard deviation by group for balanced survey scales

Group		Cognitive Engagement: In-Class	Emotional Engagement	Physical Engagement	Cognitive Engagement: Out-of-Class
Gender					
Female					
	n	264	260	265	269
	Mean	3.77	3.49	3.57	3.61
	SD	.808	.808	.868	.861
Male					
	N	99	109	110	107
	Mean	3.68	3.27	3.26	3.51
	SD	.841	.844	.858	.828
Mean Difference		.09	.22*	.31**	.10
Race/ethnicity					
White					
	N	183	185	189	191
	Mean	3.80	3.47	3.50	3.64
	SD	.811	.839	.869	.843
Non-White					
	N	184	186	188	187
	Mean	3.71	3.39	3.48	3.55
	SD	.803	.826	.886	.853
Mean Difference		.09	.08	.02	.09
Total		3.74	3.42	3.48	3.58

*Significant at .05 level **Significant at .01 level

Race / Ethnicity Group Comparison

The first comparison was to assess white versus non-white respondents to see if they acquiesced similarly or differently across the different models. In comparing whites and non-whites, model B did not even meet weak invariance, so that model likely includes some bias between these groups. Given that model B showed poor fit across the full sample, it is not surprising that the poor fit may be due to bias across groups. However, models A and 1 met strong (metric) invariance. Therefore, any difference in means between the two groups was not due to bias in the measure, but due to true differences between the groups. Model 2, which was the best fitting model across all respondents, did not converge. This was likely due to low power after dividing the sample across groups. However, given that model 1 included acquiescence and was not biased between these groups, it is plausible that acquiescence may not differ across these groups. Further research and a larger sample is necessary to obtain a comparison of model 2 and to validate the lack of difference in acquiescence between groups. The invariance comparison for emotional engagement can be seen in Table 13.

Table 13. Significance values for invariance testing by model for whites / non-whites on the emotional engagement scale

Invariance Comparison	Model A	Model B	Model 1	Model 2 No cov	Model 2 w/ cov
Metric against Configural	.538	.003*	.513	Did not converge	
Scalar against Configural	.763	.016*	.616		
Scalar against Metric	.772	.637	.567		

*$p < .05$

Not surprisingly, the physical engagement scale differed from the emotional engagement scale. Both base models met strong (metric) invariance standards, showing that the measure was unbiased based upon white / non-white grouping.

Table 14. Significance values for invariance testing by model for whites / non-white on the physical engagement scale

Invariance Comparison	Model A	Model B
Metric against Configural	.381	.435
Scalar against Configural	.385	.322
Scalar against Metric	.373	.244

It was in the cognitive engagement: in-class scale that base model A had difficulty. Although that model had pretty good fit, there is clear bias in the measure between white and non-white groups, as it did not pass any of the invariance tests. However, model B, which had excellent fit, passed all of the invariance tests. This provides further support that model B is the best model for the cognitive engagement: in-class scale.

Table 15. Significance values for invariance testing by model for cognitive engagement: in-class scale by white / non-white

Invariance Comparison	Model A	Model B
Metric against Configural	.003*	.126
Scalar against Configural	.008*	.161
Scalar against Metric	.308	.332

*$p < .05$

Given that only one model that contained acquiescence converged, it is unclear whether acquiescence differs across white and non-white groups. However, it appears that acquiescence may not differ by race / ethnicity; however, the content factors when considered without acquiescence may differ by race ethnicity. Further research is needed to validate these results.

Gender Comparison

Comparisons by gender provided very different results than those by white / non-white race/ethnicity grouping. The base models of the emotional engagement scale showed significant bias between males and females. This further supports the conclusion that the base emotional engagement models have poor fit. Model 1 did not converge, which could mean that the model had such poor fit that it could not converge or could be due to low power when splitting the sample across groups. However, both versions of model 2 did converge. Both versions of model 2 passed all of the invariance tests. This supports the conclusion that model 2 best fits the data. More importantly, it shows that acquiescence may not differ by gender. However, given that model 1 did not converge, this is a tentative conclusion that will need more research to validate. The results of the invariance tests for emotional engagement are included in Table 16.

Table 16. Significance values for invariance testing by model for females / males on the emotional engagement scale

Invariance Comparison	Model A	Model B	Model 1	Model 2 No cov	Model 2 w/ cov
Metric against Configural	.007*	<.001*	No conv.	.252	.167
Scalar against Configural	.001*	<.001*		.393	.277
Scalar against Metric	.011*	.003*		.610	.567

*$p < .05$

Up until now, all of the between group bias was in models with poorer fit. However, in the physical engagement scale, model B, which had the best fit across all respondents, did not meet any of the invariance tests. This means that bias was present in model B and that differences in means might not be due to true score differences.

Table 17. Significance values for invariance testing by model for females / males on the physical engagement scale

Invariance Comparison	Model A	Model B
Metric against Configural	.144	.057
Scalar against Configural	.340	.269
Scalar against Metric	.699	.939

In comparing the model fit by population, model B had better fit for both populations than model A. However, model B had almost perfect fit for males; whereas the fit was not as good for females. All fit indices for females were in range, but the RMSEA was out of range. The 90% confidence interval for the RMSEA also did not include 0.05. The fit index comparison between females and males can be seen in Table 18.

Table 18. Model fit of physical engagement scale by females and males

	Females		Males	
Fit Statistic	Model A	Model B	Model A	Model B
CFI	.932	.968	.944	1.00
RMSEA	.158	.116	.137	.000
X^2/df	8.03	4.75	3.14	0.97
SRMR	.045	.033	.044	.022
# free parameters	18	19	18	19
AIC	4087.965	4055.729	1705.465	1686.968

In assessing the cognitive engagement: in-class scale by females and males, the results supported the conclusion that model B had better fit than model A. However, although model B passed the invariance tests, it was approaching significance on two of the tests, as seen in Table 19. Follow-up testing may be necessary to see if model B truly passes the invariance tests.

Table 19. Significance values for invariance testing by model for cognitive engagement: in-class scale by females / males

Invariance Comparison	Model A	Model B
Metric against Configural	.032*	.223
Scalar against Configural	.012*	.061
Scalar against Metric	.064	.056

*$p < .05$

Summary

In conclusion, it is difficult to determine whether acquiescence differs by group membership. However, it is clear that responses to the content may differ by group membership. Although acquiescence in some models did not differ by group membership, others did not converge. More research and larger sample sizes in the groups are necessary to determine whether acquiescence differs across groups.

Research Question 3

Research question 3 states, "Does acquiescence differ by survey type (all positive, all negative, or balanced)?"

Descriptive Information

The three surveys were randomly distributed across all classes in order to prevent classroom effects from skewing the scores. The means for each of the surveys and scales are listed in Table 20, along with the reliability for each of the scales. There were definite differences between the three surveys, in terms of scale scores. The estimated reliabilities, using Cronbach's

alpha, were all decently high. However, the emotional and physical engagement scales were lower than the cognitive engagement reliabilities.

