Spring 2013

Student engagement in the assessment context: An examination of the Cognitive Engagement Scale-Extended Version (CES-E)

Ashley Brianne Charsha
James Madison University

Student Engagement in the Assessment Context: An Examination of the Cognitive Engagement Scale-Extended Version

(CES-E)

Ashley Charsha

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2013

Acknowledgements

I would first like to acknowledge my thesis committee, Dena Pastor and Keston Fulcher, for their thoughtful contributions in the process of completing this thesis. I would also like to give a huge acknowledgement to my advisor and committee chair, Robin Anderson, for her unwavering support and understanding throughout not only the thesis process, but throughout the last two years.

Of course, I would be remiss if I did not acknowledge the support of my fellow students during these last few years. A special thanks goes to Makayla Grays, Jeff Kosovich, Jason Kopp, Oksana Naumenko, Paula Love, and Mandy Swanson for all the encouragement they've provided during the stressful times and for making the fun times unforgettable.

Finally, I must acknowledge the support my amazing family has given me not only throughout my master's program, but throughout my entire academic career. Sometimes they believed in me more than I believed in myself, which, coupled with the family sense of humor, help me to conquer challenges I otherwise wouldn't so readily face. There are simply not enough words to express my gratitude for that.

Table of Contents

List of Tables

List of Figures

Abstract

Increasing pressure on institutions of higher education to demonstrate what students are learning has resulted in an increase in assessment testing. Because these assessments are often low-stakes for students, educators often question whether inferences based on the resulting student scores are valid. Not unexpectedly, questions often arise regarding the extent to which students are engaged on low-stakes assessments. Additionally, how their level of engagement impacts their performance is also questioned. These questions are empirical in nature. Before such questions can be examined, a psychometrically sound instrument of cognitive engagement appropriate for the assessment context must be identified. This study sought to gather validity evidence for such an instrument, the CES-E (Appendix A) using Benson's (1998) model for construct validation.

Chapter One

Introduction

Imagine two students reading their assigned text for the week. Holly uses learning strategies such as pausing to make sure she understands the material, thinking about how the new material fits in with what she already knows, and planning out answers to the associated homework. In contrast, Henry "reads" his textbook but only the first and last paragraph of each chapter and the first and last sentence of each paragraph in between. He sees studying as simply doing his homework. He completes all of his homework assignments but does so by simply trying to locate answers either in his text or online. Both students are cognitively engaged but in very different ways. While Holly is utilizing *deep* cognitive engagement learning strategies (Lublin, 2003), Henry is using *shallow* strategies (Meece, Blumenfeld, & Hoyle, 1988).

Just a few months before, Holly and Henry had arrived in a college classroom for the first time. They were not there for class, but to complete assessments designed to assess their knowledge in core areas of the institution's general education curriculum. Their university uses an assessment day model designed to assess students' learning for both the purpose of program improvement and institutional accountability. Holly and Henry took an assessment battery containing both knowledge-based and attitudinal measures. While these assessments are high stakes for the university, because it reports the results to accrediting bodies as evidence of its effectiveness, the assessments are low-stakes for Holly and Henry (e.g. don't appear on their transcripts or impact their grades).

The university randomly assigned both Holly and Henry to the same assessment room based on the last two digits of their student identification numbers. Unbeknownst to

Holly and Henry, this room assignment determined their testing configuration. This process also allows the university to track students in order to administer the same tests to each of them later in their college career. While Holly and Henry were waiting for the testing session to begin, proctors walked around the room to greet and encourage each student to do his or her best on the assessments. As instructed by the proctors, Holly and Henry both read and completed a consent form while waiting for the first assessment instrument to be distributed. At the start of the testing session, they handed in their signed consent forms and the proctors handed out the Scantron sheets needed for the assessment session. Holly and Henry both filled in the necessary information on the Scantron (i.e. ID number, test name) and the proctors read aloud the instructions for the first assessment, a test of *quantitative reasoning*. Once the lead proctor read the instructions, Holly and Henry began the assessment. Up until this point, Holly and Henry had nearly identical experiences. Holly and Henry both read the first item and marked their individual answers on their respective answer sheets with the provided pencils.

From the perspective of the proctors, both Holly and Henry appeared to give good effort on the assessment. They both appeared to be reading the items. When the proctors passed by, the proctors did not observe any specific pattern to the answers either student was providing (e.g. neither was marking all C’s). As if to confirm this, when asked to complete a scale at the end of the session regarding the effort each put into the assessment, both Holly and Henry self-reported that they had put forth good effort. However, undetected by the proctors, the two students engaged quite differently with the assessment instruments. Recall that Holly and Henry took a quantitative *reasoning* assessment. This assessment was designed to require students to reason through a variety of quantitative problems and then to record their results using a selected-response format. For example, an item may require a student to pull relevant information from a scenario and apply it to a formula to determine the velocity of a falling object. This assessment was not a simple recall test. While working through the quantitative reasoning assessment, Holly used strategies such as reviewing items multiple times before answering and “working” out her response when appropriate for the given item. In this assessment context, because she was not trying to learn anything new, Holly modified her learning strategies to help her understand what was being presented to her in order to provide the best answer. In contrast, Henry read through the material, but his effort was mostly directed at looking through the material presented in an attempt to find the correct answer. When that was not effective he would implement other strategies such as eliminating an answer or two and randomly selecting from those remaining. Henry modified his learning strategies to use mental shortcuts to complete the assessments. In other words, both Holly and Henry translated their learning strategies into test taking strategies. Such test-taking strategies can be reflective of deep (e.g. Holly) or shallow (e.g. Henry) engagement.

Holly and Henry both believed and reported that they had exerted effort when completing their assessments. After all, to their way of thinking, neither had simply randomly selected answers nor slept through the assessments like the girl they had observed in the back left corner of the room had done. However, it is clear that the strategies they used were qualitatively different. Therefore, while asking students to self-report their motivation tells us if students *believe* they demonstrated motivation in the form of effort on the assessment test, it does not give us information regarding how deeply they actually engaged with the assessment items. By examining cognitive engagement in addition to motivation, we may be more clearly able to understand how students interact with the assessment tests and how that engagement influences the validity of the inferences we make from assessment scores.

Unfortunately, cognitive engagement has not been studied in the assessment context. The majority of research related to the construct has focused on engagement related to learning in the classroom. Deep cognitive engagement in the classroom has been characterized as using learning strategies such as performing metacognitive checks and planning out answers (Lublin, 2003), such as those used by Holly. Shallow cognitive engagement has been characterized as using learning strategies such as skimming readings or simply searching for homework answers within the text (Meece et al., 1988). Students, such as Henry, expend effort to complete their academic tasks, but at only a surface level. Because so little research has been done to examine cognitive engagement in an assessment setting, the research that has been completed in the classroom setting (see Chapter 2) is important for laying out a foundation on which we can build an understanding of the construct in an assessment setting.

However, to understand 1) how the classroom research may be applied to the assessment setting and 2) the limits in the application, it is first important to understand how the two contexts relate. First and foremost, the classroom context focuses on the process of learning new material, whereas the assessment context serves the purpose of measuring what students have learned. Second, the stakes for the students can be dramatically different. In a classroom context, the students' performance is often directly tied to grades; therefore, the stakes are much higher for students than in an assessment context where their performance does not affect them personally. However, in both contexts, they interact in some way (e.g. reading, listening) with the material, then decide what, if anything, to do with that material.

Thus, in both contexts there lies the possibility that students may utilize different strategies, whether shallow or deep, that help the students to process and apply the material, whether it is a homework assignment or a knowledge-based assessment test. While there are methods for assessing classroom-based cognitive engagement, there is a noticeable lack of measures for assessing cognitive engagement in the assessment context. The lack of such measures makes it impossible for us to fully understand how students are engaging with such assessments and what impact students' cognitive engagement (or lack thereof) has on the validity of the inferences we draw from assessment findings._Therefore, it is important to develop a scale that measures cognitive engagement in a low-stakes assessment setting.

A logical step to creating such an instrument might be to adapt a cognitive engagement measure used in classrooms. Unfortunately, such existing instruments cannot be easily transferred to the assessment setting. Such classroom-based measures often pertain to only one subject area. For example, the Science Activity Questionnaire (Meece, Blumenfeld, & Hoyle, 1988) is used to assess the extent of student engagement as it pertains to a specific classroom science lesson. Such contextual misalignments make it necessary to develop a measure specific to the assessment context. Though no previously existing measure can be easily transferred in its entirety to an assessment setting, researchers have been able to draw from the classroom-based measures when developing items for the assessment context. For example, researchers developing the

Cognitive Engagement Scale (CES; Smiley & Anderson, 2011) and the Cognitive Engagement Scale-2 (CES-2; Charsha, Smiley, & Anderson, 2012) adopted several items from the Motivation and Strategy Use Survey (Greene & Miller, 1996). Like its predecessors, the Cognitive Engagement Scale-Extended (CES-E) combines items adapted from classroom-based measures with newly created items to address cognitive engagement in a low-stakes assessment setting. The following literature review examines the existing cognitive engagement measures and details the extent to which each contributed to the development of the CES-E.

The current researcher used Benson's (1998) strong program of construct validation as a framework for examining the CES-E. The program consists of three stages: substantive, structural, and external validity. The substantive stage consists of defining and operationalizing the construct of interest, including identifying other constructs in which the construct of interest is theoretically related. Included for example is a discussion of cognitive engagement and its relationship to motivation. The literature review provided in Chapter 2 summarizes the research that has been performed in regards to cognitive engagement and constitutes much of the work that represents the substantive stage of development. As mentioned previously, much of the research related to cognitive engagement has occurred in a classroom setting.

The second stage in Benson's model advocates for the importance of establishing the structure of the scale in order to understand how best to score the instrument (e.g. single score versus sub scores). The theory of cognitive engagement (Greene & Miller, 1996; Meece et al., 1998; McLaughlin, 2005) advocates for two levels of engagement, shallow and deep. However, there is no consensus on whether there is a continuum between these facets (i.e. students engage in one type at the expense of engaging in the other), or whether there is no continuum (i.e. students can demonstrate both in the same context). Greene and Miller (1996) found through path analysis that meaningful engagement (i.e. deep engagement) suppressed the effects of shallow engagement when predicting achievement in the classroom. This finding suggests that the construct of cognitive engagement may be unidimensional. However, other research findings suggest that the construct may be multidimensional. For example, a previous iteration of the Cognitive Engagement Scale suggests a third aspect of cognitive engagement; no engagement (Charsha, Smiley & Anderson, 2012). After initially testing a two factor (deep and shallow) model that did not fit the data, Charsha et al. discovered through an examination of fit indexes and item correlations that two of the items designed to assess shallow engagement actually formed their own factor. Item content indicated that the two items were more representative of none or no engagement than shallow engagement. Thus, the researchers modified the model post hoc and tested a three-factor model (deep, shallow and no) and the model was found to fit the data. However, given that the two factor model was modified post hoc and capitalizes on chance (MacCallum & Austin, 1992), the three factor model needs to be replicated on an independent sample. Additionally, the shallow and no engagement scales contained only two items each; therefore, they were not likely to be representative of their respective factors. Consequently, two new items were added to the shallow factor, and one item added to the no engagement factor to better represent the breadth of the facets. Because the possibility still exists that cognitive engagement is a construct that lies on a continuum, both a unidimensional model and a three factor model will be tested through confirmatory factor analyses as part of the current study.

To clarify how a three-factor model would differ from a one-factor model, Holly and Henry are revisited. In addition, a student named Heidi is introduced. Heidi was also in the same testing session and room with Holly and Henry, but she spent most of the session napping, undeterred by the proctors' insistence that she try her best. In the little bit of time that she paid attention to the assessments, she utilized strategies such as choosing answers randomly. If a question looked remotely difficult, Heidi did not bother to consider it and simply left the answer blank on her Scantron. Therefore, Heidi can be described as having no cognitive engagement during the session. If a three-factor model of the CES-E is supported, each of the three students will have a "profile" of cognitive engagement, which means that it is possible for Holly, Henry, and Heidi to endorse test-taking strategies that represent different factors. However, each student is likely to endorse more strategies on one factor over the other factors. For example, a three-factor model would indicate that it is possible for Holly to use test-taking strategies that demonstrate deep engagement, but she may still endorse a few strategies that indicate shallow engagement. Additionally, Henry uses shallow test-taking strategies for the most part, but may use a few deep or no engagement strategies as well. Finally, Heidi could use a few shallow strategies, but for the most part uses strategies that indicate no engagement. In contrast, if the scale is unidimensional, that would indicate a continuum that runs from deep to shallow to no engagement. As a result, Holly would fall on the end that indicates deep engagement, Henry would fall in the middle with shallow engagement, and Heidi would fall on the other end of no engagement.

If a factor structure is supported, the third and final step, according to Benson's model would involve examining external validity. This stage concerns testing whether or not the construct is related to other constructs in expected theoretical ways as outlined in the substantive stage. Previous research has provided possible theoretical relationships to test. For example, Greene, Miller, Crowson, Duke, and Akey (2004) conducted a path analysis in which they found that strategy use (using items reflecting deep engagement) was predicted by self-efficacy, mastery goals, and perceived instrumentality (i.e. the recognition of the instrumental relationship between an activity and the attainment of a personal goal). This finding suggests that students who believe that they are able to do the work, seek to master material, and value the activities needed to meet a goal are more likely to deeply engage with the material. For the current study, assuming a scoreable solution is supported, CES-E scores will be correlated with scores from three measures that align with previous research: the Academic Entitlement Questionnaire (Appendix B), the Expectancy-Value Cost Scale (EVC; consisting of three subscales that measure Expectancy, Value, and Cost. See Appendix C), and the effort subscale on the Student Opinion Survey (Appendix D). It is expected that deeper cognitive engagement will be positively correlated with the effort, expectancy scores, and value scores. It is also expected that deep cognitive engagement will be negatively correlated with cost and academic entitlement. Each of these constructs, their theoretical relationship to cognitive engagement and the instruments used to measure each are discussed further in Chapter 2.

Thus, it is expected that Holly would score high on the expectancy and value subscales of the EVC scale and on the effort subscale of the motivation measure, while scoring low on the cost subscale of the EVC scale and on academic entitlement.

Conversely, Henry would be expected to score on the lower end of expectancy and value, and higher on cost and academic entitlement. When it comes to effort, Henry might still believe that he put in effort (as described earlier), but would still be expected to score lower than Holly because of the mental shortcuts he used throughout the testing session. However, Heidi would be expected to score low on expectancy, value, and effort, and high on cost and academic entitlement.

The purpose of this study was to gather construct validity evidence for the CES-E. First, the model-data fit of the CES-E was evaluated using confirmatory factor analysis (CFA). Two a priori models were tested using two independent samples: (a) a one-factor (unidimensional) model and (b) a three-factor model. This study also examined whether the CES-E scores relate to other constructs in theoretically predicted ways. If supported, revised versions of the CES-E may be employed by future researchers to answer empirical questions regarding how students engage with material in a low-stakes assessment context and how engagement impacts performance

Chapter 2

Literature Review

The purpose of the presented study was to gather validity evidence on the Cognitive Engagement Scale – Expanded (CES-E) through an examination of the instrument’s factor structure. The researcher also planned to investigate whether scores on the CES-E related to scores on other constructs in anticipated ways. This evidence is essential to establish confidence when it comes to making inferences from the CES-E scores. Having confidence in the CES-E scores would allow researchers to examine other empirical questions such as the degree to which students are cognitively engaged while taking low-stakes assessments, and how students’ level of cognitive engagement may impact performance on low-stakes assessment tests. The assessment context is important because nearly all of the existing literature on cognitive engagement examines cognitive engagement in a classroom setting. The CES-E was specifically developed for use in an assessment setting in which the process is low-stakes for students, but high-stakes for the university. The presented study employed Benson’s (1998) model of a strong program of construct validation as the framework for examining the validity of inferences made based on CES-E scores. In an attempt to summarize work done by previous theorists and researchers who have contributed to the substantive stage (Benson, 1998) of the CES-E’s development, this chapter will examine the various types of student learning engagement, definitions of cognitive engagement, and previous attempts to measure students’ cognitive engagement. However, before making the case for examining cognitive engagement it is important to note that researchers have conducted a great deal of research aimed at better understanding test taking behavior and attitudes in the assessment context. The presented examination of cognitive engagement is meant to fit within this body of literature and to provide a more comprehensive picture of how students behave in a low-stakes assessment context.

Understanding testing taking behavior and attitudes

Researchers have examined a variety of factors that affect student performance in assessment settings. These factors include test taking context (low-stakes versus high-stakes), level of reporting results (in aggregate versus to individuals), proctor behavior, and students' attitudes toward assessments. Much of the previous research examines the effects of these factors on student motivation and thus performance.

Testing context

Research comparing high-stakes and low-stakes testing conditions has highlighted the impact stakes can have on student performance as well as the psychometric properties of assessment measures. For example, Barry and Finney (2009) examined the effects of testing context in low-stakes assessments situations and how it affected the psychometric properties of a non-cognitive measure that assessed college self-efficacy. Five groups of students participated in the study: two groups of incoming freshmen who completed the measure on their own time, unsupervised, the summer before their arrival to the university; one group of upperclassmen who completed the measure in a small classroom containing about 20 students and closely monitored by a trained proctor; one group of upperclassmen who completed the measure in a large classroom (ranging from 63 to 250 seats) and monitored by trained proctors; and one group of upperclassmen who completed the measure in a large classroom, monitored by proctors, but item order on the measure was randomized (in the other four groups, items were not randomized). Barry and Finney found that there was less misfit associated with the factor structure as groups became more controlled (i.e. trained proctors present, students in small classrooms). Thus, the authors recommended that in order to make correct decisions about instrument development, it is essential to place students in a setting that it is as controlled as possible.

Napoli and Raymond (2004) developed a measure meant to assess basic knowledge of psychology. In order to collect reliability and validity evidence, the researchers examined two groups of students: an “ungraded” group that was not linked to course outcomes and a “graded” group that was. Researchers found that in the ungraded condition, test scores were much less reliable than in the graded condition. In addition, the students in the graded condition scored significantly and practically higher (*d* = 1.27) than those in the ungraded condition on the assessment. The researchers concluded that linking assessments to course outcomes provided a much more accurate picture of student ability than not linking the assessment.

Finally, DeMars (2000) analyzed math and sciences scores from a high school proficiency test. The high school students either completed the proficiency test in a low-stakes condition (i.e. they were told the results did not affect them in any way) or a high-stakes condition (i.e. received an “endorsement” toward their diploma if they scored high enough to demonstrate proficiency). Both the math and science portions contained a mix of multiple choice and constructed response items. DeMars found that when the stakes of the test were increased, scores increased, though they increased more for constructed response items than for multiple choice items. Thus, the results of the study would seem to suggest that students are thinking differently about low-stakes tests versus high-stakes tests. DeMars suggested differences in motivation as a possible reason scores differed between the two contexts. These studies demonstrate that the testing context can impact both the psychometric properties of the measure and students' performance.

Assessment score reporting

Research on the effects of score reporting suggests that when students know they will receive individual feedback, they will perform better on their assessments. For example, Sundre, Erb, and Russell (2009) examined two groups of students who completed a measure of quantitative and scientific reasoning during a required university assessment day. Members of one group (*N*=218) received instructions that indicated they would each receive their individual score. In addition, students were told that they would also have the opportunity to view their scores against a faculty set standard as well as in comparison to other students. The second group of students (*N*=316) did not receive these instructions. The two groups did vary significantly in terms of performance with the experimental group (i.e. the ones who were instructed that they would receive feedback) performing significantly better on the assessment than the control group. It should be noted that the effect size between the groups (*d*=.21) was relatively small. However, considering that the manipulation was subtle, it is an interesting finding and suggests that stronger manipulations could have a bigger effect on motivation and performance in the future.

Interestingly, the two groups did not differ significantly on their level of self-reported motivation. One possible explanation is that reporting students' individual scores impacted students' cognitive engagement rather than their motivation. However, before such a hypothesis can be tested, a psychometrically sound measure of cognitive engagement must first be identified.

Proctor behavior

Another area of study that has attempted to add to our understanding of students' test taking behavior in low-stakes assessment settings is the examination of the impact of proctor behavior. Lau, Swerdzewski, Jones, Anderson, and Markle (2009) observed that when proctors adhered to standardized procedures, students reported that they put more effort into the assessments. Lau et al. (2009) outlined eight strategies that proctors should implement to further motivate students in low-stakes assessment contexts. These strategies include conveying the importance of the assessments, thanking students for the effort they put into the assessments, and modeling a positive attitude toward the assessments, the students, and the institution. The researchers implemented these strategies through proctor training at their institution starting in the fall of 2007. The researchers then compared student effort before the new strategies were integrated ("traditional") and after the strategies were implemented ("strategic") for both first-year, incoming students and sophomores. The researchers found that effort did increase across groups after the new strategies were implemented. These findings not only illustrate that scores can be influenced by factors within the testing session, but that test administrators can intervene and impact motivation and performance. However, like other studies that examine only motivation, it is still not known what students mean when they indicate that they put forth good effort on the assessments. The current study aims to develop a measure that could be used alongside measures of motivation in studies such as Lau et al. (2009).