Table 20. Means and Cronbach's alphas by survey type and scale

Scale	Balanced Survey[a]	Positive Survey	Negative Survey[a]
Cognitive Engagement: In Class			
n	371	193	185
Mean	3.74	3.63	3.70
Standard Deviation	.825	.842	.962
Cronbach's Alpha	.922	.948	.960
Emotional Engagement			
n	377	199	178
Mean	3.42	3.40	3.52
Standard Deviation	.830	.767	.832
Cronbach's Alpha	.867	.842	.869
Physical Engagement			
n	383	190	179
Mean	3.48	3.49	3.59
Standard Deviation	.877	.777	.896
Cronbach's Alpha	.898	.868	.895
Cognitive Engagement: Out of Class			
n	384	190	184
Mean	3.58	3.55	3.68
Standard Deviation	.852	.787	.917
Cronbach's Alpha	.940	.951	.965

[a]Negatively worded questions recoded to match positively worded questions

In order to assess how the scales compared across all three surveys, exploratory factor analyses were considered. Given that the missing data for the cognitive engagement: out of class scale was not completely at random, that scale was not used in the factor analysis. Instead, the goal was for all three surveys to show a three factor solution, one for each of the three remaining scales. Principal components analysis with a varimax rotation was utilized. For all three surveys, the Kaiser-Meyer-Olkin Measure of Sample Adequacy was above .9, which was excellent (Tabachnick & Fidell, 2013), and over 70% of the variance was explained, as seen in Table 21.

Table 21. KMO test results and variance explained for all three surveys

	Balanced	Positive	Negative
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	.941	.935	.916
Variance Explained	73.7%	78.8%	76.1%

All three surveys provided the same factor structure, demonstrating that the factor structure was consistent, regardless of item polarization, as seen in Tables Table 22, Table 23, and Table 24. Although the three surveys were consistent, their results did not match the factor structure of the original survey, as the emotional engagement scale split into two factors on all three surveys. Emotional engagement items 1, 2, 3, and 6 loaded strongly on one factor, while items 4 and 5 loaded strongly on a separate factor. Items four and five clearly functioned differently from the rest of the emotional engagement scale. The emotional engagement scale was unidimensional in the original study, so these results are inconsistent with the original study. It is possible that the emotional engagement scale in the confirmatory factor analysis part of the original study could have shown a split factor loading, but that was not one of the models tested in the confirmatory factor analysis. For the full measure, with all four scales, the original confirmatory study only

tested for one, two, and four factors. Therefore, the only opportunity for the split factor loadings in the original study was in the exploratory factor analysis phase (Burch, Heller, Burch, Freed, & Steed, 2015).

Table 22. Rotated component matrix for balanced survey

Scale	Item (+/-)	Factor 1	Factor 2	Factor 3	Factor 4
Cognitive Engagement: In Class	1 (+)	**.797**	.190	.201	.189
	5 (+)	**.778**	.259	.202	.228
	2 (-)	**.774**	.286	.192	.054
	3 (+)	**.769**	.276	.205	.250
	6 (-)	**.754**	.308	.284	.057
	4 (-)	**.671**	.190	.372	.107
Physical Engagement	5 (-)	.227	**.784**	.200	.112
	4 (-)	.291	**.753**	.139	.046
	6 (-)	.317	**.715**	.396	.051
	1 (+)	.149	**.684**	.230	.342
	2 (+)	.289	**.679**	.177	.304
	3 (+)	.234	**.678**	.307	.319
Emotional Engagement	1 (-)	.317	.301	**.783**	.139
	6 (+)	.269	.213	**.761**	.245
	3 (-)	.245	.183	**.761**	.168
	2 (-)	.239	.293	**.758**	.130
	5 (+)	.173	.231	.220	**.828**
	4 (+)	.234	.237	.206	**.816**

Table 23. Rotated component matrix for negative survey

Scale	Item	Factor 1	Factor 2	Factor 3	Factor 4
Cognitive Engagement: In Class	2	**.875**	.133	.236	.123
	3	**.872**	.282	.178	.131
	6	**.841**	.378	.139	.097
	1	**.831**	.210	.268	.142
	5	**.822**	.224	.236	.184
	4	**.784**	.262	.307	.149
Physical Engagement	2	.218	**.757**	.200	.237
	3	.354	**.745**	.273	.116
	6	.237	**.736**	.472	.085
	1	.211	**.709**	.375	.071
	5	.329	**.645**	-.071	.494
	4	.361	**.569**	.000	**.516**
Emotional Engagement	6	.212	.254	**.808**	.180
	3	.285	.042	**.758**	.241
	1	.295	.370	**.732**	.148
	2	.322	.421	**.625**	.140
	5	.172	.181	.230	**.851**
	4	.108	.167	.412	**.769**

Table 24. Rotated component matrix for positive survey

Scale	Item	Factor 1	Factor 2	Factor 3	Factor 4
Cognitive Engagement: In Class	2	**.869**	.135	.257	.135
	3	**.861**	.192	.240	.111
	5	**.848**	.224	.203	.080
	1	**.842**	.171	.199	.173
	6	**.826**	.224	.237	.080
	4	**.786**	.148	.295	.019
Physical Engagement	3	.226	**.798**	.308	.135
	4	.224	**.689**	.010	.452
	6	.224	**.676**	.453	-.142
	5	.234	**.665**	-.019	.501
	1	.038	**.659**	.364	.182
	2	.429	**.625**	.094	.225
Emotional Engagement	6	.291	.208	**.787**	.236
	3	.243	.148	**.764**	.137
	1	.428	.176	**.712**	.244
	2	.322	.223	**.695**	.034
	5	.128	.170	.217	**.888**
	4	.100	.261	.201	**.836**

Survey Model Comparison

Due to a lack of convergence on model 1 across the physical engagement and cognitive engagement: in-class scales, the only scale that could be assessed using the proposed model 1 for acquiescence across all three survey types was the emotional engagement scale. For comparison, the fit of model A was also compared across survey types, as seen in Table 25.

Table 25. Comparison of model A across all three survey types

Fit Statistic	Positive	Negative	Balanced
CFI	.751	.833	.846
RMSEA	.292	.233	.227
X^2/df	18.2	11.2	21.7
SRMR	.110	.085	.087
# free parameters	18	18	18

Model A did not fit any of the surveys well. This model fit the balanced and negative surveys better than the positive survey; however, for all three surveys, all of the fit statistics were outside of the recommended ranges. This is not surprising, given the exploratory factor loadings shown previously. Model 1 had much better fit than model A. Although not all of the fit indices were within range, both the positively worded and negatively worded surveys demonstrated moderate fit. The CFI and chi-square ratio were within the recommended ranges and the RMSEA and SRMR were fairly close. However, this model did not fit the balanced survey nearly as well. As shown previously, this is due to how well model 2 fit the emotional engagement scale for the balanced survey.

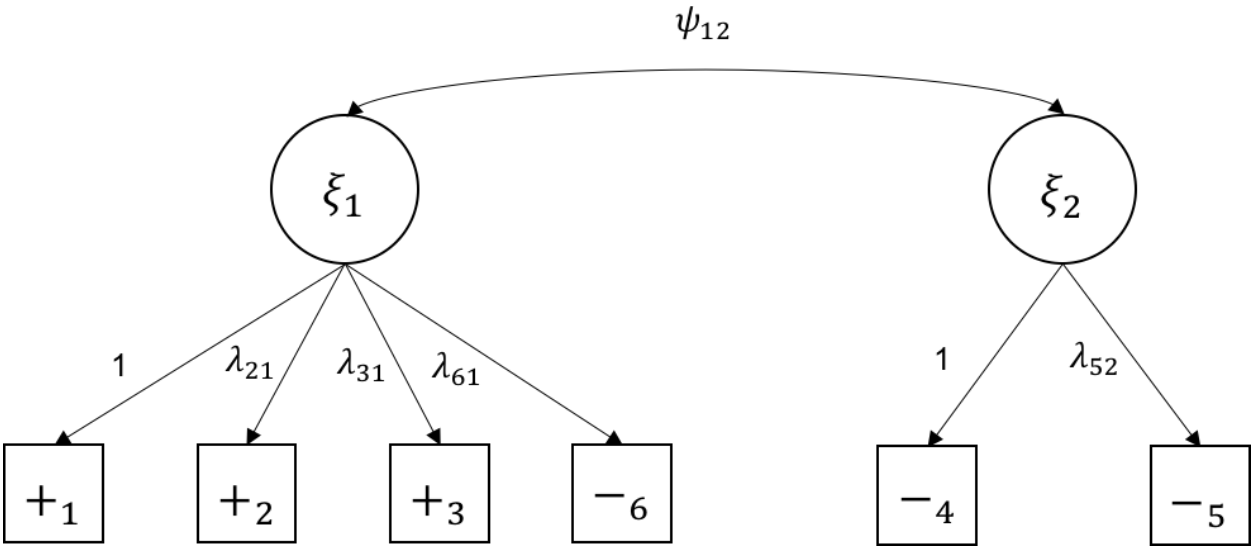
Table 26. Comparison of model 1 across all three survey types

Fit Statistic	Positive	Negative	Balanced
CFI	.960	.969	.863
RMSEA	.124	.107	.227
X^2/df	4.12	3.14	21.8
SRMR	.072	.053	.080
# free parameters	19	19	19

Although model 1 fit differently across the three survey types, the better fit of the model with the acquiescence factor supports the assertion that acquiescence affected participants' responses to the emotional engagement scale. Acquiescence affected emotional engagement, regardless of the polarity of the questions. This shows that simply eliminating negatively worded questions does not prevent acquiescence. However, the fact that the balanced survey did not fit model 1 well, but instead fit model 2 well, where the acquiescence factors were split between positive and negative items, demonstrates that people acquiesce to emotional engagement questions differently for positively worded questions that negatively worded questions. If all questions have the same polarity, model 1 is sufficient to assess acquiescence; however, combining both question types requires a more sophisticated statistical model to account for acquiescence.

Since the exploratory factor analysis demonstrated a split factor loading for the emotional engagement scale across all three survey types, shown in Figure 18. Model for emotional engagement scale from EFA, that model was also tested for goodness of fit. This enabled the researcher to determine if the split factor loading found in the EFA was due to acquiescence or to a different issue with the instrument itself.

Figure 18. Model for emotional engagement scale from EFA



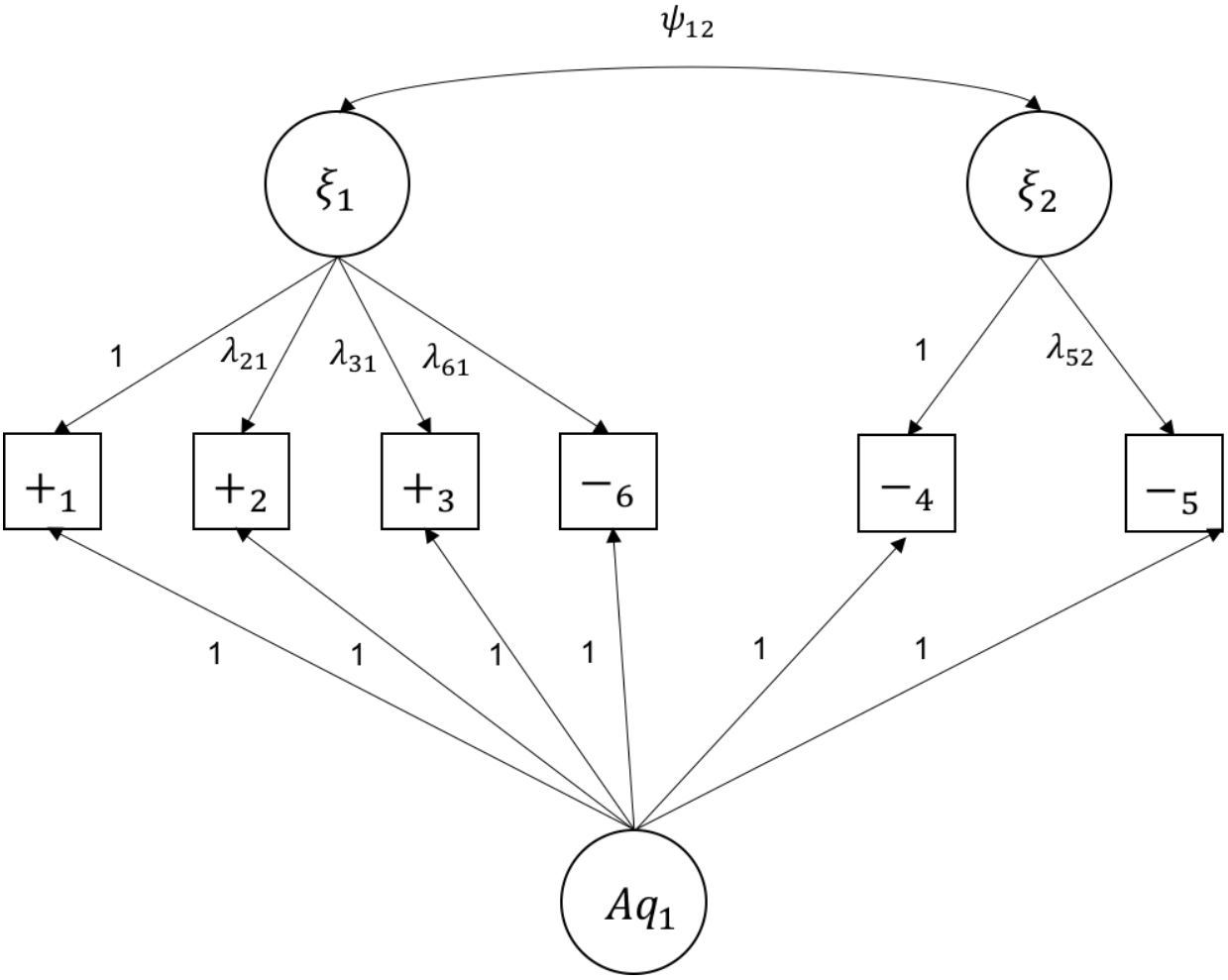
Goodness of fit indices for the EFA model, available in Table 27, revealed perfect fit for the balanced survey and good fit for the positively and negatively worded surveys. The perfect fit for the balanced survey, without an acquiescence factor, demonstrates that acquiescence may not exist in the emotional engagement scale after all. Or, if it is, it is masked by a problem with the instrument itself.

Table 27. Goodness of fit statistics for EFA model by survey type

Fit Statistic	Positive	Negative	Balanced
CFI	.972	.974	1.00
RMSEA	.103	.097	0.00
X^2/df	3.15	2.78	0.97
SRMR	.029	.026	0.01
# free parameters	19	19	19

Given the structure of the EFA model, it is not possible to add individual acquiescence factors to match the content factors, as the model would not be identified. However, a unidimensional acquiescence factor could be added, as shown in Figure 19.