Student attitudes

Research has also examined the impact of student attitudes toward low-stakes assessments on student scores. Brown and Hirschfeld (2008) conducted a study with 3,504 high school students in New Zealand in which the researchers examined four aspects of student attitudes toward assessment: ‘student accountability’ (i.e. assessment makes students more accountable), ‘school accountability’ (i.e. assessment keeps the school accountable), assessment is fun (i.e. assessment is engaging or fun), and ‘assessment is ignored’ (i.e. personal results are ignored). Their findings indicated that the students who held the attitude that assessment makes students more accountable tended to perform better than students who scored higher on items that addressed the other three attitudes. The researchers concluded that the findings were consistent with the literature that indicated students are more likely to perform better on educational outcomes if they believe they are accountable. Another study conducted by Zilberberg, Brown, Harmes, and Anderson (2009) involved assembling a focus group of six undergraduate students scheduled to attend a university required assessment day in which the results produced were low-stakes for students. The purpose of the focus group was to gather student feedback regarding their attitudes toward the assessments. The researchers began by asking the students to individually answer two questions: 1) “Does the assessment of the university programs benefit you and if so, how?” and 2) “How did you find out about the assessment day?” (p. 265). After the students shared their answers with the rest of the group, the researchers asked more probing questions addressing topics such as students’ perception of the purpose of assessment. Student responses were coded and organized into different themes. Ultimately, two global themes emerged from the findings: the first is lack of motivation, which was characterized by test frustration, lack of knowledge about assessment, and the awareness of the low-stakes nature of the test. The second global theme that emerged was communication to students about assessment. Responses focused largely on the fact that students wanted to know more about certain aspects of assessment and how they wanted that information communicated. The fact that students were extremely vocal about their feelings about assessment, combined with several other findings, led the researchers to conclude that administrators of assessment measures should not make assumptions about students' attitudes toward assessment or about how those attitudes may impact student performance.

In response, researchers have worked to develop scales to measure students' knowledge of accountability tests and students' attitudes toward institutional accountability testing. In regards to students' knowledge of accountability tests, Zilberberg, Anderson, Swerdzewski, Finney, and Marsh (2012) conducted a study with the intent of assessing college students' understanding of testing associated with institutional accountability mandates. The researchers developed a nine item multiple choice measure that addressed the "what" (i.e. the goal and purpose), "who" (e.g. those responsible for selecting test content, who sets standards) and "how" (e.g. reporting requirements) of accountability testing. Item-level results indicated that students, both freshmen (N=3196) and sophomores (N=382) had difficulty in identifying the correct answer for all nine items. Students had particular difficulty in identifying the purpose of institutional accountability testing and the factors which were used to evaluate the effectiveness of schools. Furthermore, students were asked to rate their confidence in their answers for each item and results indicated that for four of the items, which corresponded to all three of the types of knowledge meant to be assessed by the scale (i.e. the “what”, “who, and “how), differences between students who got the items correct and those who did not showed only negligible differences in confidence ratings. The researchers concluded that students tend to be confident in their beliefs about accountability testing even when they are wrong. As a result, suggestions for future research included the continuing development of a scale that assesses student knowledge about accountability testing in order to address more misconceptions and eventually provide solutions to eradicate the problem of misinformation.

Students’ attitudes toward institutional accountability testing have been examined as well. Zilberberg, Anderson, Finney, and Marsh (under review) conducted a study in which the purpose was to gather reliability and validity evidence for two measures: one designed to measure college students’ attitudes toward K-12 accountability testing (SAIAT-K-12), and one designed to measure college students’ attitudes toward higher education accountability testing (SAIAT-HE). For the SAIAT-K-12, entering college students filled out the assessment and factor analyses indicated that items on the measure formed four distinct factors: Validity (i.e. perceived quality of the test), Parents (i.e. how supportive parents were when it came to students taking assessments), Purpose (i.e. knowing the purpose of accountability tests), and Disillusionment (i.e. student perception that there are too many assessments). Scores were calculated for each of the factors and results indicated that students tended to hold negative to neutral opinions about Validity of assessments, Parents tended to exhibit concern about the assessments their children were taking, and students were not sure about the Purpose of the assessments. Finally, students scored fairly high on the Disillusionment scale, indicated a negative attitude toward having to take so many assessments.

The SAIAT-HE was distributed to college students halfway through their college careers. Results indicated that a six-factor model fit the data: the same four factors as described on the SAIAT-K-12, and the additional two factors of Professor (i.e. students' perceptions of the professors' view of assessments) and Students (i.e. how fellow students feel about assessments). However, items written to address the last two factors tended to have floor or ceiling effects, and thus the researcher pointed out the need to revise the items. Thus, there is more support for the use of the common four factors (Validity, Parents, Purpose, and Disillusionment) to assess students' attitudes. Students tended to have a negative to neutral attitude about the Validity of assessments, were not sure about their Purpose, and were fairly high in Disillusion, indicating that they have a more negative attitude toward higher education accountability tests.

Because findings suggest students tend to have negative attitudes toward accountability testing, research has attempted to link attitudes to motivation and performance on the assessments. Zilberberg, Finney, Marsh, and Anderson (in progress) gathered data for the SAIAT-K-12 from a large group of incoming college students as part of their assessments during a university required assessment day. Recall that the SAIAT-K-12 addresses students' attitudes toward K-12 accountability testing, and thus the topic would be familiar to incoming college students. In addition, students completed a measure of motivation that addressed how much effort students believed they put into the assessments and how important they perceived the assessments to be. The researchers used a path analysis on a subset of this data (*N*= 874) to test several relationships among the four factors of student attitudes (Disillusionment, Validity, Parents, and Purpose), motivation (importance and effort), and achievement (defined as students' scores on a multiple choice quantitative and scientific reasoning assessment). SAT math scores were also incorporated into the model to predict performance on the quantitative and scientific reasoning assessment. Results indicated that attitudes that students formed during their K-12 years had a negligible effect on motivation and performance, which may be positive news for practitioners considering that students tend to hold more negative views about assessment as they progress through their college careers. However, more studies need to be conducted in order to address the question of how students' attitudes impact motivation and performance, if they do at all. Additional results from the study indicate that test-taking effort was the best predictor of performance after SAT, with importance influencing effort. Thus, the researchers propose that if results replicate in the future, some kind of emphasis be placed on the importance of the assessment in order to increase effort, and thus performance. Like other studies reviewed above, including a measure of cognitive engagement may give practitioners a more clear picture of what students mean when they report their levels of effort in regards to completing assessments.

Motivation

Sundre (2006) describes motivation within a low-stakes assessment context as consisting of two components: effort (i.e. how much mental effort was used in answering test items) and importance (i.e. how important students think it is to do well on the assessments). Both of these components stem from the expectancy value-theory of motivation (Eccles et al., 1983; Pintrich & De Groot, 1989). According to the theory, one's expectancy is how well one expects to do on a task based on his or her perceived competency, and the value portion of the theory addresses how important, interesting, or useful one perceives the task to be. Applied as test-taking motivation to a low-stakes assessment context, expectancy is conceptualized as the amount of mental effort involved in answering a test item. Value is conceptualized as how important it is to students to do well on the assessments. Thelk, Sundre, Horst, and Finney (2009) reported results from several studies that used a measure developed to address both importance and effort. The researchers determined how important students perceived the test to be and how much effort they put forth dramatically affected test performance.

In conclusion, a large amount of research has been conducted in an effort to better understand student test-taking behavior and attitudes. Much of this research has attempted to link various factors to motivation and ultimately performance in low-stakes assessment contexts. However, what strategies students use when putting forth effort is still unknown. By also including a measure of cognitive engagement as a part of such studies, researchers may develop a more complete understanding of student behavior and its impact on scores.

Student Learning Engagement

The term “engagement” as it applies to student learning has multiple forms. Fredricks, Blumenfeld, and Paris (2004) conducted a literature review of student engagement in the classroom and concluded that three types are prevalent among researchers: emotional, behavioral, and cognitive. Emotional engagement is described as students’ feelings in the classroom. Examples of such feelings include interest, boredom, happiness, sadness, and anxiety (Connell & Wellborn, 1993; Skinner & Belmont, 1993). Researchers of emotional engagement view the construct in several different ways. For example, Finn (1989) characterizes a student who is emotionally engaged as one who feels like he or she belongs to the school and values the school, whereas other researchers (Lee & Smith, 1995; Stipek, 2002) view emotional engagement as something that students can have not only toward the school as a whole, but toward their individual teachers as well. Emotional engagement, defined as university mattering (France & Finney, 2009) and sense of belonging (Young & Finney, 2007), has been well-researched, and is already assessed as part of the university's assessment day process. As a result, this will not be further examined in the current study.

The construct of behavioral engagement is also viewed differently by various researchers. Fredricks et al. (2004) point out that three common views emerge from the behavioral engagement literature. One definition characterizes behavioral engagement as following the rules and the avoidance of disorderly behaviors such as skipping school (Finn, 1993). A second definition defines behavioral engagement as the act of participating in school activities, such as sports teams (Finn, 1993; Finn et al., 1995). The third definition stresses involvement in learning. For example, a student who displays behaviors such as putting effort into assignments, paying attention in class, and contributing to class discussions is viewed as behaviorally engaged (Birch & Ladd, 1997; Finn et al., 1995; Skinner & Belmont, 1993). Behavioral engagement can be seen as an outcome manifestation of cognitive engagement. In other words, students who appear to observers to be "concentrating" on a task (behavioral engagement) could be using multiple learning strategies (cognitive engagement) in order appear that way. While behavioral engagement is an external observable process, cognitive engagement is an internal one.

The literature that defines cognitive engagement emerges from two areas of study. One area, school engagement, stresses that cognitive engagement is characterized by a psychological investment in learning, going beyond requirements, and desiring challenges (Fredricks, Blumenfeld, & Paris, 2004). The second area, which comes from the learning literature, characterizes cognitive engagement as a utilization of learning strategies (Fredricks et al., 2004). Further exploration of the underlying arguments of each is presented below.

The area of school engagement stresses that cognitive engagement is primarily one's psychological investment in learning. Newmann, Wehlage, and Lamborn (1992) define cognitive engagement as “the student's psychological investment in and effort directed toward learning, understanding, mastering the knowledge, skills or crafts that academic work is intended to promote” (p. 12). Marks (2000) defines cognitive engagement in a largely similar way stating it is “a psychological process, specifically, the attention, interest, investment and effort students expend in the work of learning” (p.155). Wehlage, Rutter, Smith, Lesko, and Fernandez (1989) also focus on cognitive engagement as psychological investment, describing it as the psychological investment needed to master the knowledge and skills that are taught in school. In other words, these researchers assert that cognitive engagement, in a classroom setting, is a psychological state resulting in students' willingness to exert a high amount of mental effort in order to complete their schoolwork. Some school engagement researchers, when defining cognitive engagement, focus more on the outward behaviors exhibited by cognitively engaged students. For example, Connell and Wellborn (1991) provide a definition of cognitive engagement that includes flexibility when it comes to problem solving, preferring hard work, and demonstrating positive coping strategies after perceived failure. The work of cognitive engagement researchers such as Connell and Wellborn demonstrate how difficult it is at times to distinguish between behavior and cognition. All of these researchers are united in the belief that students who are cognitively engaged exert mental effort when it comes to learning.

The second area of study, which concerns the process of learning, characterizes cognitively engaged students as those who use learning strategies in order to help them fully understand the material. Examples of such strategies used by cognitively engaged students include rehearsing, summarizing, and elaborating on the material to fully understand it (Pintrich & De Groot, 1990; Zimmerman, 1990). Additionally, cognitively engaged students are able to avoid distractions in order to sustain their engagement (Corno, 1993; Pintrich & De Groot, 1990). As a result, engaged students understand the material better than non-engaged students and create more connections between ideas (Weinstein & Mayer, 1986). Greene, Miller, Crowson, Duke, and Akey (2004) found that meaningful strategy use is influenced by mastery goals, but not by performance goals. In other words, students who seek to master the material are more likely to use learning strategies that help them get the most out of the material. Some meaningful strategies students endorsed included doing practice problems, planning study time for class, putting new information into their own words, and making sure that they understand ideas while they are studying. It is important to note, however, that not all strategies are employed to actually enhance learning. While some learning strategies are employed in an attempt to assist the learner to engage with material in a meaningful way (deep cognitive engagement), other learning strategies are employed for the purpose of minimizing student effort (shallow cognitive engagement).

Deep and Shallow Engagement

Research has suggested that there are two components of cognitive engagement: meaningful and shallow (Greene & Miller, 1996). Other researchers (McLaughlin, 2005, McLaughlin et al., 2005; Nystrand & Gamoran, 1991) refer to these components as substantive and procedural, respectively. Smiley and Anderson (2011) used the terms deep and shallow cognitive engagement. Despite variations in the labels used for these two types of cognitive engagement, the definitions for each component are essentially the same.

Deep cognitive engagement refers to the extent that students interact with material in a meaningful way. According to the psychological investment literature, students who seek to master material are demonstrating deep cognitive engagement. Referencing back to the learning strategies definition, a student who is deeply engaged with academic material may use strategies such as metacognitive checks when reading new material and relate new ideas to previous knowledge (Lublin, 2003). For example, students may use previous exams to predict what type of questions will be asked on an upcoming exam; as a result, students become aware of which material needs to be mastered before the exam.

Shallow cognitive engagement refers to the extent that students avoid exerting mental effort, or use quick and easy learning strategies in order to complete a task as quickly as possible. Meece, Blumenfeld, and Hoyle (1988) characterize shallow engagement as the use of learning strategies to get work done with minimal effort. Examples of such strategies include copying down other students' answers and guessing on items they are unsure of so they can finish quickly. In other words, shallowly engaged students tend to do the bare minimum in order to finish what is before them. In an academic setting, a student who is high in shallow engagement could enlist strategies such as reading the questions first, then finding the answers in the material or using rote memorization to prepare for a test.

Deep and Shallow Engagement in an Assessment Context

Up until this point, cognitive engagement has been described generally. This section will describe what deep and shallow cognitive engagement is expected to look like in a low-stakes assessment setting.

A student who is deeply cognitively engaged when taking assessments could be described as one who uses strategies such as metacognitive checks and repetition to ensure that he or she thoroughly understands the questions and material associated with the questions before providing an answer. For example, a student attempting to answer a question in response to a scenario might ask him or herself if he or she fully understands the scenario before even thinking about the answer to the question. Another example is that for a quantitative and scientific reasoning assessment, a deeply engaged student would spend the time required to understand the problem presented and engage in the process to solve it. Finally, for an open-response assessment, a deeply engaged student would plan out their answer carefully and elaborate on their thoughts.

A student who shallowly engages with the assessment can be envisioned as one who may utilize testing strategies such as picking the longest answer when they are unsure what the answer really is when it comes to a selected-response test. When it comes to a scenario-based test, he or she would likely look for the answer within the scenario, in contrast to the critical problem-solving that the deeply engaged student would do. For a quantitative and scientific reasoning test, the shallowly engaged student may just guess if the process to obtain the answer feels too complicated. Finally, for an open-response assessment, he or she would provide the minimally sufficient answer (e.g. if the question required at least a one paragraph response, he or she would provide one paragraph).

Measurement of Cognitive Engagement

Researchers attempting to assess student engagement in the classroom setting have used various methodologies. Methods for evaluating cognitive engagement in academic settings include observing students in specific classes. For example, Helme and Clark (2001) videotaped students during math and science lessons and interviewed 24 students who either appeared to interact heavily with their teachers and peers or who appeared uncertain about their thoughts, motivation and actions during the lessons. In particular, the researchers were looking for the reasons (cognitions) behind why students behaved the way that they did. In order to accomplish this, students watched segments of themselves on video and provided explanations as to why they engaged in certain behaviors. Follow-up interviews were conducted with teachers in order to determine which parts of the lessons the teachers viewed as particularly important to learning or teaching. The result of these interviews was a list of behavioral indicators of cognitive engagement, such as asking and answering questions, resisting distractions, and using gestures. Though this observational technique could be considered useful for evaluating engagement in the classroom, it would not make sense in an assessment setting given that students test independently and are tested on multiple subjects.

Zimmerman and Martinez-Pons (1986) also used interviews, without the corresponding observations, in an attempt to study students' learning strategies. Each student was presented with six scenarios and each scenario was based in a different learning context. The researchers asked students which strategies the students would use to accomplish the given task. For example, one scenario asked what process students go through in writing a paper. Results indicated that high-achieving students tended to use more useful learning strategies than low-achieving students. Though not all strategies identified through the interview process are fitting of an assessment context, some strategies could be relevant. For instance, "reviewing the text" could translate into reviewing all information presented related to a given assessment item or scenario in order to ensure that a key piece of information is not missed. Again, the use of interviews may be useful in a classroom setting to gain an understanding of learning strategies; however, interviews remain impractical for assessing students' degree of cognitive engagement in a low-stakes assessment situation where an institution is assessing a large number of students.

Quantitative methods, while more practical for assessing a large number of students, are sparse, can be indirect, and often pertain to only one subject area (e.g. the Science Activity Questionnaire). Another issue with existing measures is that they were developed for use in the classroom and are not directly transferable to an assessment setting. For example, Pintrich and De Groot (1990) administered a 56-item scale to 173 middle school students from particular English and science classes. The scale asked about the students' motivation, cognitive strategy use, metacognitive strategy use, and management of effort. Students responded to the items on a 7 point Likert scale *(1= Not* *at all true of me* to *7= Very true of me*) and were instructed to answer based on their experiences in the specific English or science class. Results from the study did support a 13-item “Cognitive Strategy” factor that yielded scores with a high degree of internal consistency (α = .83). However, many of these items (e.g. “I outline the chapters in my book to help me study”) focus on students’ engagement in learning new material and are not applicable to an assessment setting where the emphasis lies on evaluating students’ understanding of the material. However, a few items on Pintrich and De Groot’s original 56-item measure may be adaptable to an assessment setting. For example, “I ask myself questions to make sure I know the material I have been studying” could be rephrased as “When taking the assessments, I asked myself questions as I went along to make sure the material made sense to me.” Greene and Miller (1996) developed a “Motivation and Strategy Use Survey” that included items intended to measure deep and shallow engagement (e.g. “When learning the material, I summarized it in my own words”) that can be modified and used for an assessment setting. Greene, Miller, Crowson, Duke, and Akey (2004) developed a measure that included twelve items that addressed study strategies (e.g. “I plan my study time for this class” and “It is easy for me to establish learning goals in this class”). The researchers did find that students with mastery goals tended to endorse those strategies, while those with performance goals did not. Such a finding appears to provide some external validity evidence for the measure. However, once again, many of the items written by Greene et al. (2004), would not work well in an assessment situation as originally written (e.g. “When I finish working practice problems or homework, I check my work for errors”). A few of the items *with modifications* may, however, be adapted for use in an assessment context.

In addition to ensuring that items are representative of cognitive engagement in a low-stakes assessment setting, it is essential to consider how many facets of cognitive engagement the items are to represent. The number and nature of the facets will determine how the resulting measure is scored. For example, research by Greene and Miller (1996) demonstrated that deep cognitive engagement decreased the effects of shallow engagement. In other words, as deep engagement increased, shallow decreased. This finding suggests that cognitive engagement is a unidimensional construct. In other words, deep and shallow engagement cannot be engaged in simultaneously; if a student engages in one, it is at the expense of engaging in the other. However, other research has indicated that cognitive engagement is a multidimensional construct. In other words, it is possible for a student to engage in both at the same time. For example, Meece, Blumenfeld, and Hoyle (1988) administered the Science Activity Questionnaire, which included 8 items intended to assess deep engagement and 7 items intended to assess shallow cognitive engagement, to fifth grade students after science lessons. Results indicated that the two factors were distinct. Because previous findings regarding the dimensionality of cognitive engagement have been inconclusive, both unidimensional and multidimensional models were tested in the current study. The model championed will inform how the measure is scored. If the scale is found to be unidimensional, item scores will be added to create a total score; in contrast, a multidimensional scale will result in subscale scores.