Figure 19. EFA factor loading with unidimensional acquiescence factor



For the balanced survey, as seen in Table 28. Goodness of fit statistics for EFA model with unidimensional acquiescence factor by survey type, the model with a unidimensional acquiescence factor had excellent fit. Even though it had excellent fit, it was not as strong as the model without acquiescence. The positively and negatively worded surveys did not converge on this model, which could be due to poor fit or a lack of power.

Table 28. Goodness of fit statistics for EFA model with unidimensional acquiescence factor by survey type

Fit Statistic	Positive	Negative	Balanced
CFI	Did not converge		.999
RMSEA			.016
X^2/df			1.11
SRMR			.011
# free parameters			20

Even though model 1 did not converge for the physical engagement and cognitive engagement: in-class scales for the balanced survey, it was evaluated for these scales for the all positively and all negatively worded surveys, in order to see if these surveys supported the finding of a lack of acquiescence. As seen in Table 29. Fit statistics for model 1 on the all positively worded and all negatively worded remaining scales, acquiescence was not consistent even by survey type. For the cognitive engagement: in-class scale, acquiescence was not only present in the all positively worded survey, but had excellent fit. Although the fit of base model A was good, shown in Table 30, model 1 was significantly better. However, for the negatively worded survey, the acquiescence model did not converge. Instead, base model A had good fit, with most fit indices within or near range.

For the physical engagement scale, it was the reverse. The negatively worded survey had perfect fit for model 1, but weaker fit on base model A. The positively worded survey had good fit on base model A and did not converge when acquiescence was added in model 1.

Table 29. Fit statistics for model 1 on the all positively worded and all negatively worded remaining scales

	Cognitive Engagement: In-Class		Physical Engagement	
Fit Statistic	All Positive	All Negative	All Positive	All Negative
CFI	.996	No convergence	No convergence	1.00
RMSEA	.053			0.00
X^2/df	1.57			0.71
SRMR	.032			.028
# free parameters	19			19

Table 30. Fit statistics for base model A on the all positively worded and all negatively worded remaining scales

	Cognitive Engagement: In-Class		Physical Engagement	
Fit Statistic	All Positive	All Negative	All Positive	All Negative
CFI	.990	.986	.918	.918
RMSEA	.074	.100	.158	.170
X^2/df	2.12	2.89	6.03	6.48
SRMR	.016	.015	.046	.054
# free parameters	18	18	18	18

The comparison of model 1 and base model A on the scales that did not converge using the balanced survey shows that there may be an interaction between content and question polarity, in terms of acquiescence. Even for all positively and all negatively worded surveys, acquiescence differed by the content, being present in some scales and not in others.

Summary

The results of research question #3 demonstrate that acquiescence may not be independent of content. Although acquiescence was not present for any of the scales on the balanced survey, it was present on the positively and negatively worded surveys for some of the scales. This means that instead of acquiescence existing consistently across scales, regardless of content, there may be an interaction between content and item wording.

V. DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

Discussion

The lack of consistent modeling of acquiescence across scales administered to the same population at the same time reveals that respondents acquiesce differently depending on the content of the measure. This finding creates fissures in the assertions of prior researchers studying acquiescence. It also provides a possible explanation for why findings and recommendations for preventing acquiescence differed dramatically across prior studies.

Issues with Content

Nunnally's (1967) heavily cited book asserts that by simply including a balance of both positively and negatively worded questions eliminates acquiescence bias. Further researchers (e.g., Marsh, 1986; Mirowsky & Ross, 1991) confirmed that including positively and negatively worded questions canceled out positive versus negative bias, allowing unidimensional scaling. Although the current study showed that acquiescence is not present in every scale, all of the scales without acquiescence still had better fit with factor loadings split by item polarity than using unidimensional scaling. This study is not the first to disprove Nunnally's theory (e.g., Herche and Engelland, 1996; Magazine, Williams, & Williams, 1996), but it provides additional evidence against Nunnally's heavily cited work. A Google Scholar search showed that Nunnally's book has been cited over 11,600 times from 2015 to the date of the search (December 21, 2016). Since assessing the presence or absence of acquiescence requires proficiency with

structural equation modeling or other statistical techniques, perhaps the simple and concrete recommendation of using balanced scales makes Nunnally's work continue to persist. Given how often how heavily utilized Nunnally's book is, it is important for the research disproving his theory to continue to be widely shared.

Other researchers (Bergstrom & Lunz, 1998; Baumgartner & Steenkamp, 2001; York, 2009) showed that unidimensionality could be maintained even with the inclusion of positively and negatively worded questions, as respondents did not differ in their response patterns. Therefore, these researchers highly recommended including both types of questions. The current study clearly demonstrates that unidimensionality cannot be assumed when both positively and negatively worded questions are included, regardless of whether or not a model with acquiescence fits the data.

Weijters, Geuens, and Schillewaert (2008) recommended utilizing reversed items only when they were next to non-reversed items from other scales. In the current study, the emotional and physical engagement scales were interspersed using the method recommended by Weijters, Geuens, and Schillewaert, but a unidimensional scale did not emerge.

Several researchers (e.g., Ray, 1979 & 1983; Weijters, Baumgartner, and Schillewaert, 2013) recommended including both positively and negatively worded questions, as the bipolar scale allows for bias to be identified and assessed. However, the current study demonstrates that a latent framework can identify acquiescence bias regardless of the polarization of the questions and separate it from the content.

The current study does not support the assertions of these researchers that including negatively worded questions can prevent acquiescence bias or maintain a unidimensional content scale. However, it also does not support the assertion of many other researchers, discussed in the

next section, that only positively worded items should be used in surveys. The current study clearly shows that acquiescence differed dramatically, depending on both item polarization and scale content.

This study does confirm many of the findings from other studies that demonstrate the difficulties that arise when using bipolar scales. Specifically, the lack of unidimensionality when only considering the content of the scale supports the findings of many studies (e.g., Herche & Engelland, 1996; Magazine, Williams, & Williams, 1996; Ibrahim, 2001).

Separating Acquiescence from Content

Many researchers have focused on trying to assess acquiescence using structural equation modeling, although researchers have not yet agreed on which model best accounts for acquiescence or how best to design surveys to either eliminate or latently account for acquiescence. Although the current study attempted to answer some of the questions surrounding modeling acquiescence, it instead identified additional questions that need to be asked.

Cambre, Welkenhuysen-Gybels, and Billiet (2002) took McClendon and Billiet's (2000) study a step further, by assessing whether a two-factor content model without acquiescence (base model B, in the current study) fit better than a unidimensional content factor with a unidimensional acquiescence factor (model 1, in the current study), shown in Figure 20.