Measuring Cognitive Engagement in an Assessment Context

As outlined above, a few cognitive engagement scales have been developed to assess student engagement in the classroom. Until recently, however, no measures existed to assess cognitive engagement in a low-stakes assessment context. Smiley and Anderson (2011) created the Cognitive Engagement Scale (CES) to assess cognitive engagement specifically in an assessment setting. The CES was adapted from the Motivation and Strategy Use Survey developed by Greene and Miller in 1996. The original CES consisted of 3 items written to assess deep cognitive engagement and 2 items designed to assess shallow cognitive engagement. Additionally, researchers developed the CES scale to be completed after students took an *open-response* assessment instrument designed to assess student knowledge and reasoning related to the fine arts and humanities. A two-factor structure (deep, shallow) was subsequently supported (Smiley & Anderson, 2011). Coefficient alpha was .56 for the three-item deep subscale, which is below the recommended cutoff of .70 for individual-level inferences (Nunnally ,1967). The shallow subscale, which contained only two items, resulted in a coefficient alpha of .71. However, additional items were added based on ideas gathered from existing measures previously described in an attempt to more fully represent the breadth of both deep and shallow engagement. As a result, researchers expanded the CES scale to 9 items (five deep, four shallow) and adapted the items to be completed directly after a *selected response* instrument, rather than in a later non-cognitive measure. The revised CES-2 was administered to college sophomores during a university-wide assessment day, after they completed a multiple-choice assessment that addressed quantitative and scientific reasoning. However, the proposed two-factor structure was not supported (Charsha, Smiley, & Anderson, 2012) for the CES-2. A review of the standardized residuals and item content revealed that items that addressed deep cognitive engagement appeared to function well. However, the two shallow items that functioned well on the original CES, did not function well as part of the CES-2. Interestingly, the two new items added by the researchers to represent shallow engagement actually appeared to form their own factor. It appeared that the two items ("I skipped the hard parts of the assessment" and "When working on the assessment I guessed a lot so I could finish quickly") actually addressed a level of engagement less than that of shallow engagement. The items appeared to address an absence of engagement or a "no engagement" factor. Not surprising given the observed correlations, a three-factor model (deep, shallow, no engagement) fit the data better than the original two-factor model. However, modifying the model post hoc and re-testing on the same sample often capitalizes on chance thus inflating fit indices (MacCallum & Austin, 2000). Furthermore, two items each for the shallow and no engagement factors are likely not sufficient to represent the construct. As a result, the CES-2 scale was lengthened again, modifying items from the existing measures previously described, to twelve items representing 3 factors (5 deep, 4 shallow, 3 no engagement), resulting in the CES-E.

The current study used data collected from a university-wide assessment day in which two samples of incoming freshmen filled out the CES-E; one sample completed the CES-E after completing a multiple choice test that addressed quantitative and scientific reasoning, and the other sample completed the CES-E after completing the multiple choice cognitive portion of an assessment that addressed knowledge about health and wellness. The researcher tested two competing, yet theoretically supported models: a multidimensional three-factor model and a unidimensional model.

Support for a three-factor model would imply that students could display more than one type of engagement (e.g. could endorse items that address deep, shallow and no engagement) in a given context. Support for a one-factor model would suggest that cognitive engagement actually falls on a continuum (Greene & Miller, 1996) between no engagement and deep engagement. In other words, if a student exhibits a high level of deep engagement, his or her level of shallow engagement will be low. Support for the one-factor model would indicate that the CES-E can be scored as a single total score with higher scores indicating more deep cognitive engagement and lower scores indicating an absence of cognitive engagement. Support for the three-factor model would indicate the need for separate subscale scores for each of the factors. However, structural validity evidence alone is not sufficient to support construct validity. Once the structure is established, the final step in Benson's model is external validity, which involves relating scores on the CES-E to scores on other measures to see if they relate in expected ways. The next section describes constructs that logic and theory indicate should be related to cognitive engagement. All constructs are described and ways in which they are expected to be related to cognitive engagement is provided.

Cognitive Engagement, Motivation and Effort

A commonality among definitions of cognitive engagement, both in the psychological investment and the learning strategies areas, is the use of the word "effort." The school engagement literature, which stresses psychological investment, characterizes a student that is high in effort as one who tries hard in school and seeks to master material. The learning strategies literature describes students who exert high levels of effort as those who make use of strategies such as metacognitive checks. Though effort is framed in different ways by the two areas, it still refers to students who seek to learn. The focus on effort as it relates to cognitive engagement is a source of confusion among researchers because it is similar to constructs in the motivation literature. For example, Brophy (1987) discusses that students who are motivated to learn are characterized by such traits as valuing learning and striving (e.g. putting forth effort) to master the material. The mastery goal (Ames, 1992; Dweck & Leggett, 1988) is another example of a construct found in the motivation literature that sounds similar to cognitive engagement. Students who set mastery goals seek to learn the material for the sake of learning, not to outperform others or to receive a grade. Additionally, mastery goals are related to preference for hard work (Ames, 1992), which usually involves effort. However, there is an important difference between motivation and cognitive engagement: motivation is more general (Appleton, Christenson, Kim & Reschly, 2006; Newmann et al., 1992), whereas cognitive engagement is context-specific (Marks, 2000). Essentially, a student can be motivated to do well in school, but their level of cognitive engagement can vary with each class. In an assessment context, this could mean that a student can be motivated to do well on the assessments, but their level of cognitive engagement varies with each assessment. In other words, cognitive engagement is context specific.

An additional problem lies within the fact that behavioral engagement is often defined as including effort as well. For example, part of Skinner and Belmont's (1993) characterization of behaviorally engaged students includes that "[behaviorally engaged students] exert intense effort and concentration in the implementation of learning tasks" (p. 572). An attempt to differentiate between cognitive effort and behavioral effort was made by Corno (1993), who states that cognitive effort places an emphasis on the student actively avoiding distractions (e.g. volition). As a result, "effort" is not directly observable. In contrast, much of the research examining behavioral engagement focuses on observable actions. For instance, several measures of behavioral engagement ask teachers to rate their students on components such as approaching assignments with real effort and being persistent when confronted with problems (Finn et al., 1995). In the behavioral context, effort appears to be characterized more as completing the work, whereas cognitive effort is described as the intent to fully understand the material with the ability to block all distractions along the way. Clearly, more work needs to be done to further examine the distinctions between effort and engagement. If effort and engagement are two distinct constructs, one would expect cognitive engagement and effort scores to be moderately related but not so highly related as to suggest they are actually the same construct.

Academic Entitlement

Academic Entitlement is “the expectation that one should receive certain positive academic outcomes (e.g. high grades) in academic settings, often independent of performance” (Kopp, Zinn, Finney, & Jurich, 2011, p. 106). In other words, students who do not put in the necessary effort into their classes but still expect decent grades are described as high in academic entitlement. Kopp et al. (2011) further explained that students who display this characteristic view themselves as customers of the university in that they pay tuition, so they believe they should receive decent grades based on that, not because of the work they do. This view can manifest itself into situations such as students justifying that they deserve a better grade in a class with such reasons as “But I came to class nearly every day” (pg. 105) or disputing a grade on an assignment by emphasizing the number of hours they put into it when the professor assigned the grade based on accuracy. Although no previous work has been done to demonstrate the relation between academic entitlement and cognitive engagement, the researcher expects there to be a strong negative relationship between deep cognitive engagement and academic entitlement, and at least a moderate positive relationship between shallow and no engagement and academic entitlement. In other words, students who employ deep testing strategies are less likely to display academic entitlement due to their already formed habits in the classroom of striving to understand the material presented and providing thoughtful answers. In contrast, students who employ shallow strategies or do not engage at all with the assessments are more likely to be higher in academic entitlement due to their habits of minimizing effort in order to complete work quickly with the expectation of still receiving a good score. While the AEQ does not address entitlement in an assessment context specifically, it does probe students regarding entitlement within the broader learning experience, including in regards to testing and the evaluation of student performance.

Expectancy-Value

The Expectancy-Value theory (EV) of motivation states that "individuals' choice, persistence, and performance can be explained by their beliefs about how well they will do on the activity and the extent to which they value the activity" (Wigfield & Eccles, 2000, p. 68). In a college setting, this means that students' class performances depend on how well they think they will do and how much they value the classes that they take. Although EV provides a good framework for evaluating student performance, Hulleman, Barron, and Lazowski (2011) also evaluated the component of cost. Cost is described as what a student has to give up in order to do well in a class (e.g. they may not like the idea of having to give up time with friends to study for a difficult class). When students have high expectations, high value, and perceive little cost, optimal motivation occurs. In terms of relation to cognitive engagement, those who are deeply engaged are theorized by the current researcher to possess the high expectations, high value, and perceived low cost described above. In contrast, students who are shallowly engaged or not engaged at all with the assessments are more likely to have low expectations of themselves and to perceive the assessments as having little value. In addition, they are likely to view putting in the effort to fully process the material as costly. While the EVC does not address expectancy, value and cost in an assessment context specifically, it does ask students regarding their overall expectancy, value and cost associated with their classes, not just within the classroom itself.

Conclusion

Despite a large amount of research designed to aid in the task of understanding students' test-taking behavior and attitudes, the information remains incomplete. One possible reason is the absence of an examination of cognitive engagement from such studies. While there have been efforts made to assess cognitive engagement in classroom settings, efforts related to assessing cognitive engagement in assessment contexts have been minimal. Making such examinations difficult is the lack of a measure for cognitive engagement in the assessment context. Using Benson's framework (1998), the researcher aimed to establish structural and external validity evidence for such a measure (CES-E). This study attempted to evaluate, using confirmatory factor analysis, the model-fit of the CES-E (i.e., Benson's structural stage) by determining whether the theoretically supported one-factor or three-factor model best fit the CES-E data. Given an interpretable and empirically supported factor structure, the researcher also planned to examine whether CES-E scores relate to other constructs in anticipated ways (i.e. Benson’s external validity stage).

Chapter 3

Methods

Whereas the previous chapter examined what Benson (1998) termed the substantive stage of the development of the CES-E, the purpose of this chapter is to describe the participants, procedures and analyses proposed to begin to address both the structural and external validity stages of Benson's model. The chapter outlines the methods for the three separate phases of the current study. In Phase I, the researcher tested two competing models. Each model is theoretically supported; however, based on previous examinations (Smiley & Anderson, 2011; Charsha, Smiley & Anderson, 2012), the three-factor model was expected to fit the data better than the one-factor model. In Phase II, both models were cross validated on a second independent sample to ensure the models were not simply capitalizing on idiosyncrasies of the data from sample one. In Phase III, CES-E scores were to be correlated with scores on other measures to examine if CES-E scores relate to these other variables in anticipated ways. Like the two models from Phase I, the anticipated relationships with external variables were theoretically driven. All phases of the study used archival data collected as part of the university's ongoing assessment initiatives. Specifically, the study used data collected in the fall of 2012.

Phase I: Factor Structure

Participants and procedure

Participants were 602 entering first year students who participated in university-wide assessment day activities. The sample was composed of 65% females and 87% were Caucasian. The mean age of participants was 18.42 with a standard deviation of .35.

Data collected during assessment day activities is used to assess student learning and development. Students first complete the assessments when they enter the university as in-coming first year students and complete the same assessments again when they have accumulated 45-70 credits. Assessments are completed either during a morning or afternoon session and students are randomly assigned to rooms based on the last two digits of their student ID numbers. The assessment instruments students complete are based on students' room assignments. The assignment of students to assessment instruments by ID number enables the University to track students and administer the same assessment instruments to students during their sophomore year as they completed as incoming students. Students complete the assessments under proctored conditions using standardized testing procedures. Additionally, because assessments are high stakes for the university, but low-stakes for students, proctors are trained to motivate students (Lau et al., 2009).

Students upon entering their assigned room received welcome sheets and signed consent forms. Before each assessment, proctors read aloud test instructions. Additionally, students were instructed to complete one measure at a time and are not allowed to start on the next assessment until all students in the room are finished with the previous test or time is called.

Instruments

Students first completed the knowledge-based component of the Knowledge of Wellness and Health (KWH), Version 7, a 31-item selected-response subscale of the 75-item KWH instrument. The KWH is used by the university to assess student learning associated with a specific sub-set of the university's general education objectives that address health and wellness. The KWH also contains a series of questions about students' health and wellness habits and attitudes. Coefficient alpha for the 31-item knowledge based subscale was .30. Though this indicates poor reliability, this is not particularly important for the current study because the researcher was not attempting to use KWH-7 scores in any way. Instead, students were simply asked to consider this assessment while completing the instrument of interest, the Cognitive Engagement Scale-Extended (CES-E).

The CES-E consists of 12 items: five developed to assess deep cognitive engagement, four developed to assess shallow cognitive engagement and three developed to assess an absence of cognitive engagement. The complete CES-E scale can be found in Appendix A. An example of a deep CES item is "When preparing to answer the questions on the assessment, I stopped to reflect on the information provided." An example of a shallow CES item is "If I was not sure about the answer to a question on the assessment, I picked the longest answer." Finally, an example of a 'no" or absence of engagement item is "I skipped the hard parts on the assessment." All items are answered on a 7-point Likert scale (1 =*Strongly Disagree*, 7 = *Strongly Agree*). Because the CES-E is designed to assess student engagement on knowledge or reasoning assessment tasks, students participating in this study first completed the 31 knowledge-based selected-response items, then completed the CES-E. After completing both the knowledge-based KWH items and the CES-E items, students completed the attitudes and habits portion of the KWH. In order to smoothly transition between the Cognitive Engagement Scale-Extended and the KWH, the CES-E items were presented to students as items 32 through 43 of the KWH.

Data screening for CES-E data

Any cases that were missing data were removed from the dataset. An examination of item distributions revealed that there were no univariate outliers Item distributions were also examined to determine if every option was selected at least once for every item. This was done in order to establish whether the data could be treated as continuous. Data were then screened for multivariate outliers through an examination of Mahalanobis distances. Because the data modeled through the CFAs are expected to be linear, linearity between every possible pair of items were assessed through bivariate scatterplots. Univariate normality was be evaluated by using absolute values of 2 for skew and 10 for kurtosis; exceeding these cutoffs indicates a violation of the univariate normality assumption (West, Finch & Curran, 1995). Finally, multivariate normality was assessed by Mardia's normalized multivariate kurtosis. Values bigger than 3 could bias results (Bentler & Wu, 2003); in such a case, an adjustment was be used along with the chosen estimation method, described later in this section.

Data Analysis

In order to assess the factor structure of the scale, two competing models were tested for each sample. The variance for the latent variables was scaled to 1. The one-factor model (Figure 1) suggests that all twelve items represent a single underlying factor of cognitive engagement and supports the theory that cognitive engagement is represented by a continuum from no engagement to shallow engagement to deep engagement. The three-factor model (Figure 2) is a multidimensional model where the three factors (no engagement, shallow engagement and deep engagement) are correlated. It should be noted that although the error variances are not represented for each model, they were estimated. Evidence that the one-factor model fits the data would suggest that a single total cognitive engagement score should be calculated while championing the three-factor model would support the calculation of three separate sub-scale scores.

In addition to looking at measures of global fit, localized misfit was also examined through standardized correlation residuals. Both the three-factor and one-factor models were tested in both samples and any consistent issues noted. An example of this could be that a shallow item that appears to have a stronger relationship with the items meant to address deep cognitive engagement rather than other shallow engagement items. If movement of the item to that factor could be justified based on substantive considerations, then the item would have been moved to the deep subscale in a modified model. However, if modification of the model or scale could not be theoretically justified, then the modification was not made; in other words, modifications were not based solely on fit.

Figure 1. *One-Factor Model*

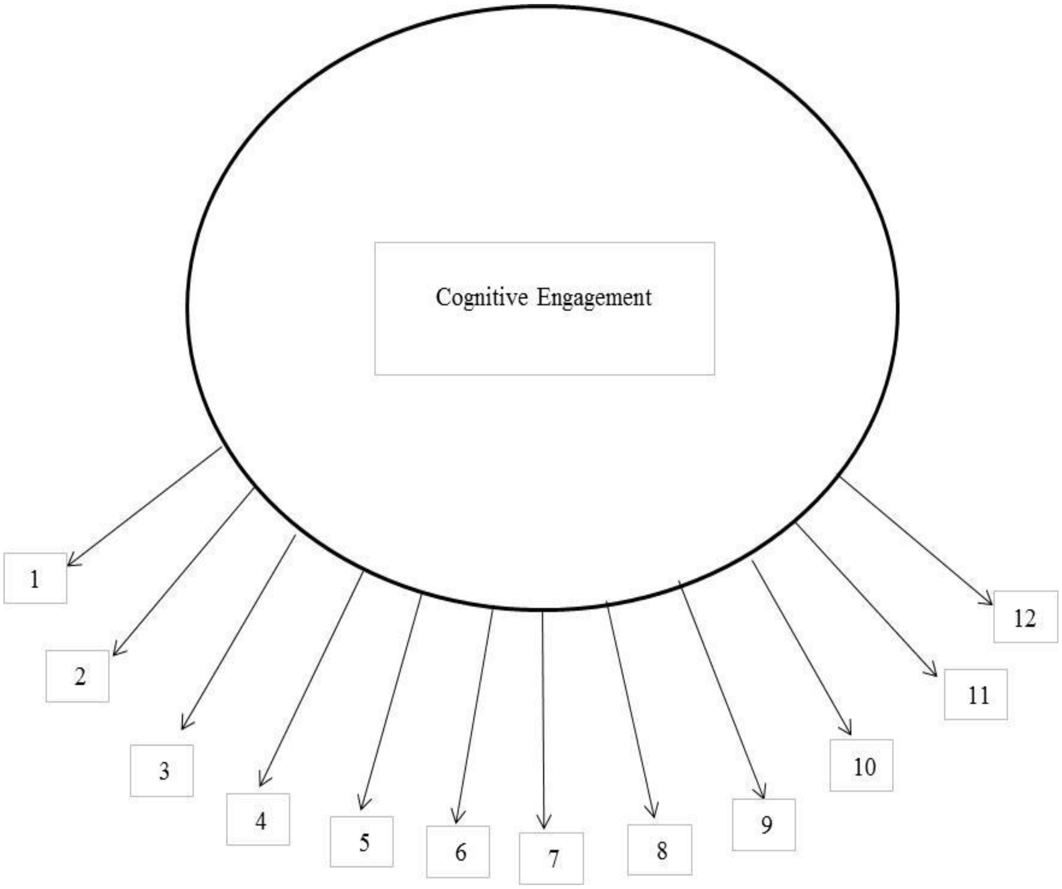
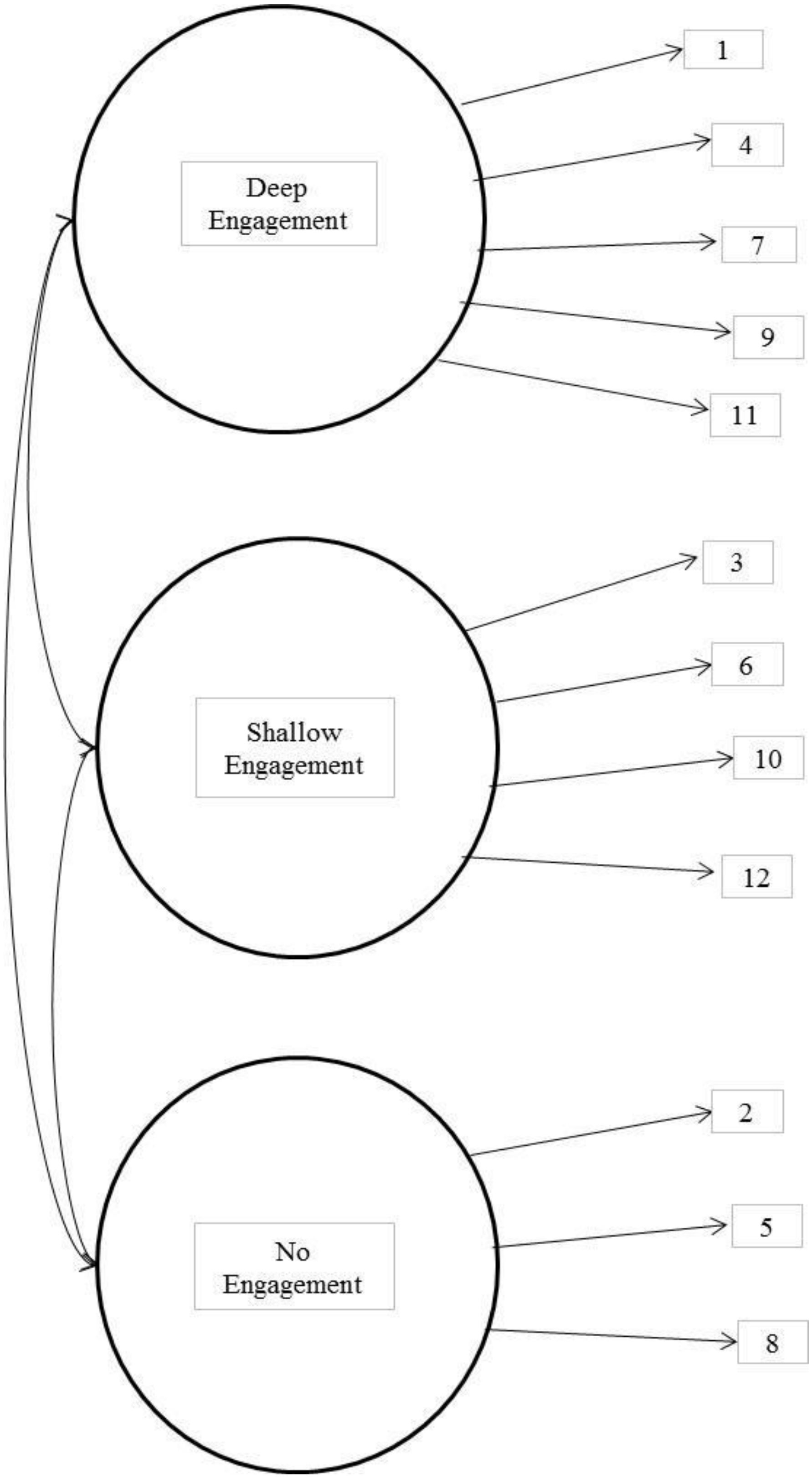


Figure 2. *Three-Factor Model*



Estimation method

Models were estimated in LISREL 8.80. Maximum Likelihood estimation was used due to its robustness (i.e. ability to yield unbiased parameter estimates) to smaller sample sizes and misspecification of models (Olsson, 2000). Additionally, because the data proved to be non-normal, the estimation method was used with an adjusted Satorra-Bentler chi-square and Robust Standard Errors to compensate for the non-normality (West, Finch & Curran, 1995).