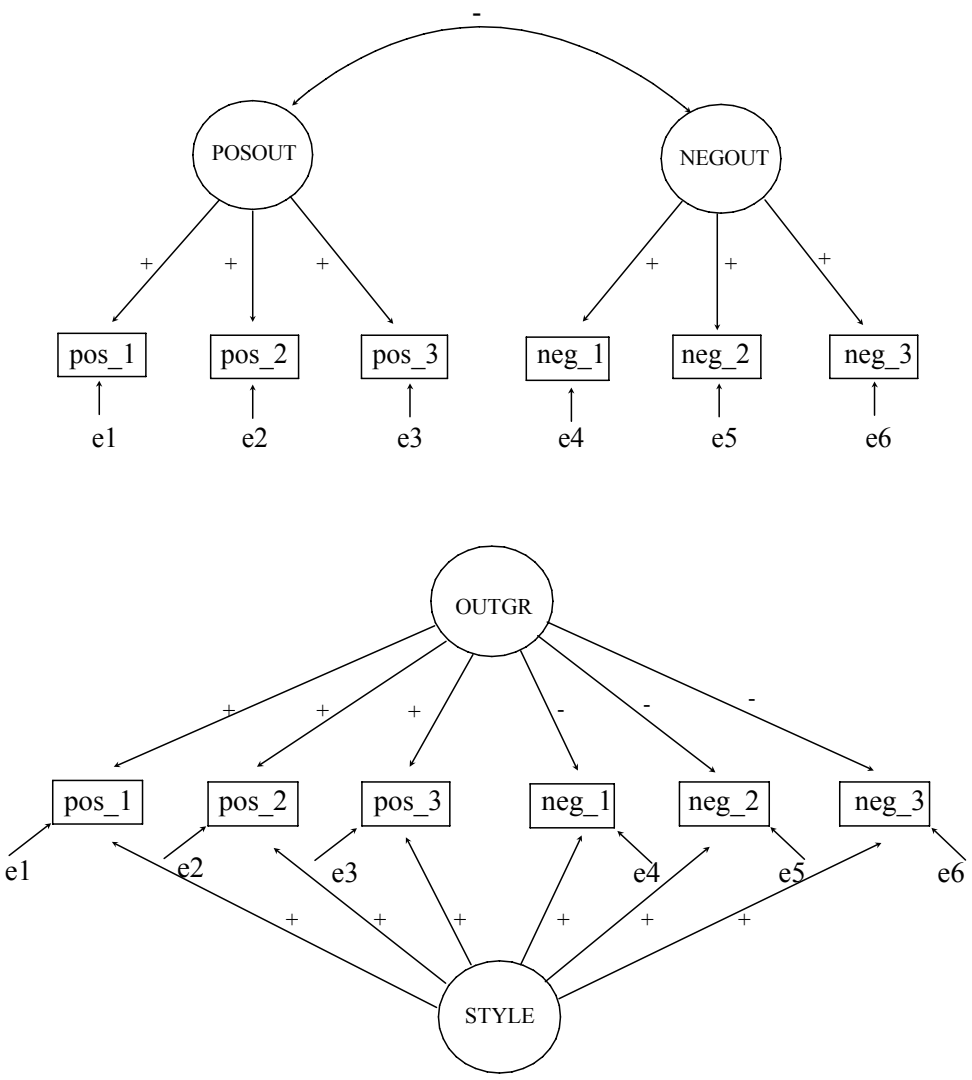


Figure 20. Cambre, Welkenhuysen-Gybels, and Billiet (2002) models (p. 3)

The results of Cambre, Welkenhuysen-Gybels and Billiet's international study were mixed, as some countries better fit base model B better and others fit model 1 better, although none of the countries assessed were the United States. The current study supports and furthers these findings, as acquiescence differed based upon some group memberships and by scale content. For the balanced survey, the physical engagement and cognitive engagement: in class scales

provided evidence for base model B. However, the all positively and all negatively worded items supported model 1 on some content scales, but not others.

Motl, Conroy, and Horan (2000) and Distefano and Motl (2006) found that a single content factor with two acquiescence factors, one for each set of polarized questions, had the best fit, shown in Figure 21.

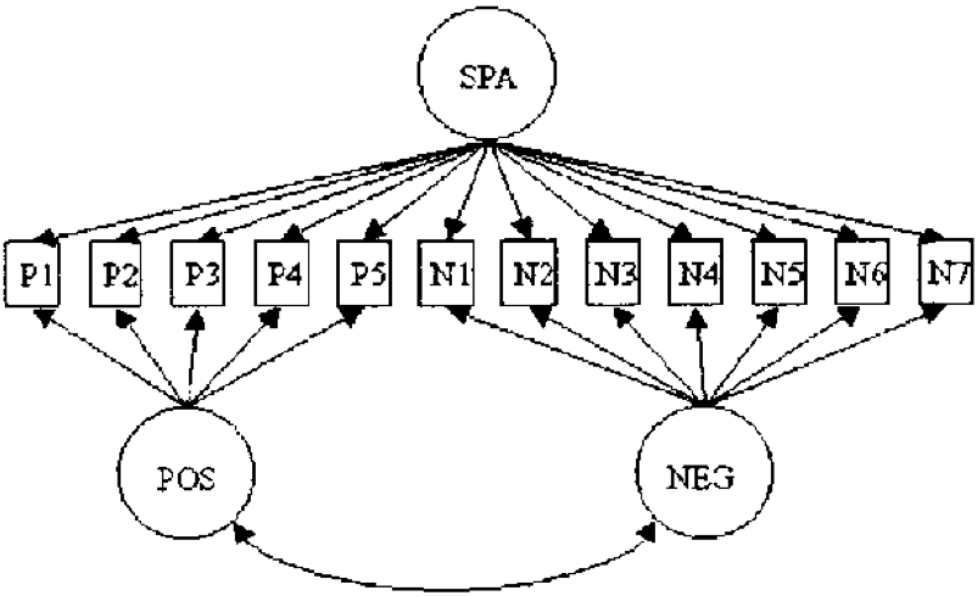


Figure 21. Motl, Conroy, and Horan (2000) study model (p. 337)

The current study's emotional engagement scale originally supported this assertion, until the exploratory factor analysis revealed that splitting the factor differently created better fit and adding a latent acquiescence factor worsened the fit. This finding reflects that confirmatory factor analysis is only as accurate as the model specified. If a model is misspecified, where the best fitting model is not tested, then the results of the confirmatory factor analysis are not

accurate. Acquiescence was originally found when using confirmatory factor analysis to test the emotional engagement scale, but this was an erroneous conclusion, as the best fitting model was not tested. In this instance, acquiescence was falsely assumed to be a method effect, when, in actuality, acquiescence was not present. A correctly specified model is essential for accurate model testing. Researchers may want to consider using both exploratory and confirmatory factor analyses as part of confirmatory studies to ensure that the best models are being tested.

Unlike in Horan, DiStefano, and Motl's (2003) later study, the presence of the acquiescence effects was not consistent across scales. Instead, the current study showed that acquiescence differed dramatically, and in some scales was even non-existent, dependent on the content of the scale. Given the recency of studies assessing acquiescence using structural equation modeling, further research is clearly necessary to ascertain how best to model acquiescence and get closer to the respondents' true scores or if survey design techniques can be developed to prevent acquiescence.

Differences by Demographic Group

Supporting the findings from Falthzik and Jolson (1974) and DiStefano and Motl (2009), respondents did not differ by demographic group on the models including acquiescence. However, more research is needed to confirm these findings, given the small sample and lack of convergence in some of the between-groups acquiescence models. Unlike Bachman and O'Malley's (1984) study, which showed that blacks were more likely than whites to acquiesce, whites versus non-whites in the current study did not differ in their acquiescence. However, the current study included all minority races in the non-white category. It is possible that a larger study could directly compare blacks to whites. The current study also did not support the

assertion that acquiescence is a heritable characteristic, as acquiescence did not differ by group membership. However, given that group membership was not comprised of people with familial relationships, it is possible that family members might acquiesce similarly. Future researchers might consider assessing acquiescence amongst those with familial relationships, in order to assess the stability of acquiescence across generations.

Conclusions

This study supports prior research showing that acquiescence is an important method effect that needs to be evaluated in order to eliminate one form of bias. However, this study shows that acquiescence is not consistent across content, making it difficult to generalize about acquiescence as a whole or make concrete recommendations to researchers on how to design surveys that either prevent acquiescence bias or statistically control for it. Instead, all survey researchers should evaluate acquiescence in their scales and populations before analyzing the results of the content of the study.

Further, additional research is needed to understand the interaction between item polarization and scale content. Across all three survey types, respondents acquiesced in some circumstances, but not in others, even though the item order was the same for all three surveys. The lack of similar acquiescence across scales in the same survey format further demonstrates that simply using an all positively worded, all negatively worded, or balanced survey does not eliminate acquiescence. Until a method for assessing and effectively controlling for acquiescence can be developed, researchers will need to assess each scale individually.

Implications for Future Research

Since the factor loadings in the exploratory factor analysis were consistent across all three versions of the instrument, it is possible that the emotional engagement scale of the instrument is flawed and has two items that are not consistent with the rest of the emotional engagement scale. Although this scale was previously validated, further development of the scale may be necessary for it to be consistent across different university populations. Beyond the flaw in the emotional engagement scale, this study demonstrated that acquiescence was not present in the balanced survey scales. However, it is possible that the small sample size limited the functionality of the more complex models, preventing convergence. Further research with a large sample size is necessary to determine if the lack of convergence was due to poor fit or a lack of power. A follow-up for this study would be to utilize the models developed by previous researchers to determine how they fit the data from this study.

Although this study did not find a difference in acquiescence by demographic group, the limited power and lack of acquiescence overall may have compromised the ability of this study to detect group differences. Additional research is needed to ascertain if different groups acquiesce differently. It is also possible that groups may acquiesce differently not only by group membership, but by content.

The presence or absence of acquiescence in all positively and all negatively worded scales requires more research. The research to date focused on the presence of acquiescence in balanced scales, but did not examine all positively or all negatively worded scales. The current study showed that acquiescence can be present in all positively worded and all negatively worded scales, but that acquiescence differed within the same population, depending on the content of the scale. Further, since acquiescence was present with one polarization but not the

other on the physical engagement scale and then the reverse on cognitive engagement: in class scale, more research is necessary to determine if there is an interaction between content and the scale polarization that determines whether or not respondents acquiesce.

Implications for Survey Design

The findings of this study provide limited guidance for survey designers. It is unclear how acquiescence is related to content and item polarization. Since prior researchers (e.g., Motl, Conroy, and Horan, 2000) found that acquiescence was present in balanced scales and that acquiescence was more severe for negatively worded items, it cannot be assumed that balanced scales eliminate acquiescence bias, even though acquiescence was not present on balanced version of these engagement scales. In addition, since this study found acquiescence on both the all positively and all negatively worded versions of the survey, it cannot be assumed that acquiescence is limited to all negatively worded or balanced scales. Therefore, survey designers cannot assume that using all positively worded, all negatively worded, or a balance of both will prevent acquiescence bias, even though prior literature has recommended each of these as a way to prevent acquiescence bias.

Since acquiescence bias differed by content within the same population during a survey administration, survey designers cannot assume that acquiescence bias will or will not be present in a population, just because of prior results with that population. Survey designers, as part of their pilot process for a specific survey, need to evaluate whether or not acquiescence bias is present. This presents a serious challenge for survey designers, as many designers may not have the statistical knowledge to do structural equation modeling. In that case, designers may need to use a balanced survey, so that they can compare the means of the negatively worded questions to the means of the positively worded questions. Using a *t*-test, the designer could assess whether

statistically significant differences exist between the two formats. Designers could also go back to Winkler, Kanouse, and Ware, Jr's (1982) method of creating and controlling for respondents' Acquiescence Response Set scores. Exploratory factor analyses could also demonstrate whether or not the scale was unidimensional. Although these methods are less rigorous and effective than using structural equation modeling, they could at least give survey developers an idea of what impact acquiescence might be having on their scale(s) with their specific population(s). Based upon the results of these methods, the designers could adjust their surveys before administration.

Finally, survey designers should consider using exploratory and confirmatory factor analyses as part of their confirmatory analyses, in order to ensure that the best models are being tested. Since acquiescence can be falsified by a misspecified model, survey designers need to carefully test for alternative models before asserting that acquiescence is or is not present.

Limitations

The major limitation of this study was the use of a convenience sample, as it may not have been representative of the entire undergraduate population at this institution or of college students. Although the racial diversity was representative of the university's published statistics, the gender breakdown was not. In addition, many other demographic characteristics were not evaluated that could have made this population different from the university's undergraduate population. A random sample of the entire undergraduate population would have allowed for more generalizability. Further, since this was a large, urban, public university, a sample that included other institutions that served different populations would have increased the external validity of the study.

A second limitation was the sample size. Although sufficient power was attained for the balanced survey, a larger sample could have been helpful when doing the multiple group

comparison and the comparison across survey types. In order to keep students from being identified who chose not to participate in the study, IRB required that every student in the class be given a survey. In turn, that led to unequal numbers of students completing the three survey types, especially for the negatively worded survey.

Since all respondents received the questions in the same order, it is possible that order effects were present. In addition, the emphasized negative statements, although good item development practice, could have caused other method biases. Additional research using randomized question wording and surveys with and without negative emphasis are necessary to ascertain the impact of these possible issues. Further, given that the negatively worded items were developed for this study, cognitive interviews could help understand how respondents react to the negatively worded items.

A final limitation was the measure itself. Although the wording did not impact the factor structure of the survey, the results of this study did not match the factor structure of the original, previously validated instrument, even on the all positively worded version, which was the same as the original instrument. Although the issues with the emotional engagement scale presented the finding that the presence of acquiescence can be erroneous when a model is misspecified, using a measure with more validation work may have allowed the better testing of the acquiescence models.

REFERENCES

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P. M., Barbaranelli, C., Medda, E., Nisticò, L., Stazi, M. A., & Caprara, G. V. (2010). Much more than model fitting? Evidence for the heritability of method effect associated with positively worded items of the Life Orientation Test Revised. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(4), 642-653.

Allison, P. (2001). *Missing Data*. Thousand Oaks, CA: Sage Publications.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, J. C., Gerbing, D. W., & Hunter, J. E. (1987). On the assessment of unidimensional measurement: Internal and external consistency, and overall consistency criteria. *Journal of Marketing Research, 24,* 432-437.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*, 491-509.

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.

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143-156.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*(3), 231-240.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bergstrom, B. A., & Lunz, M. E. (1998, April 13-17). *Rating scale analysis: Gauging the impact of positively and negatively worded items.* Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608-628.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley Interscience.

Burch, G. F., Heller, N. A., Burch, J. J., Freed, R., & Steed, S. (2015). Student engagement: Developing a conceptual framework and survey instrument. *Journal of Education for Business, 90*(4), 224-229.