Assessing model-data fit

Hu and Bentler (1999) recommend reporting at least one absolute fit index (i.e. stand-alone indexes that do not compare the model to another model) and one relative fit index (i.e. examines if one model fit the data better than another model). The fit indexes used for this study included the absolute fit index of χ^2, which is actually a "badness of fit test" indicating that the model fits the data well if the test is non-significant. Additionally, root mean square error of approximation (RMSEA), which takes model complexity into account, was examined. The standardized root mean square residual (SRMR), which is sensitive to misspecification of factor correlations and misspecification of factor pattern coefficients, was used to evaluate model fit as well. Additionally, the comparative fit index (CFI), a relative fit index, was examined. The CFI is also sensitive to misspecification of factor correlations and misspecification of factor pattern coefficients; it ranges from 0 to 1 and assesses the degree to which the model fits the data better than a model in which all of the indicators are uncorrelated. Though Hu and Bentler provided their own suggested cutoffs for all of the fit indexes, the current study used the suggested adjusted cutoffs provided by Yu & Muthén (2002) because of the fact that the Satorra-Bentler adjustment was used. Suggested adjusted cutoff values for each of the fit indices are as follows: χ^2: *p*>.05, RMSEA: <.05, SRMR: <.07, and CFI: >.96. Comparisons of these fit indices between the two models determine which model fits the data best. For example, if the one-factor model met the cutoff restrictions and the three-factor model did not, it would be clear that the one-factor model fit the data better. Should both models fit, a change of χ^2 ($\Delta\chi^2$) test would have been computed between the two models. If significant, it implies that the more parsimonious model fits worse than the more complex model. The model with the best fit would have been championed. Because a $\Delta\chi^2$ test can only be performed if models are nested, it is worth noting that the three-factor and one-factor model are nested; the correlations between the factors in the three-factor model can be set to 1 to change the model from a three-factor to a one-factor model. If neither model fits, this test is not performed and no model is championed. Instead, correlation residuals are examined in order to determine why the models do not fit.

Phase II: Cross Validation

Participants and Procedure

Participants were 372 entering first year students who participated in the same university-wide assessment day described in Phase I. This sample was again mostly female (59.4%) and Caucasian (89%). The mean age was 18.46 with a standard deviation of .36.

In Phase I, participants responded to items on the CES-E after completing the knowledge-based portion of the KWH. Here, participants first completed a different test, the Natural World, version 9 (NW-9). Additionally, students completed the Student

Opinion Survey (SOS) immediately following the NW-9, followed by the CES-E. The SOS is discussed more in Phase III below. In order to ensure that students were thinking about the NW-9 while answering items on the CES-E, the SOS and CES were printed onto one page and stapled to the back of the NW-9 with instructions to answer all 22 items (10 SOS items and 12 CES-E items) while thinking about the assessment (NW-9) they just completed.

Instruments

Students first completed the Natural World, Version 9 test (NW-9), which consists of 66 selected-response items. The university uses the NW-9 to assess student learning associated with a specific sub-set of the university's general education objectives that address quantitative and scientific reasoning. Reliability coefficients have been demonstrated to be above .7 in paper and pencil administrations (Sundre, 2008).

Students completed the same version of the CES-E as described in Phase I above.

Data screening for CES-E data

The same data screening procedures used in Phase I were also used for Phase II. Once again, the archival data was examined to ensure that every option was selected at least once for every item. Additional procedures included examining linearity, checking for outliers, viewing bivariate scatterplots to determine if data is linear, and assessing normality.

Data Analysis

After the data analysis was completed for Phase I, both models (one-factor and three-factor) tested in Phase I were tested in Phase II in order to determine if findings regarding the factor structure replicated when students completed the CES-E after completing a different, more difficult instrument. This is important given that the ultimate goal was to establish a cognitive engagement scale that would reflect students' levels of engagement across different tests. Omega was only calculated for each factor in the model if the same model was championed in the two phases of the study. Additionally, variance extracted (i.e. the amount of overall variance in the indicators accounted for by the latent variable) was also calculated. All estimated parameters for a championed model were interpreted.

Phase III: External Validity

Participants and Procedure

Participants and procedure will be the same as described in Phase I. The Phase I sample was selected because it is larger than the independent sample used in Phase II. Following the completion of the KWH and cognitive engagement items as described in Phase I, students responded to several non-cognitive measures. These measures included the Academic Entitlement Scale, the Expectancy-Value Cost scale, and the Student Opinion Survey (SOS).

Instruments

The Academic Entitlement Questionnaire (Appendix B) is an eight item scale that measures how students feel regarding the grades they receive. Sample questions include "Because I pay tuition, I deserve passing grades" and "It's the professor's responsibility to make it easy for me to succeed." Previous research by Kopp et al. (2011) found that a unidimensional model fit the data. In other words, a single latent construct does appear to drive the responses to the items. In addition, reliability was demonstrated by omega coefficients of .81 and .84 in the two samples.

The Expectancy-Value Cost Scale (EVC, Appendix C) consists of 16 items that measure student attitudes toward their classes. More specifically, the scale gauges if the students believe that classes will be worth their time and effort. Sample items include “I value the classes I am taking this semester”, “I think my classes will be interesting”, and “I see a purpose for taking my classes this semester.” Students choose their responses on a 1 (strongly disagree) to 8 (strongly agrees) scale. The EVC scale was recently developed by Flake et al. (2011) from existing questionnaires. In their previous forms, these questionnaires yielded reliable scores and support for valid inferences. In the current form, Barron, Hulleman, Lazowski, Flake and Grays (2012) discovered that Expectancy-Value was one of the constructs that accounted for the most variance in course grades for the classes students find the least and most motivating. The researchers used both qualitative and quantitative methods. A three-factor model was supported, meaning that answers to the items are being driven by three latent factors (expectancy, value, cost). As a result, three subscale scores are calculated.

The Student Opinion Survey (SOS, Appendix D) is a ten-item instrument used to assess student motivation during assessment activities. The SOS is a two-factor scale, which consist of effort and importance. Each subscale contains five items. The university uses the SOS to monitor students’ motivation while taking student learning outcome assessments. SOS scores help to inform assessment professionals regarding the validity of the inferences made from scores produced from the assessments. Reliability coefficients across samples consistently fall between .8 and .9 for both subscales. Additionally, the correlation between the two is moderate across multiple samples (r =

.41 for 2008), indicating that the two subscales are related but not redundant (Sundre, 2008).

Data analysis

Data was to be analyzed through a series of simple Pearson correlations between observed scores that illustrated the relationship of the identified measures with CES-E scores. In a situation where the three-factor model would be championed, effort scores were expected to more highly positively correlate with deep engagement subscale scores and negatively with both shallow and no engagement subscale scores. Also, the researcher predicted that the expectancy and value scores from the EVC to be positively correlated with deep cognitive engagement scores and negatively related to shallow and no engagement scores. Cost scores from the EVC and academic entitlement scores were expected to be negatively related to deep cognitive engagement, and positively related to shallow and no engagement scores. In a situation where a unidimensional model is supported, higher cognitive engagement scores would reflect deeper engagement. In this case, the researcher predicted that cognitive engagement scores would be moderately positively correlated with effort scores from the SOS, positively correlated with expectancy and value scores, and negatively correlated with cost scores and Academic Entitlement scores.

Conclusion

In summary, the purpose of the presented study was to examine the validity of inferences made from scores on the CES-E, a cognitive engagement scale developed to assess students' levels of cognitive engagement within a low-stakes assessment situation. Using Benson's framework (1998), two theoretical factor models were tested and cross-validated on an independent sample. Finally, given an interpretable factor structure, CES-E scores would be correlated with other theoretically related constructs in an attempt to examine external validity.

Chapter Four

Results

Results for Phase I and Phase II are presented below. Both phases examined whether a theoretically supported model (one-factor or three-factor) fits the data. Phase I used the data obtained from students who completed the KWH cognitive items before filling out the CES-E in an attempt to determine which model best fit the CES-E data. Phase II used the data obtained from students who completed the NW-9 before filling out the CES-E in an attempt to cross-validate findings from Phase I.

Phase I: Of the two proposed models, which model fits the data best?

Data Screening and Descriptive Statistics

Before running CFA analyses, the data were screened for missing responses. Seven cases with missing data were removed from the dataset, bringing the initial sample size of 609 to 602. After the data was screened for missing data, item frequencies were examined in order to establish that all seven answer options were used for all 12 items in order for the data to be treated as continuous. Item frequencies indicated that all seven response options were used for all 12 items.

Data were then screened for both univariate and multivariate outliers. No univariate outliers were identified. Additionally, Mahalanobis distances indicated that there were no multivariate outliers for the sample. Furthermore, linearity between every possible pair of items was assessed through bivariate scatterplots and no curvilinear relationships were discovered.

Finally, univariate and multivariate normality were examined. Univariate normality for each item was evaluated by determining whether skewness and kurtosis values exceeded the absolute value of 2 for skewness and absolute value of 10 for kurtosis; exceeding these cutoffs indicates a violation of the univariate normality assumption (Finney & DiStefano, 2006West, Finch & Curran, 1995). Two of the items, “If I was not sure about the answer to a question on the assessment, I picked the longest answer” (skew value of 2.1) and “I skipped the hard parts on the assessment” (skew value of 2.98) exceeded the cutoff for skewness. Additionally, the latter item exceeded the kurtosis cutoff with a value of 10.30. The data also violated the multivariate normality assumption as evidenced by a Mardia’s normalized multivariate kurtosis coefficient of 26.15. Given that this value exceeded 3, it was concluded that the data were multivariately non-normal (Bentler & Wu, 2003).

In order to simplify the discussion of items, item names are abbreviated according to the facet of cognitive engagement they were written to represent. Items meant to represent deep are noted with a *D*, shallow items with an *S,* and no engagement items with an *N*. Additionally, a number is attached to each item which signals its placement on the test. For example, the item “When preparing to answer the questions on the assessment, I stopped to reflect on the information provided” was written to measure deep engagement and was the first deep engagement item to appear on the CES-E; thus, its label is D1.

Item means had a wide range of values. Specifically, means ranged from 1.51 (N1) to 5.28 (S3). Item N1 had the least amount of variance (*SD*=1.05) and item S2 had the most variance (*SD* = 1.96). In the sample, items contained a mix of positive and negative correlations, with absolute values ranging from .01 to .35. Because of the low correlations between items, it is likely that any model would fit the data, at least according to some of the absolute fit indexes. There are three possible reasons as to why the correlations are so small: 1) some items have low variance compared to one another, which restricts the range between item scores and thus they cannot correlate highly 2) the items are not measuring the same construct, and 3) items appear to address the same construct, but item distributions vary across items, which also restricts range and leads to low correlations. These reasons will be explored in Chapter 5. Correlations and descriptive statistics for all 12 items are reported in Table 1.

Table 1. *Correlations and Descriptive Statistics for Phase I sample*

	D1	D2	D3	D4	D5	S1	S2	S3	S4	N1	N2	N3	M	SD	Skew	Kurtosis
D1	1												5.60	1.41	-1.67	2.83
D2	0.24	1											5.08	1.49	-0.72	0.08
D3	0.28	0.21	1										4.95	1.73	-0.72	-0.42
D4	0.17	0.20	0.24	1									4.77	1.55	-0.60	-0.27
D5	0.15	0.24	0.28	0.35	1								4.01	1.62	-0.10	-0.65
S1	0.06	-0.02	-0.06	-0.07	-0.09	1							4.70	1.75	-0.59	-0.66
S2	0.03	0.11	0.09	0.12	0.13	0.06	1						3.92	1.96	-0.08	-1.23
S3	0.21	0.23	0.28	0.31	0.33	0.11	0.17	1					5.28	1.43	-0.93	0.54
S4	-0.15	-0.08	-0.04	-0.05	0.03	0.01	0.04	-0.04	1				1.71	1.15	2.10	4.70
N1	-0.31	-0.14	-0.01	-0.05	0.03	-0.06	0.07	-0.03	0.24	1			1.51	1.05	2.98	10.30
N2	-0.15	-0.16	-0.29	-0.16	-0.23	0.16	-0.03	-0.15	0.08	-0.03	1		4.15	1.92	-0.10	-1.24
N3	0.11	-0.19	-0.2	-0.15	-0.16	0.29	-0.01	-0.07	0.14	-0.01	0.30	1	3.73	1.71	0.05	-0.99

N=602

Estimation Method

Because every item had all seven answer options used at least once, it is appropriate to treat the item responses as continuous in the factor analyses. Therefore, Maximum Likelihood estimation was used to estimate model-data fit. However, because the data violated the multivariate normality assumption, a Satorra-Bentler (S-B) adjustment to the χ^2 statistic, fit indexes, and standard errors was used along with the Maximum Likelihood estimation to correct for the non-normality (Satorra & Bentler, 1994). Model-data fit was estimated in LISREL 8.80.

One-Factor Model

As mentioned previously, because of the low correlations between items, it is likely that any model would likely fit the data, at least according to some of the absolute fit indexes. However, the incremental fit index would likely indicate that the model is poor fit. As predicted, the model did fit the data according to two (RMSEA= .07 and SRMR =.08) of the absolute fit indexes. However, if the model provided excellent fit, the chi-square value would be close to the value of the degrees of freedom; this is not the case with this model ($\chi^2_{S\text{-}B}$ = 231.60, df=54), contradicting what the other two absolute fit indexes indicate about model-data fit. Also as predicted, the CFI value of .75 indicated that the model does not provide a better fit to the data than a model in which the indicators are uncorrelated. This finding is not surprising given the low observed correlations.

In addition to global fit, localized misfit was also examined in the form of correlation residuals. Values greater than |.10| indicate large misfit. Few of the residuals are large, which is not surprising given the small observed correlations between items. However, there were a few places of misfit. Negative residuals, which indicate that the model overestimated the relationship between items, were found between items D1 and N1, and S3 and N2. Of particular note is the negative residual between D1 and N1, which would definitely seem to suggest that the items need to be on separate factors. Positive residuals indicate that the model underestimated the relationship between items and a total of five large positive residuals were found. These large positive residuals were found between items S1 and S3, S1 and N2, S1 and N3, S4 and N3, and N2 and N3. Of particular note is the residual between items S1and N3, which suggests that the model extremely underestimated the relationship between the two items and that they may belong on the same factor. However, given such small observed correlations it is difficult to make definitive conclusions. A full table of the residuals can be found in Table 2.

Table 2. *Residuals for Phase I One-Factor Model*

	D1	D2	D3	D4	D5	S1	S2	S3	S4	N1	N2	N3
D1	0.00											
D2	0.06	0.00										
D3	0.06	-0.03	0.00									
D4	-0.04	-0.02	-0.04	0.00								
D5	-0.07	0.00	-0.02	0.07	0.00							
S1	0.10	0.03	0.00	-0.02	-0.03	0.00						
S2	-0.04	0.03	-0.01	0.03	0.03	0.08	0.00					
S3	0.01	0.00	0.01	0.05	0.05	**0.16**	0.08	0.00				
S4	-0.10	-0.03	0.03	0.01	0.10	0.00	0.06	0.02	0.00			
N1	**-0.27**	-0.09	0.05	0.01	0.09	-0.07	0.09	0.02	**0.22**	0.00		
N2	0.02	0.03	-0.06	0.05	0.00	**0.12**	**-0.11**	0.06	0.03	-0.07	0.00	
N3	0.03	-0.04	-0.01	0.03	0.03	**0.25**	0.05	**0.11**	0.09	-0.05	**0.15**	0.00

N=602

Three-Factor Model

If the proposed three-factor model was appropriate for the data, correlations between items associated with the same factor would positively correlate and be at least moderate in nature. Additionally, correlations between items of different factors would be small. However, these correlation patterns were not found (see Table 1). Instead, the correlations between the items are consistently low regardless of whether the correlations are between items written to be on the same factor or items written to be on different factors. It is possible for a model to fit in such situations due to low correlations among the items; however, the model fit would not have any real meaning. In other words, the fact that the model would reproduce small correlations would not be meaningful to the theory of cognitive engagement. However, the three-factor model did not converge to an admissible solution due to a non-positive definite phi matrix, which is the matrix that describes the correlations between factors. Essentially, the items written to address each of the three factors shared little common variance.

Examination of the estimated correlations between factors from the three-factor model that did not achieve convergence indicated an out-of-range correlation of 15.44 between the proposed shallow and deep factors. This finding makes sense considering that the shallow items were poorly correlated with each other and one shallow item (S3) correlated more strongly with deep items than it did with other shallow items. In other words, the shallow items did not intercorrelate, so a factor based on their common variance is not meaningful. Instead, the factor correlation reflects the association between specific deep and shallow items. In general, the out of-range result may indicate an over-factoring (Rindskopf, 1984) of the data (i.e. the results suggest that shallow and deep should be one factor).

Another out-of-range factor correlation (-8.84) was found between no engagement and shallow engagement. Upon closer examination of the correlation matrix, it was discovered that this factor correlation was driven by the fact no engagement items were correlated negatively with shallow items in greater magnitude than the relationships between no engagement items. The only relationship of value between no engagement items occurred between items N2 and N3. Correlations between N1 and N2, and N1 and N3 were essentially zero. In other words, the no engagement items had the same problem as the shallow items: they did not intercorrelate, so a factor based on their common variance is not meaningful. Because of the non-convergent model, fit information is not provided.

Phase II: Cross-Validation

Data Screening and Descriptive Statistics

The researcher employed the same data screening procedures in Phase II as used in Phase I. Eighteen cases with missing data were removed from the dataset, which reduced the sample size from 391 to 373. Item frequencies revealed that all answer options were chosen. No univariate outliers were identified. Mahalanobis distances indicated 1 multivariate outlier which was removed from the dataset. Thus, the final sample size was 372. Scatterplots indicated linear relationships between all pairs of items. Univariate normality was met but a Mardia's value of 8.61 was found, indicating a violation of the multivariate normality assumption.

Like in Phase I, item means had a wide range of values. Specifically, means ranged from 1.78 (N1) to 5.54 (D1). Item N1 had the least amount of variance (*SD*=1.13) and item D3 had the most variance (*SD* = 1.76). In the sample, items contained a mix of positive and negative correlations, with absolute values ranging from .01 to .40. Because the correlations are similar in magnitude, a one-factor model would likely have good fit. However, because the correlations are so small, this finding would have little meaning. Similar to the Phase I sample, this could be due to low item variance, items not measuring the same construct, or varying item distributions. Explorations to the causes of such low correlations will be discussed in Chapter 5. Correlations and descriptive statistics for all 12 items are reported in Table 3.

Table 3. *Correlations and Descriptive Statistics for Phase II sample*

	D1	D2	D3	D4	D5	S1	S2	S3	S4	N1	N2	N3	M	SD	Skew	Kurtosis
D1	1												5.54	1.21	-1.10	1.02
D2	0.32	1											4.80	1.45	-0.47	-0.32
D3	0.39	0.14	1										4.65	1.76	-0.45	-0.94
D4	0.28	0.31	0.21	1									4.60	1.32	-0.41	-0.22
D5	0.29	0.32	0.25	0.40	1								3.97	1.41	-0.18	-0.47
S1	-0.03	0.01	-0.02	-0.03	0.03	1							4.45	1.65	-0.57	-0.60
S2	0.02	0.02	0.02	0.10	0.21	0.10	1						4.02	1.71	-0.26	-0.98
S3	0.28	0.24	0.14	0.37	0.39	0.05	0.09	1					4.79	1.47	-0.60	-0.17
S4	-0.23	-0.01	-0.14	-0.02	0.07	0.20	0.07	0.02	1				1.93	1.32	1.61	2.03
N1	-0.25	-0.17	-0.09	-0.07	-0.06	0.22	0.03	-0.07	0.15	1			1.78	1.13	1.92	3.70
N2	-0.20	-0.12	-0.24	-0.17	-0.27	0.10	0.08	-0.08	0.09	0.09	1		4.73	1.62	-0.43	-0.75
N3	-0.30	-0.18	-0.12	-0.03	-0.12	0.37	0.02	-0.01	0.27	0.32	0.19	1	3.23	1.59	0.47	-0.73

N=372

Estimation Method

Maximum Likelihood estimation with a Satorra-Bentler adjustment was used once again for the Phase II sample for the same reason as Phase I (i.e. a violation of multivariate normality). The models were estimated again in LISREL 8.80.