Cabre, B., Welkenhuysen-Gybles, J., & Billiet, J. (2002). Is it content or style? An evaluation of two competitive measurement models applied to a balanced set of ethnocentrism items. *International Journal of Comparative Sociology, 43*(1), 1-20.

Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement, 55*(6), 991-997.

Cohen, G., Forbes, J., & Garraway, M. (1996). Can different patient satisfaction survey methods yield consistent results? Comparison of three surveys. *British Medical Journal, 313*(7061), 841-844.

de Vaus, D. (2014). *Surveys in social research* (6th ed.). New York, NY: Routledge.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys.

DiStefano, C., & Motl, R. W. (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.

Dr. Seuss (1960). *Green eggs and ham.* New York: Random House.

Falthzik, A. M., & Jolson, M. A. (1974). Statement polarity in attitude studies. *Journal of Marketing Research, 11,* 102-105.

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research, 38*(3), 353-374.

Garson, D. G., & Statistical Associates Publishing. *Missing Values Analysis and Data Imputation*. Retrieved from http://www.statisticalassociates.com/missingvaluesanalysis_p.pdf.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*, 186-192.

Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation and the Health Professions, 15*(2), 198-220.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Heaven, P. C. L. (1983). Authoritarianism or acquiescence? South African findings. *The Journal of Social Psychology, 119*, 11-15.

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science, 24*(4), 366-374.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling: A Multidisciplinary Journal, 10*(3), 435-455.

Horn, L. R. (1989). *A natural history of negation.* Chicago: University of Chicago Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 1-55.

Hughes, G. D. (2009). The impact of incorrect responses to reverse-coded survey items. *Research in the Schools*, *16*(2), 76-88.

Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88,* 497-500.

Identifying Reference (2016).

Indiana University School of Education (2016). *National Survey of Student Engagement.* Retrieved from http://nsse.indiana.edu/

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York, NY: The Guilford Press.

Kunda, Z., & Fong, G. T. (1993). Directional questions direct self-conceptions. *Journal of Experimental Social Psychology, 29*, 63-86.

Lam, T. C. M., & Stevens, J. J. (1994). Effects of content polarization, item wording, and rating scale width on rating response. *Applied Measurement in Education, 7*(2), 141-158.

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.

Marsh, H. W. (1986). *Multidimensional self concepts: Do positively and negatively worded items measure substantively different components of self*. Author.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810-819.

McMillan, J. H. (2012). *Educational research: Fundamentals for the consumer* (6th ed.). Boston, MA: Pearson Education, Inc.

McPherson, J., & Mohr, P. (2005). The role of item extremity in the emergency of keying-related factors: An exploration with the life orientation test. *Psychological Methods, 10*(1), 120-131.

Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly, 14*(3), 31-36.

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) Affective Commitment Scale: Effects of negatively worded items. *Journal of Business Psychology, 27,* 421-436.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 x 2 index. *Social Psychology Quarterly, 54*(2), 127-145.

Mitchell, M. L., & Jolley, J. M. (2013). *Research design: Explained* (8th ed.). Belmont, CA: Wadsworth.

Motl, R. W., Conroy, D. E., & Horan, P. M. (2000). The Social Physique Anxiety Scale: An example of the potential consequence of negatively worded items in factorial validity studies. *Journal of Applied Measurement, 1*(4), 327-345.

Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 562-578.

Nardi, P. M. (2006). *Doing survey research: A guide to quantitative methods* (2nd ed.). Boston, MA: Pearson Education, Inc.

Nunnally, J. C. (1967). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill, Inc.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & 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.

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://quantpsy.org/.

Ray, J. 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.

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

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

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

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*, 116-131.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Boston, MA: Pearson Education, Inc.

Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development, 34,* 166-176.

Weems, G. H., Onwuegbuzie, A. J., & Collins, K. M. T. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation and Research in Education, 19*(1), 3-20.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320-334.

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing, 26*, 2-12.

Wheaton, B., Muthen, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D. Heise (Ed.), *Sociological methodology* (pp. 84-136). San Francisco, CA: Jossey-Bass.

Winkler, J. D., Kanouse, D. E., & Ware, Jr, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67*(5), 555-561.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: Mesa Press.

Wong, N., Rindfleisch, A., Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research, 30,* 72-91.

Yorke, M. (2009). 'Student experience' surveys: Some methodological considerations and an empirical investigation. *Assessment and Evaluation in Higher Education, 34*(6), 721-739.

APPENDIX – SURVEYS

Balanced Survey

INTRODUCTION – IRB STUDY # HM20008570

You are invited to participate in a research study to enhance good research practices by studying student engagement in the classroom. **This survey contains 28 questions and should take less than 10 minutes to complete. Your responses will be anonymous.** This is not an evaluation of your instructor and will not affect your grade for this class; however, your honest assessment of your engagement in this class could help further research at VCU. An anonymous, aggregated report of student engagement may be provided to your instructor for informational purposes, but all demographic indicators will be removed.

Participation in this study is voluntary. I hope you will choose to participate, as participation is important for obtaining valid results. You do not have to participate in this study. If you choose to participate, you may stop at any time without any adverse consequences. You may also choose not to answer particular questions in the survey.

If you have any questions or concerns about your participation in this research, please contact: Amy Hutton at achutton@vcu.edu. If you have any general questions about your rights as a participant in this or any other research, you may contact: Office of Research, Virginia Commonwealth University, Telephone: (804) 827-2157

Thank you for your help! It is much appreciated and will help advance knowledge about how to conduct good research.

DEMOGRAPHICS

What is your age?

- ○ 17 and under
- ○ 18 to 23
- ○ 24+

How do you describe yourself? (check one)

- ○ Female
- ○ Male
- ○ Transgender
- ○ Do not identify as female, male, or transgender

Are you Hispanic or Latino?

- ○ Hispanic or Latino
- ○ Not Hispanic or Latino

In addition, select one or more of the following racial categories to describe yourself:

- ○ American Indian or Alaska Native
- ○ Asian
- ○ Black or African American
- ○ Native Hawaiian or Other Pacific Islander
- ○ White

REFLECTION ON THIS CLASS

To what extent do you agree or disagree with the following statements. When I am in the classroom for this class….

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
My mind is focused on class discussion and activities.					
I DO NOT pay a lot of attention to class discussion and activities.					
I focus a great deal of attention on class discussion and activities.					
I am NOT absorbed by class discussion and activities.					
I concentrate on class discussion and activities.					
I DO NOT devote a lot of attention to class discussion and activities.					

Please rate to what extent you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
I am NOT enthusiastic about this class.					
I work with intensity on assignments for this class.					
I DO NOT feel energetic when I am in this class.					
I exert my full efforts towards this class.					
I am NOT interested in material I learn in this class.					
I devote a lot of energy toward this class.					
I am proud of assignments I complete in this class.					
I DO NOT try my hardest to perform well for this class.					
I feel positive about the assignments I complete in this class.					
I DO NOT strive as hard as I can to complete assignments for this class.					
I am excited about coming to this class.					
I DO NOT exert a lot of energy for this class.					

To what extent do you agree or disagree with the following statements. When I am reading or studying material related to this class….

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
My mind is NOT focused on class discussion and activities.					
I pay a lot of attention to class discussion and activities.					
I DO NOT focus a great deal of attention on class discussion and activities.					
I am absorbed by class discussion and activities.					
I DO NOT concentrate on class discussion and activities.					
I devote a lot of attention to class discussion and activities.					