One-Factor Model

Because correlations were small for the Phase II sample, the unidimensional model was again expected to show decent fit according to absolute fit indexes, but not according to the incremental fit index. This prediction was supported. One of the values (SRMR = .09) indicated decent absolute fit, but another (RMSEA = .11) did not. Additionally, the chi-square value ($\chi^2_{\text{S-B}}$ = 241.12, df = 54) was again not close to the degrees of freedom. Finally, the low CFI value of .69 implies that the fit of the one-factor model over a model in which all of the indicators are uncorrelated is not an improvement.

Localized misfit was once again examined for the model and although a number of residuals that exceeded |.10| were found, few of them could be classified as very large (i.e. bigger than |.15|). This is once again due to the observed low correlations among items. One negative residual was found between items D1 and S4, indicating that the model overestimated the relationship between those two items. However, the residuals that exceed the |.10| guideline are mostly positive, indicating that the model underestimated the relationships between the items. These larger positive residuals tend to take place between items meant to measure shallow cognitive engagement and no cognitive engagement, notably when it comes to items S1, N3, and S4. This could suggest that these items could form their own factor; however, because the correlations are so low, no definitive conclusions can be drawn. A table of all correlation residuals can be found in Table 4.

Table 4. *Residuals for Phase II One-Factor Model*

	D1	D2	D3	D4	D5	S1	S2	S3	S4	N1	N2	N3
D1	0.00											
D2	0.01	0.00										
D3	**0.11**	-0.08	0.00									
D4	-0.06	0.04	-0.03	0.00								
D5	-0.08	0.03	-0.01	0.08	0.00							
S1	0.03	0.05	0.02	0.02	0.08	0.00						
S2	-0.05	-0.03	-0.03	0.04	**0.14**	**0.11**	0.00					
S3	-0.02	0.00	-0.07	**0.11**	0.10	0.09	0.04	0.00				
S4	**-0.14**	0.06	-0.07	0.06	**0.16**	**0.19**	0.09	0.09	0.00			
N1	-0.08	-0.04	0.03	0.08	0.10	**0.20**	0.06	0.06	**0.11**	0.00		
N2	0.02	0.05	-0.08	0.02	-0.06	0.07	**0.12**	0.09	0.04	0.00	0.00	
N3	-0.10	-0.02	0.02	**0.14**	0.07	**0.34**	0.05	**0.14**	**0.22**	**0.24**	**0.24**	0.00

N=372

Three-Factor Model

Recall that if the proposed three-factor model was to fit the data, correlations between items of each factor would positively correlate and be at least moderate in nature. Additionally, correlations between items of different factors would be small. However, this is not what was illustrated in Table 3. Not surprisingly, the model once again did not converge to an admissible solution due to a non-positive definite phi matrix.

Reasons behind the non-convergence were similar to those found in the Phase I sample. In particular, shallow items have low correlations with one another, indicating that they have little common variance. Additionally, item S3 ("When answering the questions on the assessment, I looked for clues of how to respond within the test itself") was more highly correlated with the deep items than with other shallow items as was also the case in Phase I. Inspection of the phi matrix indicated correlations that were in bounds (unlike the Phase I sample) Because of the non-admissible solution, no fit information can be reported.

Conclusion

Inter-item correlations were low in both samples. Low inter-item correlations have three main causes: low item variance, items not representing the construct, or possibly item distributions restricting the range for items' correlations. Inspection of item variances revealed that none of the variances were particularly low compared to each other and all items had a variance of at least one. Recall that students responded to CES-E items on seven point scale; therefore, variances of one or more signal that items have good variability. So, low item variance does not appear to be a plausible cause for low correlations for the current study. Therefore, the other two causes will be discussed in

Chapter 5. Because the inter-item correlations were so small, the inferences that can be made about model-data fit are limited. The findings related to the one-factor model produced fairly similar results across the Phase I and Phase II samples, with some absolute fit indexes supporting fit but the incremental index and χ^2 indicating poor fit. These findings are not surprising given the small correlations between the items. As for the three-factor model, there were some similar results across the two samples that may explain why the model did not converge to an admissible solution with either sample. For example, shallow items had nearly non-existent correlations with one another in both samples. Therefore, it is difficult to advocate for any type of change to the scale.

Because of the many issues with the scale, no scorable solution for the CES-E was found. Therefore, Benson’s structural stage was not met and Phase III (external validity) could not be completed. Given that the primary issue with the data was the low relationships between items, this is discussed more extensively in Chapter 5. Additionally, a proposal for the future development of a cognitive engagement scale for the assessment context is provided.

Chapter 5

Discussion

The purpose of the current study was to examine construct validity evidence for the CES-E using Benson's strong model of construct validation. Of particular focus was an examination of the structure of the CES-E as well as an examination of whether CES-E scores were related to scores on other construct measures in expected ways (external validity). Neither of the models tested (one-factor or three-factor) using CFA fit the CES-E data in an interpretable way. The failure to support an interpretable factor structure made it impossible to examine external validity. That is, any scores, subscale or total, obtained from the scale would not be meaningful due to a lack of evidence supporting a meaningful factor structure. Therefore, it is unwise to correlate such scores with other constructs. This chapter examines the possible reasons for why the findings did not support either of the theoretically driven models, the potential for developing a unidimensional scale consisting only of items written to assess deep cognitive engagement, and what findings from this study may mean in regards to the application of cognitive engagement theory in an assessment context. Finally, limitations of the current study and suggestions for future research are provided.

Phase I: Factor structure

Testing of the unidimensional model in Phase I resulted in decent absolute model-data fit, but poor incremental fit. Testing of the three-factor model yielded no admissible solution. Both of these findings can be attributed to low inter-item correlations. Generally, low inter-item correlations can be attributed to low item variances, items not measuring the same construct, or varying item distributions. However, in the current study item variances do not appear to be a plausible cause, and item distributions appear to play only a small role, leaving item content as the primary explanation for the low correlations for both samples.

One-Factor Model

In terms of the one-factor model, some item distributions do vary from others, particularly those for items N1 (“I skipped the hard parts on the assessment”) and S4 (“If I was not sure about the answer to a question on the assessment, I picked the longest answer”). Both of these items are more positively skewed than the rest of the items. This suggests that the items may be extreme and thus not expected to correlate well with the rest of the items. Distributions of each of the items can be found in Appendix E.   While item distributions may have played a small role, the most likely reason for the low correlations is that the items do not measure the same content. .The one-factor model was tested due to a theoretical suggestion from the literature; however, the items were not written to measure a single factor. The items were written to address three factors (e.g. deep, shallow, no engagement). Therefore, it makes sense to assume that not all of the items are addressing the same content. For instance, the item (N1) “I skipped the hard parts on the assessment” is likely measuring something different than the item (D5)“When approaching the questions on the assessment, I planned out or organized my response prior to providing my answer.” Thus, it is possible the items are not measuring the same construct.

Three-Factor Model

When examining the three-factor model, item distributions as a cause of low correlations are discussed for certain factors. However, like the one-factor model, it is item content that is the most likely cause of the low correlation. Items were expressly written to represent one of three factors. However this does not guarantee that the items written for a given factor are actually measuring the same content.

Deep Items

Although still low, the items meant to address deep cognitive engagement had higher inter-item correlations (.15 to .35) than the shallow items (-.04 to .17) or the no engagement items (-.01 to .30). Despite slightly higher inter-item correlations of deep over shallow and no engagement items, several concerns regarding item content still exist among the deep items. For example, recall that students completed the CES-E in Phase I after completing a selected-response assessment instrument related to health and wellness. One of the CES-E items (D5),“When approaching the questions on the assessment, I planned out or organized my response prior to providing my answer” had a distribution in which the “Neither Agree Nor Disagree” option was selected the most. This could be an indication that the item is confusing to students. This may be the case given the content of the cognitive assessment instrument that the students took just prior to the CES-E. In other words, it probably does not take a lot of planning and organizing to answer a selected-response cognitive item written to address health and wellness. Therefore, it may not be that item D5 itself is unclear, but that is not applicable to the test content and/or format of the cognitive test that preceded the CES-E administration. If there is a problem with the item content in relation to the content/format of the cognitive assessment instrument to which the CES-E items refer, this could be problematic given that researchers designed the CES-E to be used in conjunction with all types of assessment tests (e.g., health and wellness, quantities reasoning, information literacy).

Shallow Items

As previously mentioned, no correlation between any shallow items was higher than .17 and a few reasons may help to explain why this is the case. To start, in terms of item distributions, the strong selection at one end of the response scale gave item S4 a skewness value of 2.1. The rest of the shallow items had small, negative skews. Given this information, it is not difficult to see why S4 has the lowest correlations within the items designed to measure shallow engagement. Overall, it appears like an extreme item and therefore does not fit well with the rest.

However, the bigger problem with this facet of cognitive engagement is that the items may in fact not be measuring the same thing. In other words, after more careful consideration of item content, these items may not be representative of the same construct. For example, item S3"When answering the questions on the assessment, I looked for clues of how to respond within the test itself" correlated much more highly with items meant to represent deep cognitive engagement than with other shallow items, which could indicate that students consider it a strategy to deeply process material. This idea does make sense given that many K-12 teachers suggest to their students that they use this strategy to more efficiently complete their standardized tests. This is not necessarily a bad thing since it does cause students to engage with the assessments, but in terms of the CES-E, the item may not be addressing shallow cognitive engagement at all. For item S2,"when answering the questions on the assessment, I considered how those reviewing the answers would want me to respond", students chose "Slightly Agree" most often, but only by a slim margin: "Neither Agree Nor Disagree" was chosen by only a few less students, suggesting that the item may be confusing to students (Velez & Ashworth, 2007) and thus also not measuring the same construct as the rest of the items.

Finally, item S1, "If I didn't understand a question on the assessment, I narrowed down the answers and then randomly picked one of the remaining responses" students tended to choose a variety of answers, which means that the item shows promise in terms of actually measuring shallow engagement. To clarify, having variability in answers helps to discriminate students who have little of a characteristic from those who have a high amount; item S1 meets the variability criteria. Furthermore, the items shows promise because students did not seem to find the item confusing as the midpoint was the least chosen answer, and item content appears to address a strategy representative of shallow cognitive engagement. Researchers may wish to retain this item in future iterations of the scale should the measure continue to include shallow items.

Though one shallow item may show promise in tapping into the construct of shallow cognitive engagement, the same cannot be said of the other shallow items. Instead, the remaining three items appear to be extreme, confusing, or simply refer more to being assessment savvy (i.e. utilizing strategies previously taught by teachers). Obviously, these are major issues as they ultimately mean that the items are not measuring what they were intended to measure. What is unclear at this point is whether the items were just in fact not representative of the construct, or if shallow engagement is not easily transferred to an assessment context.

No Engagement Items

As with the shallow subscale items the N1 item distribution had a fairly high positive skew with a value of 2.98, while the other two items had skew values close to zero, providing more evidence as to why the item is so poorly correlated with N2 and N3. Overall, the item appears to extreme. This finding does make sense considering that given that entering freshmen made up the sample, they may have wanted to make a good impression or were not sure how the data would be used; therefore, they indicated that they completed the entire assessment. Additionally, given that the assessment they completed was multiple choice, students that did not skip answering any items perhaps thought that they weren't skipping anything even if they did not carefully read the item.

For items N2 ("I did not check my answers for mistakes") and N3 ("When working on the assessment, I guessed a lot") students chose answers from both ends of the scale fairly evenly, which explains the fact that they have a moderate correlation for the sample and may be useful for future iterations of a cognitive engagement scale. That is, having such variability among responses is good for discriminating students who possess a large amount of a characteristic from a low amount, and items N2 and N3 have a good amount of variability in responses. Furthermore, item content appears to address the intended construct without any extreme wording. Therefore, if no engagement is examined in future studies, items N2 and N3 might be good items to retain, but it is important to keep in mind that the correlations were still low between the items. In other words, any future use of the items should be done with caution. As for item N1, it could be re-worded to be less extreme or dropped and replaced by new items.

Phase II: Cross-Validation

Just as in Phase I, testing of the unidimensional model in Phase II resulted in decent absolute model-data fit, but poor incremental fit. Testing of the three-factor model yielded no admissible solution. Item distributions and item content are discussed as plausible reasons for the low correlations.

One-Factor Model

In terms of item distributions for the one-factor model, they varied widely across the 12 items, which can be seen in Appendix E. Items S4 and N1 again were much more positively skewed than the rest of the items, suggesting again that students perceive the items as more extreme. Another example of differing item distributions comes from item D5, which yielded a high number of "Neither Agree Nor Disagree" responses. The high number of mid-point responses suggests that students could have found the item confusing in some way

When it comes to item content, Phase I results were reproduced. Items were not written to measure one overall construct, so it makes sense to assume that not all of the items are addressing the same content. In terms of item distributions, they vary widely across the 12 items, which can be seen in Appendix E. Items S4 and N1 again were much more positively skewed than the rest of the items, suggesting again that the items are perceived as more extreme to students. Another example of differing item distributions comes from item D5, which yielded a high number of "Neither Agree Nor Disagree" responses. The high number of mid-point responses also suggests that students could have found the item confusing in some way.

Three-Factor Model

Deep Items

As in Phase I, Phase II deep engagement items tended to have higher inter-item correlations (.14 to .40) than shallow engagement items (.02 to .20) or no engagement items (.09 to ,32) The Phase II data yielded the same results as Phase I.

One interesting finding came from the item D5, “When approaching the questions on the assessment, I planned out or organized my response prior to providing my answer”; participants chose “Neither Agree Nor Disagree” the most frequently. It is an interesting finding in this phase considering that students completed an assessment designed to test quantitative and scientific reasoning, and thus were required on many items to work out mathematical problems. Therefore, it would be expected that the item would not have as many midpoint responses as it did in the Phase I sample; in fact, a slightly higher percentage of students (32.5% compared to 29.1% in Phase I) selected the midpoint response. It is possible that students did not consider calculating answers to be the same as planning out or organizing responses. In other words, it is possible that the item wording is confusing or does not seem directly relevant to the test content. However, considering that the item has performed well in the past, the current researcher does not recommend immediately removing it from the scale but instead attempting to determine how students are interpreting this item.

In regards to the rest of the items meant to address deep cognitive engagement, items D2, D3, and D4 had some variability in their responses, which is sometimes good in terms of differentiating among students. For item D1 (“When preparing to answer the questions on the assessment, I stopped to reflect on the information provided”), nearly half of students endorsed the “Agree” option. Such a high percentage of endorsement on an option that was not the most extreme may be a signal that students may have been answering in a socially desirable way. Therefore, such a finding would need to be remembered in future iterations of scales that included deep engagement items. Overall, however, deep items did perform slightly better than either the shallow or no engagement items.

Shallow Items

Shallow items once again had the lowest inter-item correlations among the three proposed factors, indicating little shared variance between the items. Examination of item distributions produced results similar to those found in the Phase II sample. For example, item S4, “If I was not sure about the answer to a question on the assessment, I picked the longest answer” again had a moderate positive skew value of 1.61 while the rest of the item distributions had a negative skew. Examination of frequencies of the seven answer options indicated that students overwhelmingly chose “Strongly Disagree” as their response to the item, indicating once again that the item is probably an extreme one. Additionally, item S3, “When answering the questions on the assessment, I looked for clues of how to respond within the test itself”, correlated much more highly with deep items than other shallow items, suggesting that the item does not align with the content of the other shallow items. Recall that when discussing results for Phase I, this finding was attributed to the possibility that item S3 is more indicative of a student being savvy about taking assessments rather than being shallowly engaged with the material. Considering this finding was produced in both samples, and examination of item content does in fact diverge from the content of the other three shallow items, it seems appropriate to drop it from the shallow subscale.

In regards to the other two items, the findings related to item S2 “When answering the questions on the assessment, I considered how those reviewing the answers would want me to respond” again indicated that it may have been confusing for students. Finally, item S1 “If I didn’t understand a question on the assessment, I narrowed down the answers and then randomly picked one of the remaining responses” had a wide range of responses, reinforcing the idea that it may be useful to retain the item in any revised version of the shallow subscale.

In summary, the shallow items did not perform well in either sample. This is consistent with findings related to previous versions of the CES-E. While the original CES shallow subscale items did form a factor, there were only two items (Smiley & Anderson, 2011). In the subsequent CES-2 version when researchers attempted add shallow items in order to capture the theoretical breadth of the shallow factor, the shallow subscale performed poorly (Charsha, Smiley, & Anderson, 2012). Should any future researchers attempt to correct the item issues identified with the CES-E, it would be advised to first revisit the literature to determine how the representativeness of the items can be improved. Researchers may also wish to consider the possibility that shallow engagement, as it is conceptualized in the classroom, does not exist in an assessment context.

No Engagement Items

Given the problems associated with items in the Phase I and the deep and shallow items in Phase II, it was no surprise to find problems associated with no engagement items in Phase II as well. Results from Phase I replicated in Phase II. For example, the item (N1) “I skipped the hard parts on the assessment” was an extreme item, with the overwhelming majority of students choosing that they strongly disagreed with the statement, leading to nearly non-existent correlations between the item and the other two items meant to assess no cognitive engagement. Furthermore, for the other two items (“I did not check my answers for mistakes (N2)” and “When working on the assessment, I guessed a lot (N3)”) had a much stronger correlation (though still small) than the correlations found between most other items on the CES-E, with response options chosen fairly evenly. Thus, it would appear that the item “I skipped the hard parts on the assessment” needs to be replaced or rewritten to be less extreme and the other no engagement items have potential to work well on any future scale intended to measure no cognitive engagement. The most important point is that currently, only two items appear to represent no engagement on the CES-E, and thus no engagement runs the risk of being underrepresented.

Conclusion from One-Factor and Three-Factor Models

Given the many issues associated with the one and three-factor models that were tested, the researcher cannot move forward with examining the external validity (Phase III) of scores derived from the existing version of the CES-E. It is known that neither of the theoretically supported models fit the data, thus evidence for the structural stage for Benson’s strong model for construct validation has not been provided. However, the following section examines a possible next step in the research based on previous CES studies as well as findings from the current study.

A Potential One-Factor “Deep” Model

It is clear that there are numerous issues with the items on the CES-E. Therefore, the scale warrants major revision. However, which subscales to retain and which items to revise may warrant discussion. Studies on previous versions of the CES-E (CES; CES-2) consistently showed that items written to assess deep engagement perform better than items written to assess shallow and no engagement. For example, Smiley and Anderson (2011), who studied the original CES, found initial support for a two-factor model: deep and shallow. The deep engagement factor consisted of three items and the lowest factor loading was .56. Cronbach’s alpha was calculated to be .57. The shallow factor consisted of only two items and was believed to not represent the full breadth of shallow engagement in an assessment context. Charsha, Smiley, and Anderson (2012), who studied the CES-2, found no support for shallow items, but did find that the inter-item correlations among five items (three from Smiley and Anderson’s study, plus two new ones) written to represent deep engagement were all at least moderate in nature and that a deep engagement factor was supported. Reliability for items written to assess deep cognitive engagement was .74. Thus, items meant to assess deep cognitive engagement have performed well in the past.

In addition to the findings of previous studies, in the current study the correlations among deep items were the strongest. Based on the relative strength of the inter-item correlations between deep items in the current study and past support for the deep items (in which the correlations were stronger than in the current study), the researcher suggests consideration of a one-factor model using only the items written to assess deep cognitive engagement in the assessment context.

One-Factor Deep Model

A five-item unidimensional model, consisting of items meant to measure deep cognitive engagement was fit to the Phase I and Phase II samples. Fit indexes for both models can be viewed in Table 5.

Table 5. *Fit Indexes for One-Factor Deep Model*

	N	df	$\chi2_{S\text{-}B}$	*p*	SRMR	RMSEA	CFI
Phase I Sample	602	5	5.91	0.31	0.04	0.02	0.99
Phase II Sample	372	5	4.75	0.44	0.05	0.00	1

One important point to note is the improvement of the CFI in both samples over the CFI indices for the original proposed one (CFI= .75 for Phase I and .69 for Phase II) and three-factor models (non-convergent model in both samples). In addition to global fit, localized misfit was examined through correlation residuals and there were no particularly large residuals associated with either sample. Residuals for Phase I can be found in Table 6; residuals for Phase II can be found in Table 7. Factor loadings for both samples, shown as the correlation between the each item and the latent factor, can be found in Figures 3 and 4. Because the loadings are in a correlation metric, they can be squared to obtain the variance shared between the factor and each item.

Figure 3. *Factor Loadings for the One-Factor Deep Model (Phase I)*

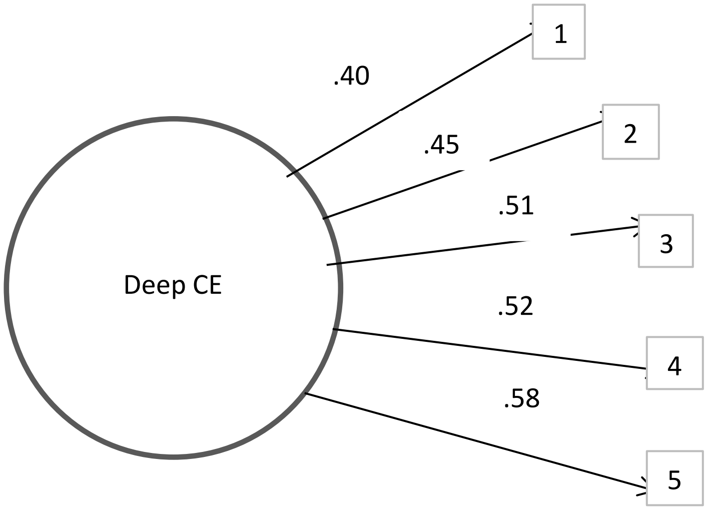


Figure 4. *Factor Loadings for the One-Factor Deep Model (Phase II)*

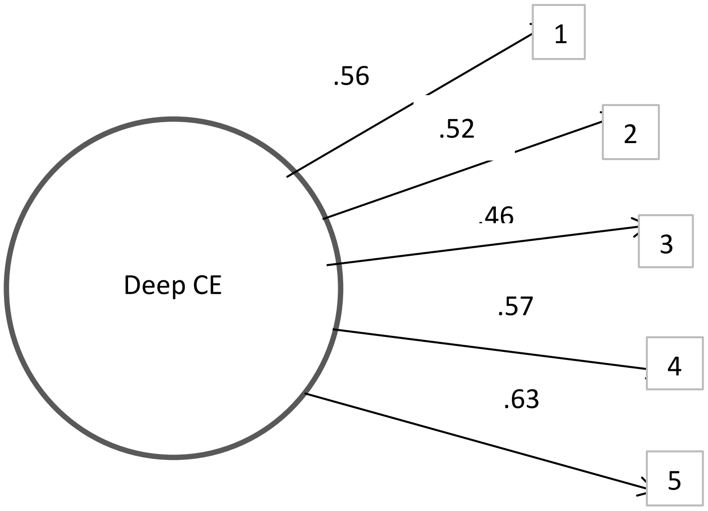


Table 6. *Residuals for Phase I One-Factor Deep Model*

	D1	D2	D3	D4	D5
D1	0.00				
D2	0.05	0.00			
D3	0.08	-0.03	0.00		
D4	-0.04	-0.04	-0.03	0.00	
D5	-0.09	-0.03	-0.02	0.05	0.00

Table 7. *Residuals for Phase II One-Factor Deep Model*

	D1	D2	D3	D4	D5
D1	0.00				
D2	0.02	0.00			
D3	**0.12**	-0.10	0.00		
D4	-0.05	0.01	-0.05	0.00	
D5	-0.07	0.01	-0.04	0.04	0.00

In addition to examining fit and factor loadings, the current researcher calculated reliability and variance extracted for each sample. Reliability was marginal for both Phase I (ω = .62) and Phase II (ω = .68) samples. Furthermore, the proportion of variance extracted from the items by the latent factor was quite low in both cases. For the Phase I sample only 24.8% of the variance from the items was extracted by the latent construct; as for the Phase II sample, only 30.74% of the variance was extracted. Because both percentages fall below 50%, this indicates that there is more measurement error than variance explained by the factor when it comes to item variance. Furthermore, the correlations among the deep total score and total scores on the external validity measures (AEQ, EVC subscales, and SOS effort) described in Chapter 3 were calculated and although all of the correlations were in the hypothesized directions (i.e. deep cognitive engagement is positively correlated with expectancy, value, and effort and negatively correlated with cost and academic entitlement), the correlations were much lower than anticipated. These correlations can be found in Table 8. Thus, although the one-factor deep model shows more promise in terms of global and localized fit than the original proposed models, there is still concern regarding the low correlations between the items and thus the complications resulting from the issue.

Table 8. *External Validity Evidence Using Deep CES-E Total Scores**

	DeepCE (Phase I)	DeepCE (Phase II)
SOS Effort	0.29	0.37
Expectancy	0.14	0.16
Value	0.26	0.20
Cost	-0.17	-0.21
Academic Entitlement	-0.10	-0.18

**Note: Listwise deletion was used for the correlations, so there are fewer participants than were included in the CFAs. N=569 for Phase I, N=347 for Phase II.*

Unidimensional Deep Model and Implications for Theory

This study, along with studies examining previous versions of the CES, has produced some evidence in support of developing a measure of cognitive engagement that consist of items researchers have labeled as measuring “deep” cognitive engagement. However, producing multiple items that successfully measure shallow or no cognitive engagement has been more difficult. Shallow items have been particularly challenging, which could indicate one of two things: either the shallow items that have been piloted are not representative of the construct, or the empirical interpretation of cognitive engagement theory as applied to an assessment context is incorrect. To be clear, Benson’s (1998) substantive stage consists of two parts: the theoretical (i.e. defining the construct) and the empirical representation of the theoretical construct (i.e. items accurately represent the construct). In the current study, it is unclear if cognitive engagement theory needs to be revisited, or if the problem lies is that items written to measure the construct are not representative of the theory in an assessment context. Disentangling what the underlying problem is has been made difficult by conflicting findings. However, if researchers were to pursue a cognitive engagement scale that consisted of only items written to measure what researchers have termed deep engagement, this would raise new questions. One of the questions being "Assuming higher scores mean deeper engagement, are students who score low shallowly engaged or just low in deep engagement?" Recall that previous classroom-based research (Green & Miller, 1996) has suggested that cognitive engagement is a unidimensional construct. It could be that cognitive engagement in an assessment context is unidimensional, but perhaps in a different way than originally hypothesized. In other words, while the possibility still exists that shallow engagement is at one end of the continuum and deep engagement at the other, it could be that a measure that assesses only "deep" engagement is capturing the continuum for an assessment setting in which the stakes are low for students. If students score low on such a measure, then they would simply not be deeply engaged with the assessments. In other words, the construct of cognitive engagement is narrower within an assessment context than it is for a classroom context. Therefore, classroom-based research on cognitive engagement may not be directly applied to the assessment context. This idea could explain why shallow items have behaved so poorly both in the current study and past studies.

Limitations, Future Directions, and Conclusion

One limitation applied to this study. The limitation was that the KWH (Phase I) and NW-9 (Phase II) assess different content, which could have affected results. The primary reason that test content likely affected results is that students may have perceived certain items as more applicable for one test over the other. This limitation makes it difficult to explain why some results differed across samples. If the content of the cognitive measure does impact how the items function, this is concerning given that one reason for developing a cognitive engagement measure was to use it to assess engagement on assessment instruments measuring different content areas.

Future research related to developing a measure of cognitive engagement could take one of two directions, the first of which is to continue to work on the scale aimed at measuring multiple factors. That is, developing new and additional items meant to address deep, shallow and no engagement. A second possibility is to further develop a unidimensional scale aimed at measuring students' cognitive engagement using only items that are developed to measure what researchers call "deep" cognitive engagement. While current classroom theory may support the first approach, recent findings support the second approach, for two reasons: 1) Shallow items have shown to be particularly difficult to interpret in terms of their misfit, as there is no clear pattern as to why they do not work and 2) Administrators and faculty would likely find results pertaining to students' deep engagement still useful for interpreting student test scores, even without information about shallow or no engagement. Put another way, it is probably sufficient to know that a student is low in what researchers call deep engagement when it comes to interpreting the validity of their assessment scores, rather than spend more resources trying to determine where shallow items belong.

Though results from the current study and previous studies seem to suggest that further development of a scale comprised of only items written to address "deep" engagement could be a fruitful endeavor, it would first be beneficial for researchers to conduct think-alouds with students while they are completing the cognitive assessments (e.g. KWH or NW9). Such a process would assist researchers in determining what cognitive engagement strategies the students employ in the assessment context. The information collected from the students could then be organized into a more accurate picture of what cognitive engagement in an assessment context looks like; this is a particularly important point given that cognitive engagement in the assessment setting may not exist in the same form as it does in a classroom setting.

Another think-aloud process could be used to collect student feedback on current CES-E items in order to establish which items students find confusing or not applicable to the assessment context. For example, think-alouds could either support or disconfirm the hypothesis that the item “When approaching the questions on the assessment, I planned out or organized my response prior to providing a answer” (D5) did not appear directly applicable to the assessments students completed as part of the University’s assessment process. Furthermore, any new items developed as a result of either the think-alouds just described could be reviewed by yet another group of students. However, as important as student feedback can be when it comes to item development, items need to be reviewed or revised by content experts as well. Because there have not been a lot of studies that have examined cognitive engagement in the assessment setting, it would likely be difficult to find a content expert in that domain. Thus, the context experts would need to consist of a mix of those who are knowledgeable about how students operate in an assessment context (e.g. test-taking motivation experts) and those with a solid background in cognitive engagement in the classroom context. Together they would be able to provide a more complete picture of what cognitive engagement in an assessment context may look like.

In conclusion, cognitive engagement has proven to be a tricky construct to measure in an assessment setting, mainly because a sound measure has yet to be developed. However, research on the CES-E and its predecessors suggest that researchers may wish to re-examine the construct of cognitive engagement as it manifests in the assessment context. This work would be aided by incorporating feedback from students and content experts in order to paint a more accurate picture of what cognitive engagement in an assessment context looks like.

Appendix A

The CES-E

Please think about the test you just completed. Please respond to the items using the 7 point scale below.

1	**2**	**3**	**4**	**5**	**6**	**7**
Strongly Disagree	**Disagree**	**Slightly Disagree**	**Neither Agree Nor Disagree**	**Slightly Agree**	**Agree**	**Strongly Agree**

1. When preparing to answer the questions on the assessment, I stopped to reflect on the information provided.
2. I skipped the hard parts on the assessment.
3. If I didn’t understand a question on the assessment, I narrowed down the answers and then randomly picked one of the remaining responses.
4. When reading the questions on the assessment, I tried to figure out how the material presented fit with what I had learned in my courses.
5. I did not check my answers for mistakes.
6. When answering the questions on the assessment, I considered how those reviewing the answers would want me to respond.
7. When taking the assessment, I went back over material provided on the test if I did not understand it.
8. When working on the assessment, I guessed a lot.
9. As I was working on the assessment, I asked myself some questions as I went along to make sure the material made sense to me.
10. When answering the questions on the assessment, I looked for clues of how to respond within the test itself.
11. When approaching the questions on the assessment, I planned out or organized my response prior to providing my answer.
12 If I was not sure about the answer to a question on the assessment, I picked the longest answer.

***Items intended to measure deep engagement are 1,4,7,9, and 11 (D1-D5). Items intended to measure shallow engagement are 3,6,10, and 12(S1-S4). Items intended to measure no engagement are 2,5, and 8 (N1-N3).**

Appendix B

Academic Entitlement Questionnaire (AEQ)

The following items are asking about your personal attitudes about the college experience. Not all students feel the same way or are expected to feel the same way. Remember, **there are no right or wrong answers. Just answer honestly.**

Please respond by indicating how much you agree or disagree with each statement using the response options 1 (Strongly disagree) to 7 (Strongly agree).

1	**2**	**3**	**4**	**5**	**6**	**7**
Strongly Disagree	**Disagree**	**Slightly Disagree**	**Neither Agree Nor Disagree**	**Slightly Agree**	**Agree**	**Strongly Agree**

1. If I don't do well on a test, the professor should make tests easier or curve grades.
2. Professors should only lecture on material covered in the textbook and assigned readings.
3. Because I pay tuition, I deserve passing grades.
4. If I am struggling in a class, the professor should approach me and offer to help.
5. If I cannot learn the material for a class from lecture alone, then it is the professor's fault when I fail the test.
6. I should be given the opportunity to make up a test, regardless of the reason for the absence.
7. I am a product of my environment. Therefore, if I do poorly in class, it is not my fault.
8. It is the professor's responsibility to make it easy for me to succeed.

Appendix C

Expectancy-Value Cost Scale (EVC)

For this survey we are interested in your *general*, *overall* attitudes regarding all of the classes you have this semester. Please read each item and choose the response choice, using the 1 to 8 scale below, that best represents your feelings about how true each item is. If you *Completely Disagree* with the statement, mark a 1. If you *Completely Agree* with the statement, mark an 8. Or mark any number in between. There are no right or wrong answers. Just answer as honestly as possible.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Completely Disagree	**Strongly Disagree**	**Disagree**	**Slightly Disagree**	**Slightly Agree**	**Agree**	**Strongly Agree**	**Completely Agree**

1. I'm excited about the classes I'm taking this semester.
2. I expect to do well in my classes this semester.
3. I think my classes this semester are worthwhile.
4. Because of other things I'm doing this semester, I don't have as much time for my classes as I'd like.
5. I am confident that I can learn the material in my classes.
6. I think my classes will be useful to me.
7. I think my classes require too much time and effort for me to do well.
8. I am confident I will be successful this semester in my classes.
9. I think my classes this semester are interesting.
10. I think there are other things I'd rather do with my time than just focusing on my classes this semester.
11. I know I can understand the material in my classes.
12. I don't think I can invest the time and effort that is needed to do well in my classes.
13. I value the classes I am taking this semester.
14. Doing well in my classes may not be worth other things I have to give up.
15. I think my classes this semester are enjoyable to take.
16. I think my class schedule this semester is too stressful.

Appendix D

The Student Opinion Survey (SOS)

For items 67 through 76, please think about the test that you just completed. Mark the answer using the 1-5 point scale that best represents how you feel about statements 67 through 76 below.

1	**2**	**3**	**4**	**5**
Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**

67. Doing well on these tests was important to me.
68. I engaged in good effort throughout these tests.
69. I am not curious about how I did on these tests relative to others.
70. I am not concerned about the scores I receive on these tests.
71. These were important tests to me.
72. I gave my best effort on these tests.
73. While taking these examinations, I could have worked harder on them.
74. I would like to know how well I did on these tests.
75. I did not give these tests my full attention while completing them.
76. While taking these tests, I was able to persist to completion of the tasks.