Negatively Worded Survey

N

INTRODUCTION – IRB STUDY # HM20008570

You are invited to participate in a research study to enhance good research practices by studying student engagement in the classroom. This survey contains 28 questions and should take less than 10 minutes to complete. Your responses will be anonymous. This is not an evaluation of your instructor and will not affect your grade for this class; however, your honest assessment of your engagement in this class could help further research at VCU. An anonymous, aggregated report of student engagement may be provided to your instructor for informational purposes, but all demographic indicators will be removed.

Participation in this study is voluntary. I hope you will choose to participate, as participation is important for obtaining valid results. You do not have to participate in this study. If you choose to participate, you may stop at any time without any adverse consequences. You may also choose not to answer particular questions in the survey.

If you have any questions or concerns about your participation in this research, please contact: Amy Hutton at achutton@vcu.edu. If you have any general questions about your rights as a participant in this or any other research, you may contact: Office of Research, Virginia Commonwealth University, Telephone: (804) 827-2157

Thank you for your help! It is much appreciated and will help advance knowledge about how to conduct good research.

DEMOGRAPHICS

What is your age?

- ○ 17 and under
- ○ 18 to 23
- ○ 24+

How do you describe yourself? (check one)

- ○ Female
- ○ Male
- ○ Transgender
- ○ Do not identify as female, male, or transgender

Are you Hispanic or Latino?

- ○ Hispanic or Latino
- ○ Not Hispanic or Latino

In addition, select one or more of the following racial categories to describe yourself:

- ○ American Indian or Alaska Native
- ○ Asian
- ○ Black or African American
- ○ Native Hawaiian or Other Pacific Islander
- ○ White

N

REFLECTION ON THIS CLASS

To what extent do you agree or disagree with the following statements. When I am in the classroom for this class….

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
My mind is NOT focused on class discussion and activities.					
I DO NOT pay a lot of attention to class discussion and activities.					
I DO NOT focus a great deal of attention on class discussion and activities.					
I am NOT absorbed by class discussion and activities.					
I DO NOT concentrate on class discussion and activities.					
I DO NOT devote a lot of attention to class discussion and activities.					

Please rate to what extent you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
I am NOT enthusiastic about this class.					
I DO NOT work with intensity on assignments for this class.					
I DO NOT feel energetic when I am in this class.					
I DO NOT exert my full efforts towards this class.					
I am NOT interested in material I learn in this class.					
I DO NOT devote a lot of energy toward this class.					
I am NOT proud of assignments I complete in this class.					
I DO NOT try my hardest to perform well for this class.					
I DO NOT feel positive about the assignments I complete in this class.					
I DO NOT strive as hard as I can to complete assignments for this class.					
I am NOT excited about coming to this class.					
I DO NOT exert a lot of energy for this class.					

To what extent do you agree or disagree with the following statements. When I am reading or studying material related to this class….

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
My mind is NOT focused on class discussion and activities.					
I DO NOT pay a lot of attention to class discussion and activities.					
I DO NOT focus a great deal of attention on class discussion and activities.					
I am NOT absorbed by class discussion and activities.					
I DO NOT concentrate on class discussion and activities.					
I DO NOT devote a lot of attention to class discussion and activities.					

Positively Worded Survey

INTRODUCTION – IRB STUDY # HM20008570

You are invited to participate in a research study to enhance good research practices by studying student engagement in the classroom. **This survey contains 28 questions and should take less than 10 minutes to complete. Your responses will be anonymous.** This is not an evaluation of your instructor and will not affect your grade for this class; however, your honest assessment of your engagement in this class could help further research at VCU. An anonymous, aggregated report of student engagement may be provided to your instructor for informational purposes, but all demographic indicators will be removed.

Participation in this study is voluntary. I hope you will choose to participate, as participation is important for obtaining valid results. You do not have to participate in this study. If you choose to participate, you may stop at any time without any adverse consequences. You may also choose not to answer particular questions in the survey.

If you have any questions or concerns about your participation in this research, please contact: Amy Hutton at achutton@vcu.edu. If you have any general questions about your rights as a participant in this or any other research, you may contact: Office of Research, Virginia Commonwealth University, Telephone: (804) 827-2157

Thank you for your help! It is much appreciated and will help advance knowledge about how to conduct good research.

DEMOGRAPHICS

What is your age?

- ○ 17 and under
- ○ 18 to 23
- ○ 24+

How do you describe yourself? (check one)

- ○ Female
- ○ Male
- ○ Transgender
- ○ Do not identify as female, male, or transgender

Are you Hispanic or Latino?

- ○ Hispanic or Latino
- ○ Not Hispanic or Latino

In addition, select one or more of the following racial categories to describe yourself:

- ○ American Indian or Alaska Native
- ○ Asian
- ○ Black or African American
- ○ Native Hawaiian or Other Pacific Islander
- ○ White

P

REFLECTION ON THIS CLASS

To what extent do you agree or disagree with the following statements. When I am in the classroom for this class….

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
My mind is focused on class discussion and activities.					
I pay a lot of attention to class discussion and activities.					
I focus a great deal of attention on class discussion and activities.					
I am absorbed by class discussion and activities.					
I concentrate on class discussion and activities.					
I devote a lot of attention to class discussion and activities.					

Please rate to what extent you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
I am enthusiastic about this class.					
I work with intensity on assignments for this class.					
I feel energetic when I am in this class.					
I exert my full efforts towards this class.					
I am interested in material I learn in this class.					
I devote a lot of energy toward this class.					
I am proud of assignments I complete in this class.					
I try my hardest to perform well for this class.					
I feel positive about the assignments I complete in this class.					
I strive as hard as I can to complete assignments for this class.					
I am excited about coming to this class.					
I exert a lot of energy for this class.					

To what extent do you agree or disagree with the following statements. When I am reading or studying material related to this class….

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
My mind is focused on class discussion and activities.					
I pay a lot of attention to class discussion and activities.					
I focus a great deal of attention on class discussion and activities.					
I am absorbed by class discussion and activities.					
I concentrate on class discussion and activities.					
I devote a lot of attention to class discussion and activities.					

VITA

Amy Christine (Baumgartner) Hutton was born on April 27, 1982 in Dallas, Texas and is an American citizen. She graduated from Pembroke Hill School in Kansas City, Missouri in 2000. She received her Bachelor of Musical Arts in Music with a double major in Communication Arts and Sciences from DePauw University, Greencastle, Indiana in 2004 and her Master of Fine Arts in Theatre Pedagogy from Virginia Commonwealth University, Richmond, Virginia in 2008. Her work experience includes being a professional stage manager, working on- and off-Broadway and in theatres across the United States, being an Assistant Professor of Theatre and Head of Stage Management at Virginia Commonwealth University, and serving as the Director of Admissions for the Department of Music at Virginia Commonwealth University. Amy currently works as the Senior Applications and Data Analyst for Strategic Enrollment Management at Virginia Commonwealth University. She has presented her research at local, national, and international research conferences, including conferences organized by the American Association of Collegiate Registrars and Admission Officers (AACRAO), the American Educational Research Association (AERA), the Potomac and Chesapeake Association for College Admission Counseling (PCACAC), and the Association for Theatre in Higher Education (ATHE).