Appendix E

Phase I item distributions

Figure 5a. *Item D1*

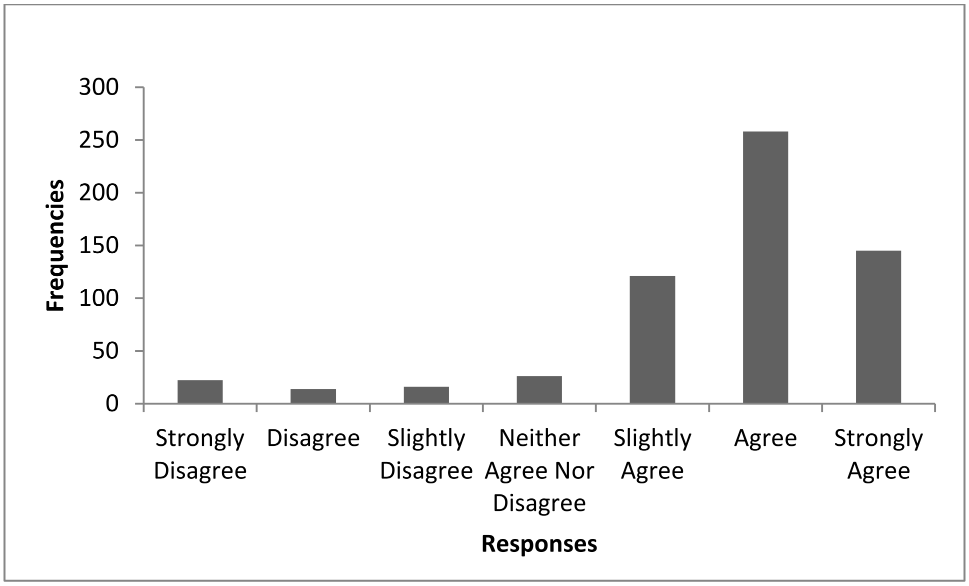


Figure 5b. *Item D2*

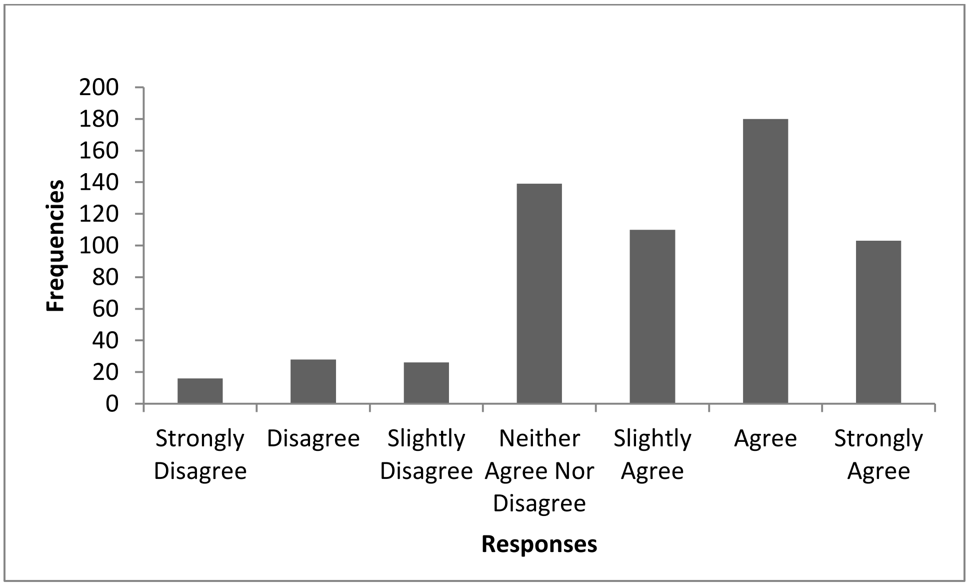


Figure 5c. *Item D3*

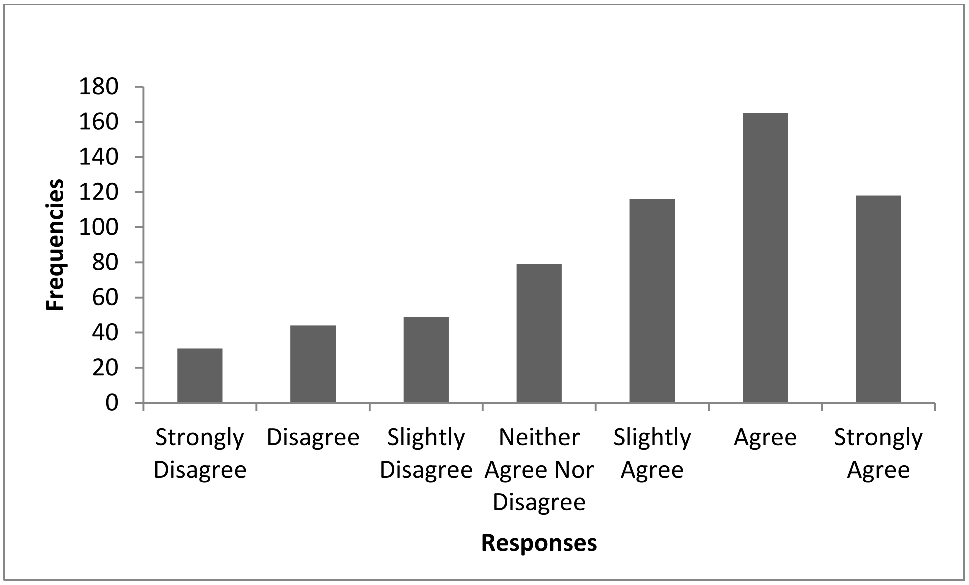


Figure 5d. *Item D4*

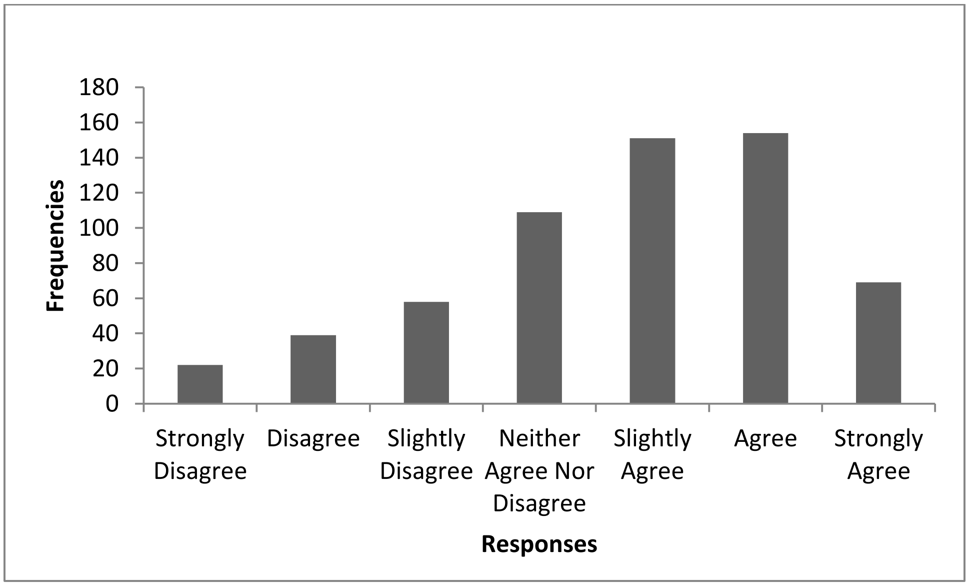


Figure 5e. *Item D5*

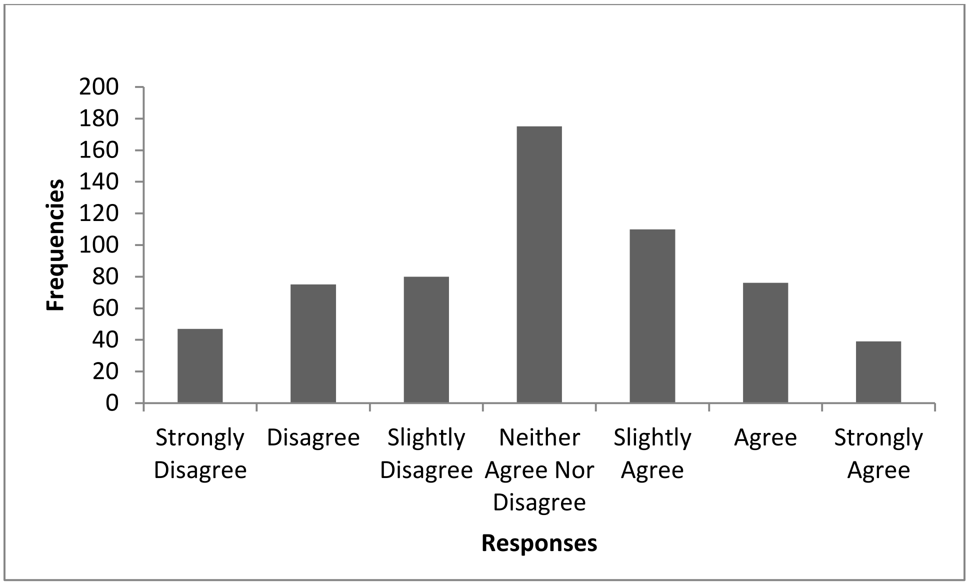


Figure 5f. *Item S1*

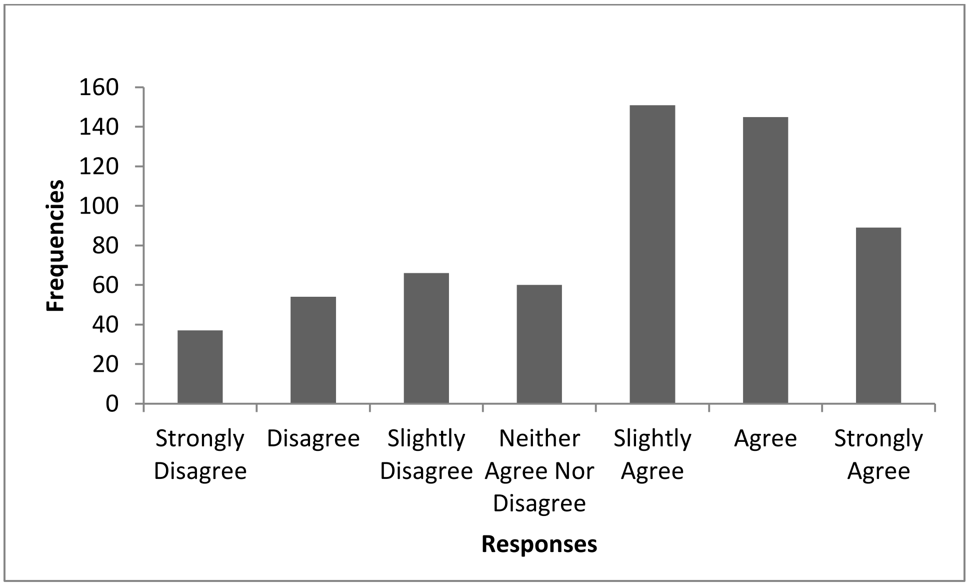


Figure 5g. *Item S2*

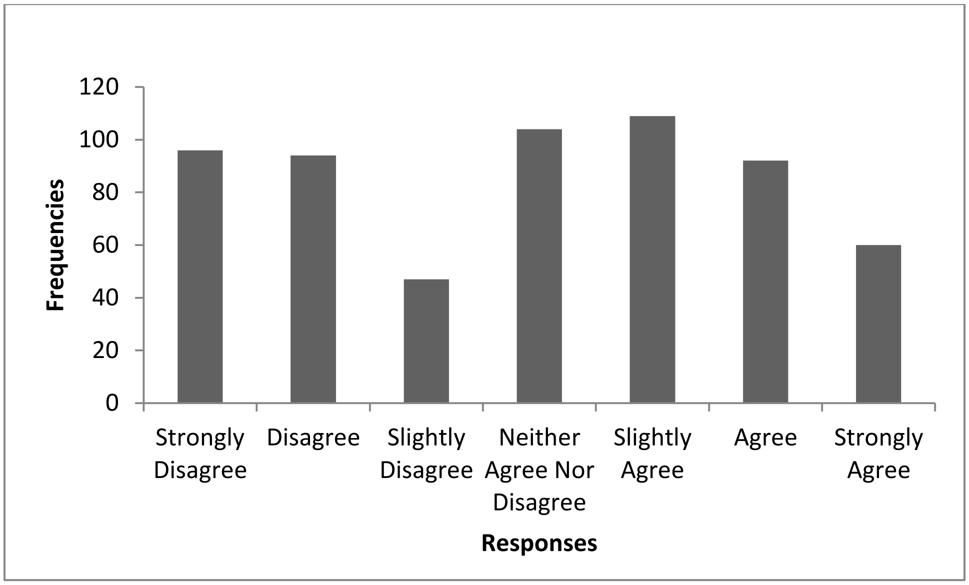


Figure 5h. *Item S3*

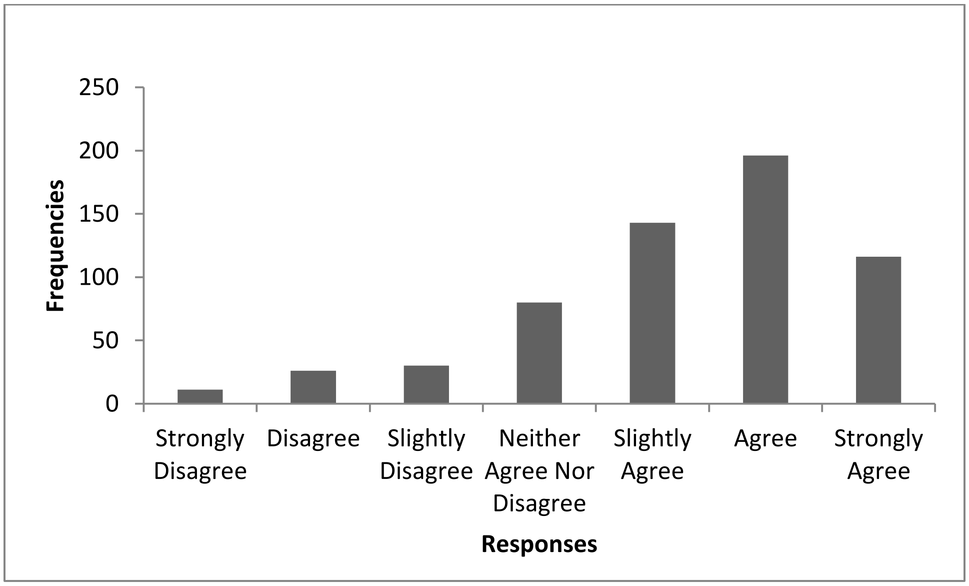


Figure 5i. *Item S4*

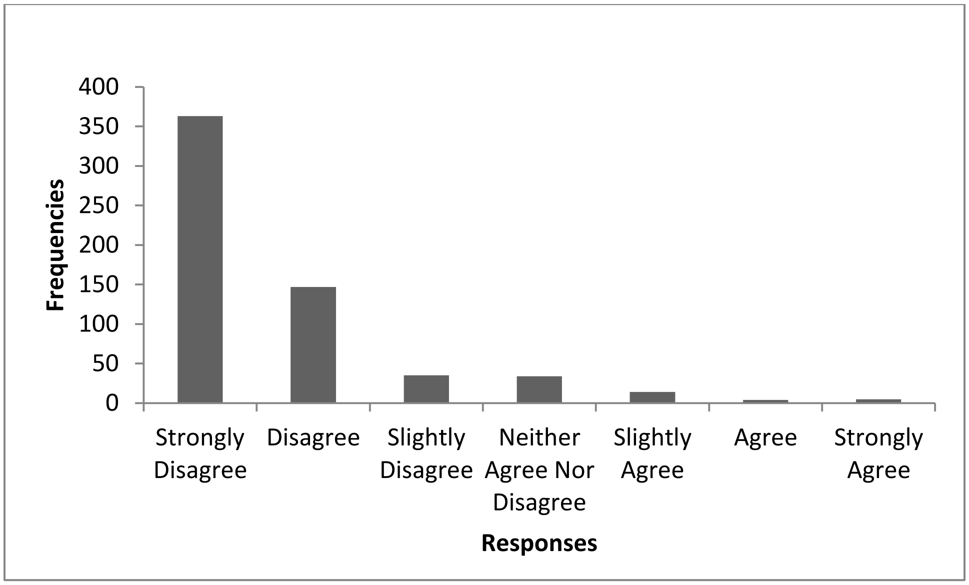


Figure 5j. *Item N1*

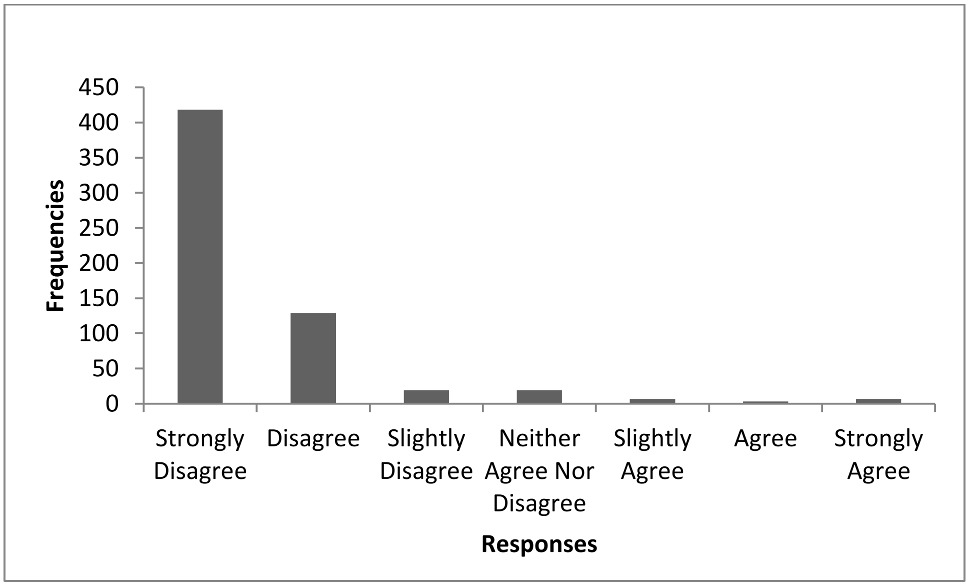


Figure 5k. *Item N2*

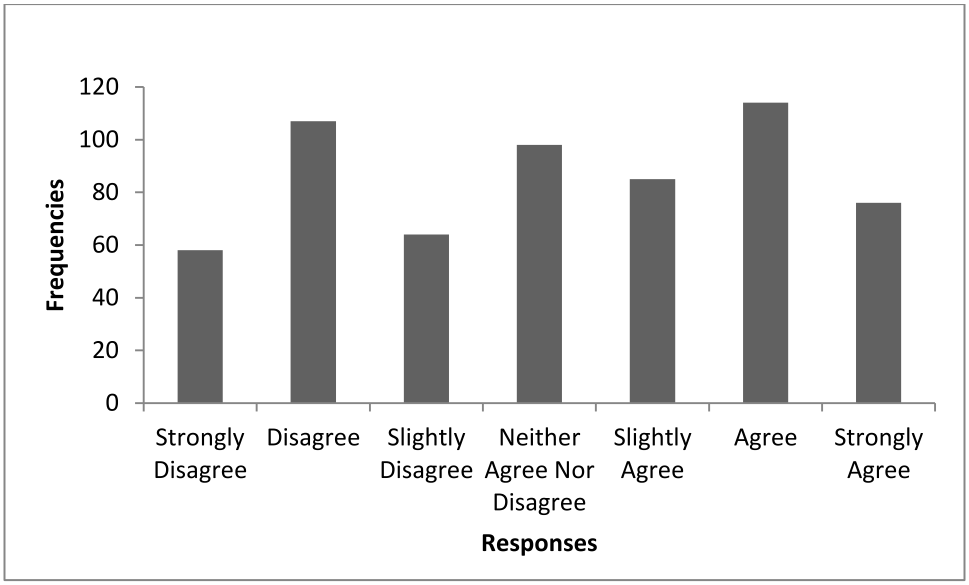
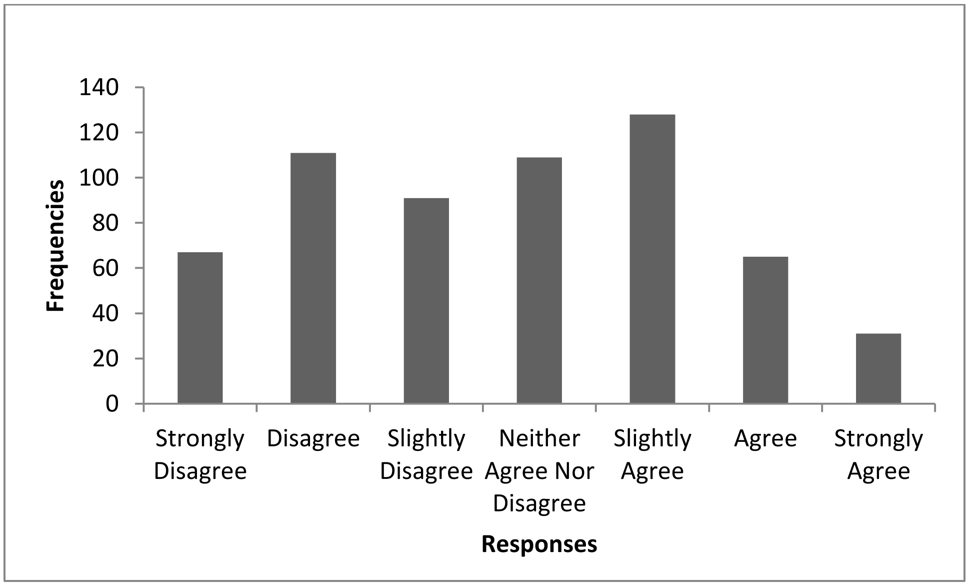


Figure 5l. *Item N3*



Phase II Item Distributions

Figure 6a. *Item D1*

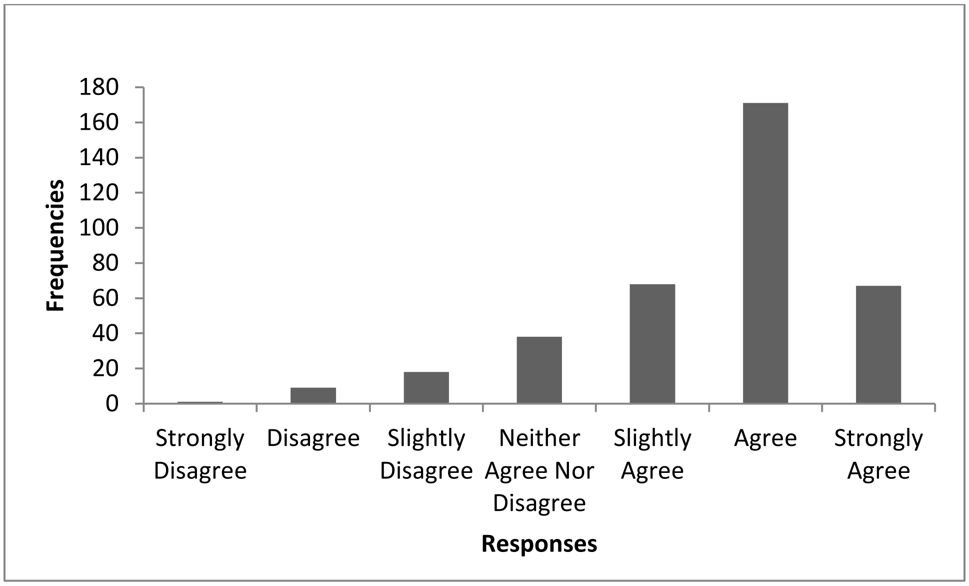


Figure 6b. *Item D2*

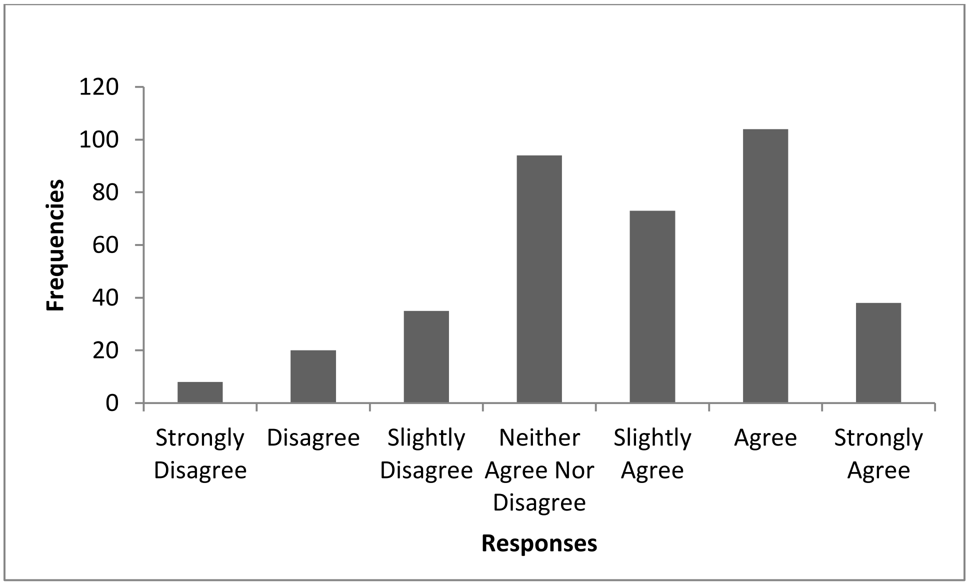


Figure 6c. *Item D3*

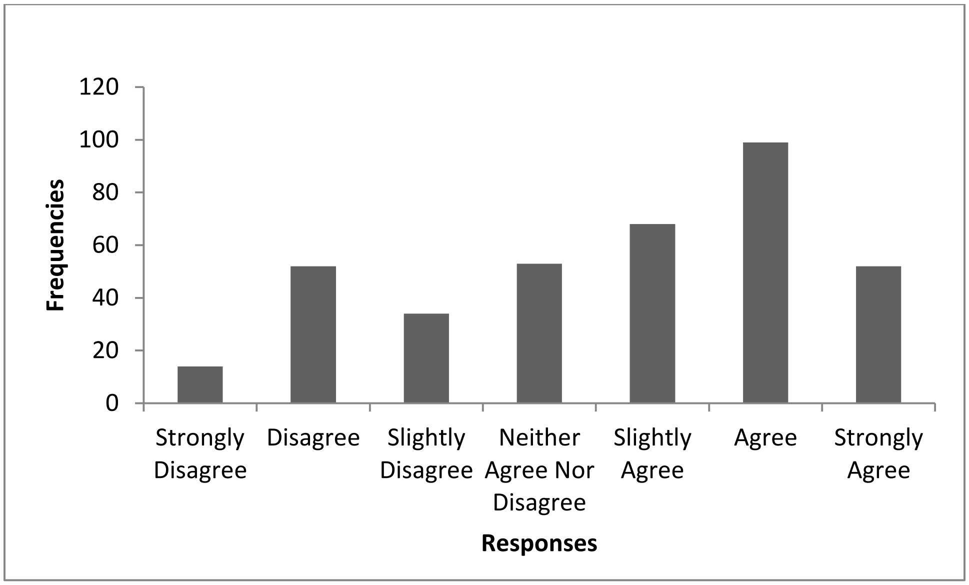


Figure 6d. *Item D4*

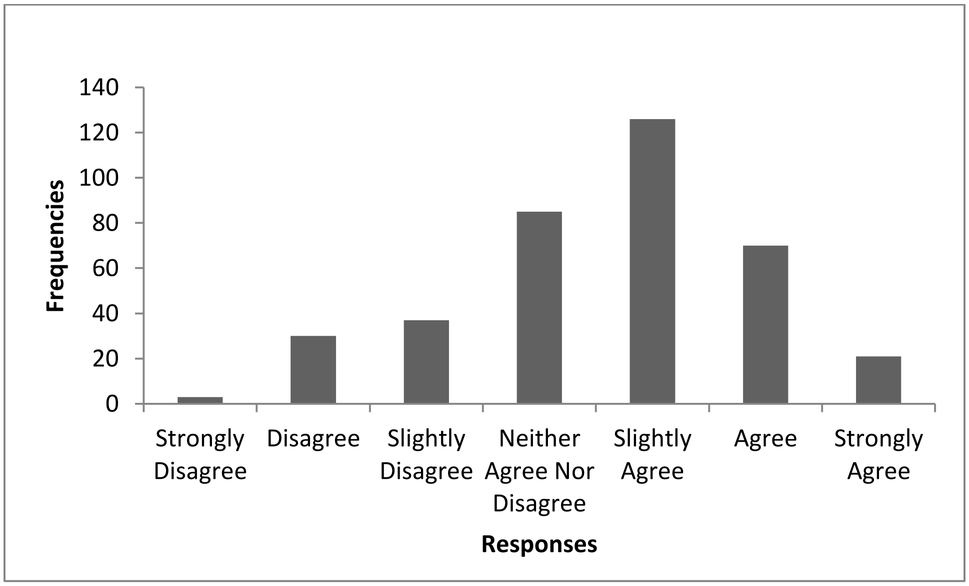


Figure 6e. *Item D5*

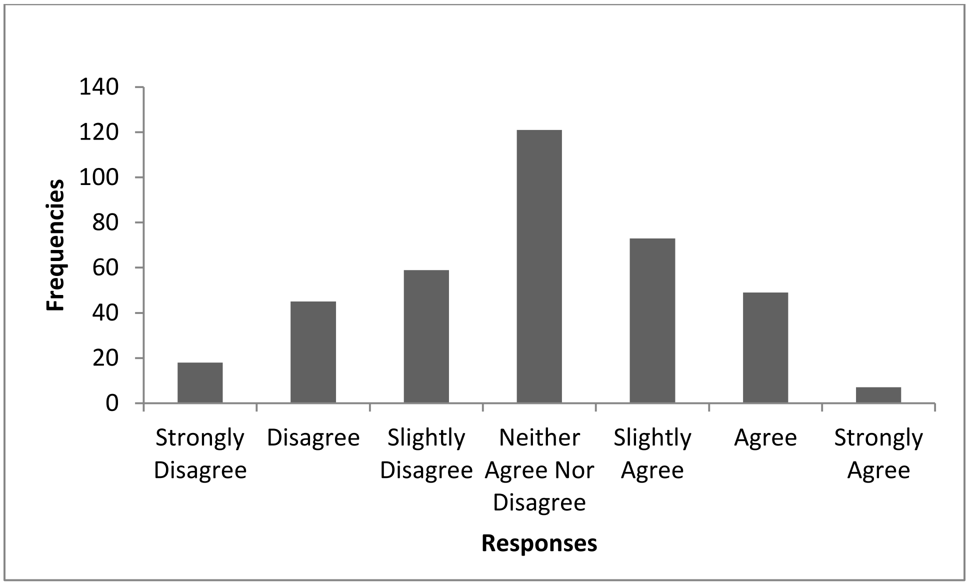


Figure 6f. *Item S1*

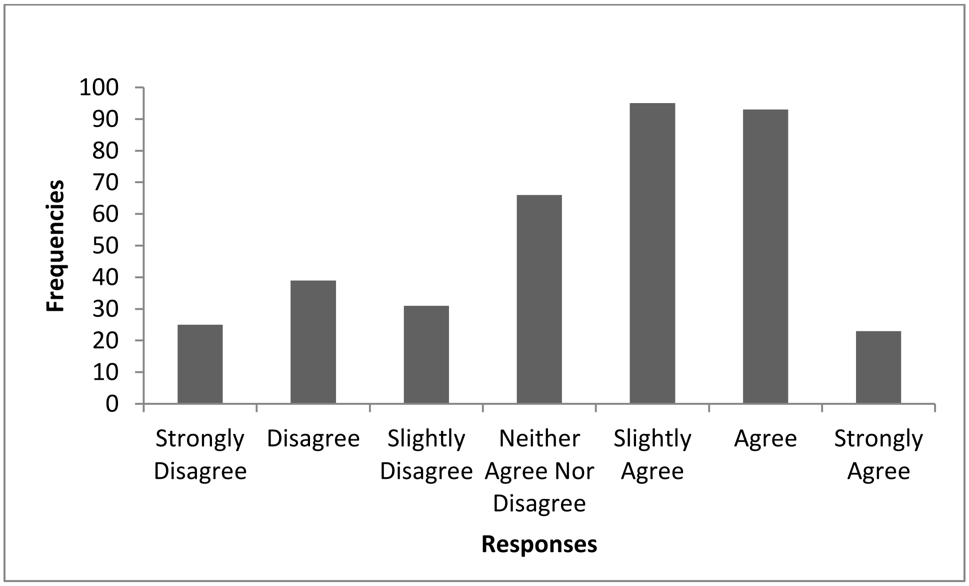


Figure 6g. *Item S2*

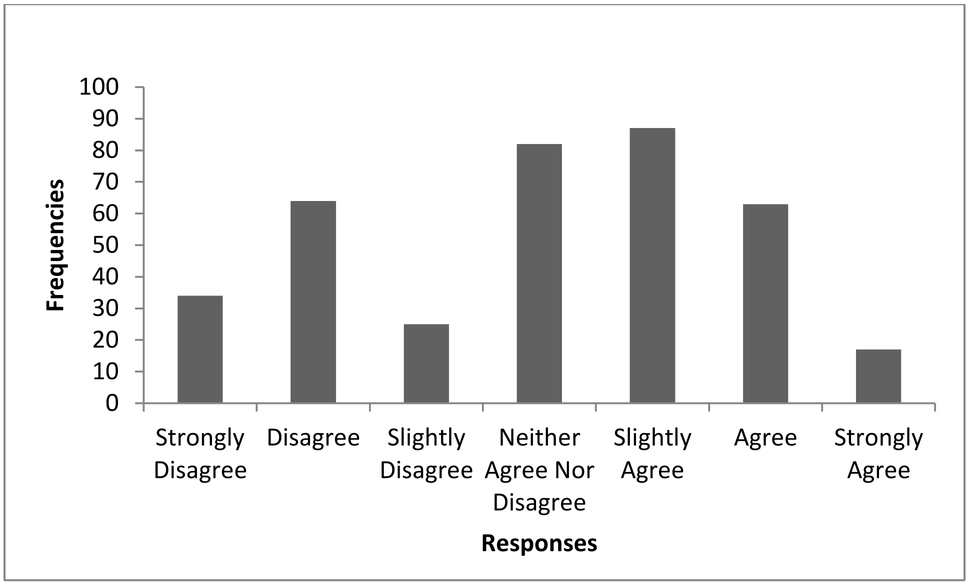


Figure 6h. *Item S3*

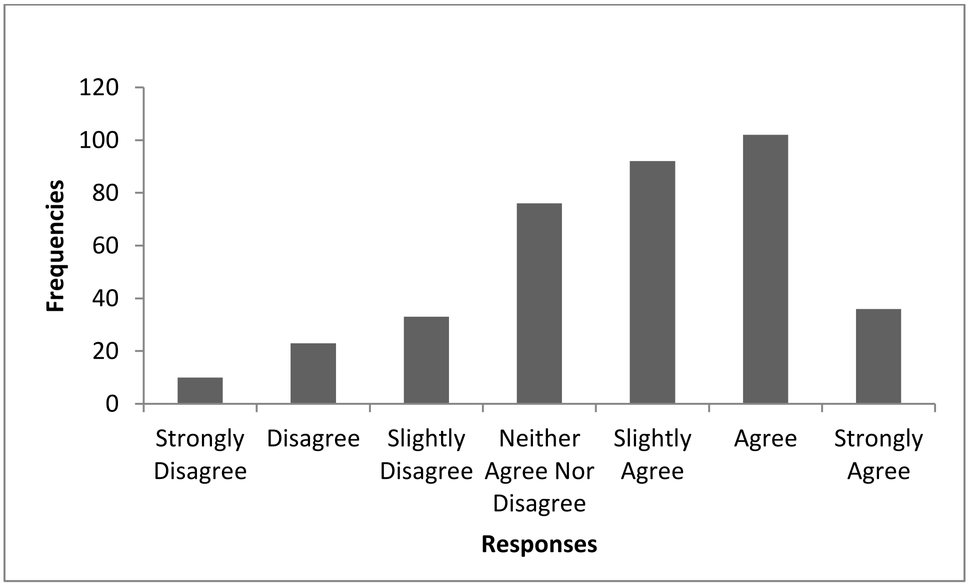


Figure 6i.*Item S4*

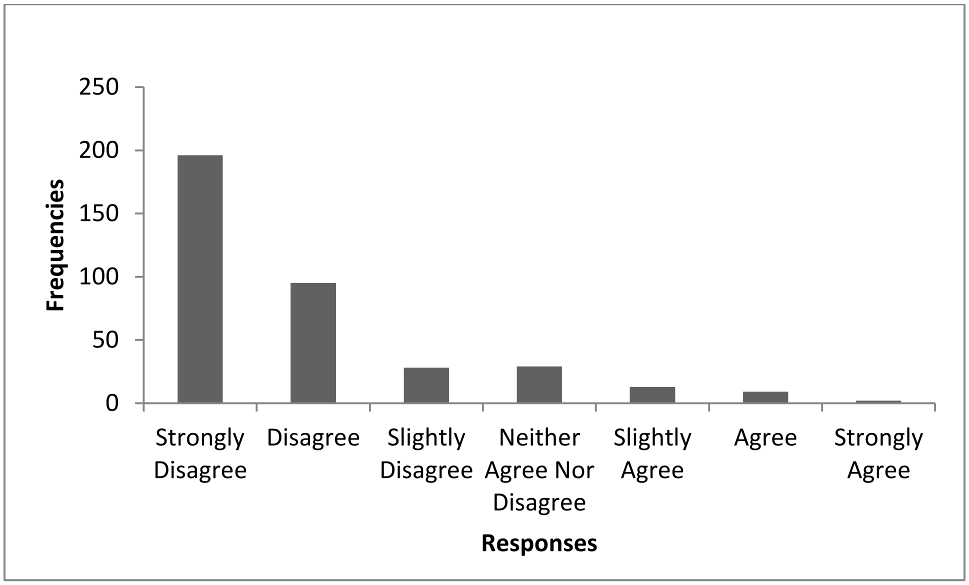


Figure 6j. *Item N1*

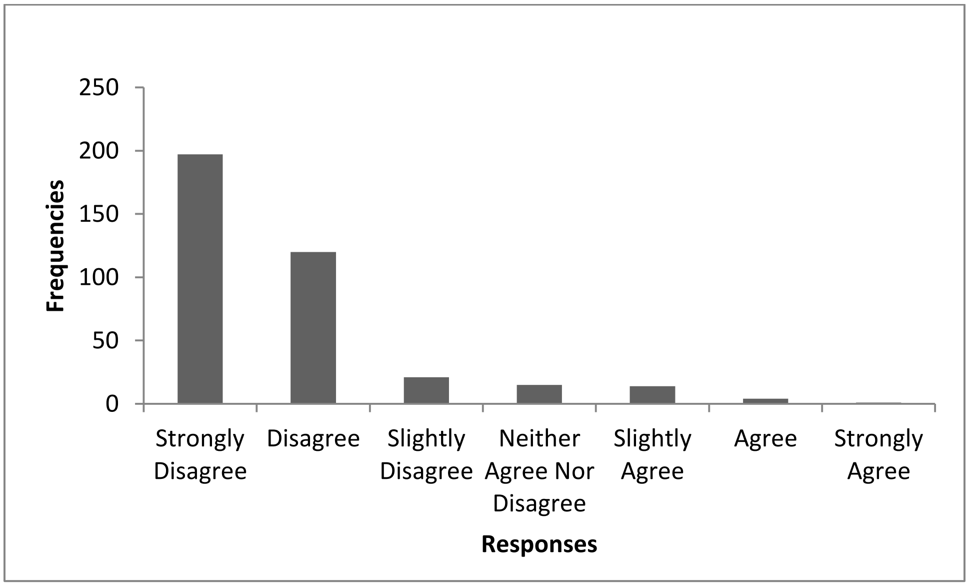


Figure 6k. *Item N2*

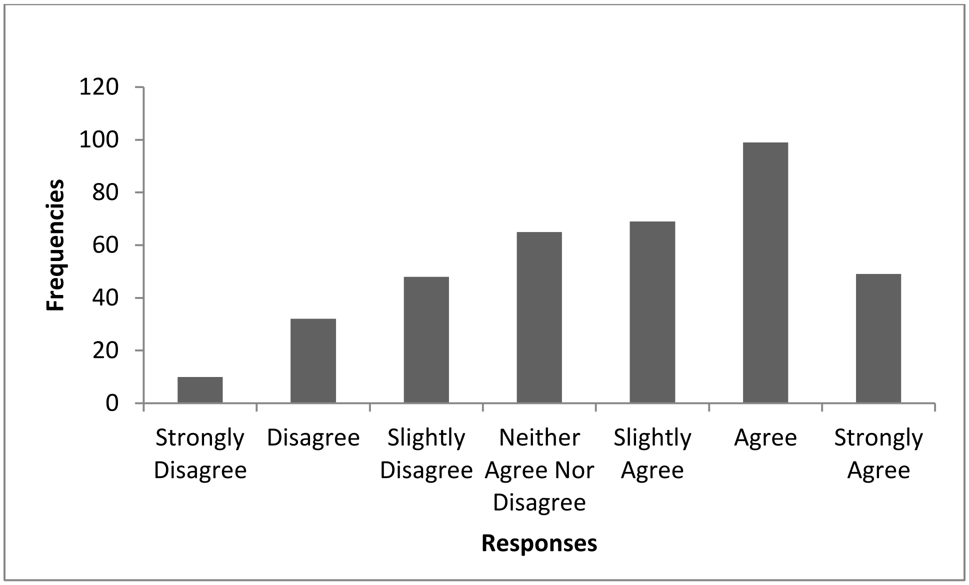
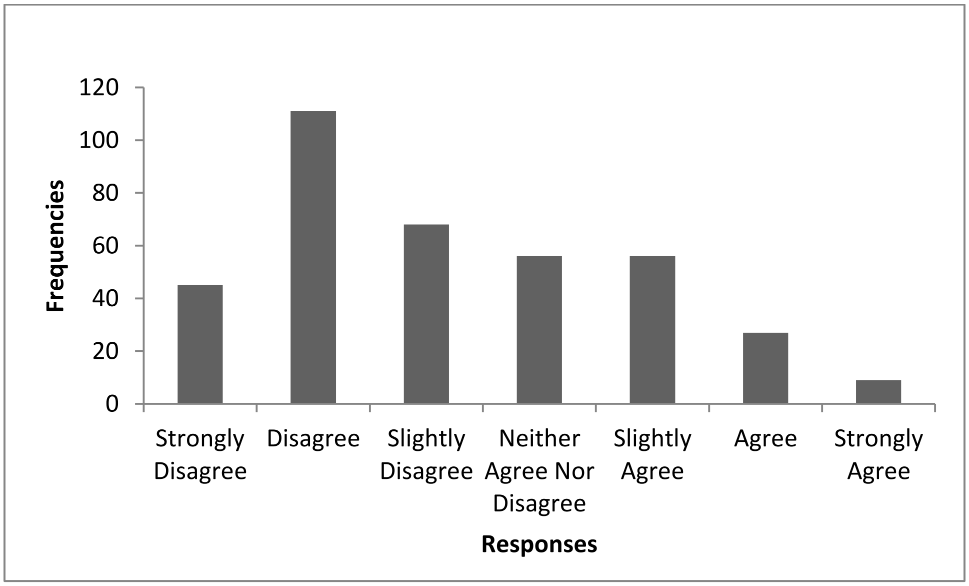


Figure 6l. *Item N3*



References

Appleton, J., Christenson, S., Kim, D., & Reschly, A. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology, 44,*427-445.

Barry, C. & Finney, S. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment, 4,* 17-26.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17,* 10-17, 22.

Bentler, P. M., & Wu, E. J. C. (2003). *EQS for Windows User's Guide*. Encino, CA: Multivariate Software, Inc.

Barron, K., Hulleman, C., Lazowski, R., Flake, J., and Grays, M. (2012). What constructs matter in academic motivation: A mixed-method investigation. Poster presented at the annual conference of the American Educational Research Association.

Bonwell, C. & Eison, J. (nd). Active learning: Creating excitement in the classroom.

Brown, G. & Hirschfeld, G. (2008). Students' conceptions of assessments: Links to outcomes. *Assessment in Education: Principles, Policy, & Practice, 15*(1), 3-17.

Charsha,A., Smiley, W. & Anderson, R. (2012, October). *Measuring cognitive engagement in low-stakes testing: Confirmatory factor analyses of the Cognitive Engagement Scale-2*. Paper presented at the annual conference of the Northeastern Educational Research Association, Rocky Hill, CT.

Corno, L. (1993). The best-laid plans: Modern conceptions of volition and educational research. *Educational Researcher, 22,* 14–22.

DeMars,C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77.

Flake, J., Barron, K. E., Hulleman, C. S., Grays, M., Lazowski, R., & Fessler, D. (2011, May). Evaluating cost: The forgotten component of expectancy value theory. Poster presented at the 2011 Association for Psychological Sciences Annual Convention, Washington, DC.

France, M. & Finney, S. (2009). What matters in the measurement of mattering? A construct validity study. *Measurement & Evaluation in Counseling and Development*, *42*, 104 – 120.

Fredricks, J. Blumenfeld, P. & Paris, A. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59-109.

Greene, B. & Miller, R. (1996). Influences on achievement: Goals, perceived ability, and cognitive engagement. *Contemporary Educational Psychology*, *21*, 181-192.

Greene, B., Miller, R., Crowson, H., Duke, B., & Akey, K. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology*, *29*, 462-482.

Helme, S. & Clarke, D. (2001). Identifying cognitive engagement in the mathematics classroom. *Mathematics Education Research Journal*, *13*(2), 133-153.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Kopp, J., Zinn, T., Finney, J., & Jurich, D. (2011) The development and evaluation of the academic entitlement questionnaire. *Measurement and Evaluation in Counseling and Development, 44*(2),105-129.

Lau, A., Swerdzewski, P., Jones, A., Anderson, R., & Markle , R. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*(3), 196-217.

Lee, V. E., & Smith, J. B. (1995). Effects of high school restructuring and size on early gains in achievement and engagement. *Sociology of Education, 68,* 241–270.

Lublin, J. (2003). Deep, surface, and strategic approaches to learning. Centre for Teaching and Learning, UCD Dublin.

MacCallum, R. Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490-504.

Marks, H. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153-184.

Meece, J., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientation and cognitive engagement in classroom activities. *Journal of Educational Psychology, 80,* 514–523.

Napoli, A. & Raymond, L. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in graded and un-graded conditions. *Research in Higher Education, 45*(8), 921-929.

Nystrand, M., & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, *25,* 261–290.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, & WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*, 557-595.

Pintrich, P. & de Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Education Psychology, 82*(1), 33-40.

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases and related problems. *Sociological Methods and Research, 13,* 109-119.

Smiley, W. & Anderson, R. (2011). Measuring students' cognitive engagement on assessment tests: A confirmatory factor analysis of the short form of the Cognitive Engagement Scale. *Research and Practice in Assessment*, *6,* 17-28.

Skinner, A. & Belmont, M. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*(4), 571-581.

Stipek, D. (2002). Good instruction is motivating. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation.* San Diego, CA: Academic Press.

Sundre, D. (2008). Student Opinion Scale (SOS): A measure of examinee motivation. Test Manual Center for Assessment and Research Studies, James Madison University.

Sundre, D, Erb, P. & Russell, J. (2009). Motivation in low stakes testing conditions: What's the feedback on feedback? Presented at the annual conference of the Northeastern Educational Research Association, Rocky Hill, CT.

Thelk, A., Sundre, D., Horst, J., & Finney, S. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performances. *The Journal of General Education*, *58*(3), 129-151.

Velez, P., & Ashworth, S. (2007). The impact of item readability on the endorsement of the midpoint response on surveys. *Survey Research Methods, 1*(2), 69-74.

Wehlage, G., Rutter, R., Smith, G. Lesko, N., & Fernandez, R. (1989). Reducing the risk: Schools as communities of support. Philadelphia: Farmer Press.

Weinstein, C., & Mayer, R. (1986). The teaching of learning strategies. In M. C. Wittrock (Ed.), *Handbook of research on teaching and learning* (3rd ed., pp. 315–327). New York: Macmillan.

West, S.G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: concepts, issues, and applications*, 57 – 75.

Wise, S. & DeMars, C. (2005): Low examinee effort in low-stakes assessment: Problems and potential solutions, *Educational Assessment*, *10*(1), 1-17.

Wigfield, A, & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68-81.

Young, W., Finney, S. J., & Bacon, J. (2007, June). Mentoring as a judicial sanction: Assessing sense of belonging. Paper presented at the International Assessment & Retention Conference, St. Louis.

Yu, C., & Muthén, B (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Zilberberg, A., Anderson, R., Finney, S., & Marsh, K. (under review, *Educational Assessment).* American college students' attitudes toward institutional accountability testing: Developing measures.

Zilberberg, A., Anderson, R., Swerdzewski, P., Finney, S. & Marsh, K. (2012). Growing up with no child left behind: An initial assessment of the understanding of college students' knowledge of accountability testing. *Research & Practice in Assessment, 7,* 12-25.

Zilberberg, A., Finney, S., Marsh, K. & Anderson, R. (in progress). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 21,* 3–17.

Zimmerman, B. & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology, 80(*3), 284-290.