Title
Studying the validity concept using a unified framework: A case study of a community college student engagement scale

Author
Quinones, Patricia

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Studying the validity concept using a unified framework:

A case study of a community college student engagement scale

A dissertation submitted in partial satisfaction

of the requirement for the degree of Doctor of Philosophy

in Education

by

Patricia Quinones

2015

ABSTRACT OF THE DISSERTATION

Studying the validity concept using a unified framework:

A case study of a community college student engagement scale

by

Patricia Quinones

Doctor of Philosophy in Education

University of California, Los Angeles, 2015

Professor Christina A. Christie, Chair

The use of student survey in higher education is vast and gathered for a variety of purposes, including accountability, assessment of learning outcomes, and increasingly, to measure student engagement (Pike, 2012; Porter, 2011). The crucial role student feedback plays in higher education begs the question, just how valid are college student surveys? More precisely, are they being validated in a way that is consistent with current validity theory and with the criteria agreed upon and put forth by professional organizations? Recent efforts to validate instruments appear to be minimal and reflect a lack of understanding of current validity theory. Whereas theorists tend to view validity as consisting of one general form of construct validity, practitioners continue to perceive different types of validity. The result is a gap between validity theory and how validity is assessed in practice.

For this research, the validity of the newly developed Community College Student Engagement Scale (CCSES) was assessed. The concept of student engagement has garnered much attention in recent years and has been linked to student retention, motivation, and academic achievement (Fredricks, Blumfield, & Paris, 2004). Given the effects attributed to student engagement on these important educational issues, the need for a reliable and accurate measure of it is great, especially at the community college level, where there is a lack of available student engagement measures.

Using Messick's (1988, 1989, 1995) unified framework of construct validity, this research aimed to accomplish two things: (a) assess the validity of the student engagement scores in a way that is consistent with established guidelines and theory, and (b) document the barriers, if any, to assessing validity in a concerted way that could potentially speak to the existing gap between validity theory and practice.

A mixed-methods study was used to assess the content, structural, and external facets of construct validity. The sample consisted of community college instructors and students. Data collected from instructor interviews, student focus groups, and classroom observations indicate that there is sufficient evidence of the content facet of construct validity for the CCSES. Results also indicate that, structurally, the CCSES corresponds with the multi-dimensional nature of the student engagement construct, as defined by the engagement literature. Lastly, external results show convergence of student self-reported, instructor-reported, and researcher-reported engagement data. External results show that instructors' ratings of their students' class participation engagement was the only significant predictor of English GPA and overall fall 2014 GPA.

Future validity studies should move away from the fragmented and outdated framework that they continually rely on, one that states that there are three separate types of validity. Continued attempts to assess validity using a unified framework can potentially improve validity theory and practice, because they allow for the sharing of experiences, strategies, and lessons learned.

The dissertation of Patricia Quinones is approved.

Todd M. Franke

Noreen M. Webb

Marvin C. Alkin

Christina A. Christie, Committee Chair

University of California, Los Angeles

2015

To my mom and dad, Lourdes and Juan Quinones, for their unconditional love and support, and for demonstrating strength and resilience; to my siblings, closest friends, and extended family for your encouragement, understanding, and providing countless moments of laughter.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

The successful completion of my graduate program was due, in part, to the support and guidance of many wonder advisors and mentors throughout the years. Most recently, Tina Christie, my advisor for the past five years has been a fierce advocate, providing me with opportunities that have allowed me to grow personally and professionally. Her unwavering trust in my abilities provided the courage to step out of my comfort zone and try new endeavors. I cannot thank her enough for providing the opportunity to study at UCLA, and for providing the opportunity to collaborate with the amazing faculty and students of the Social Research Methodology (SRM) division in the Department of Education. The impact she has had in my life the past five years can never be fully stated.

It has been a privilege to work with, engage in wonderful conversations, and have the unconditional support of my dissertation committee members: Todd Franke, Reenie Webb, and Marv Alkin. I thank them tremendously for being part of my dissertation committee and for helping shape this study. My fellow SRM friends provided an amazing supportive community. To my best friend, Debbie Grodzicki, thank you for so many unforgettable moments, for so many moments of laughter that made the past four years seem to pass too quickly. I am also grateful to have met other SRM students: Jenn, Lisa, Tim, Ale, and Minh. They are incredibly smart, and I have learned so much from each and every one of them. I look forward to continued friendship.

Incredible thanks is owed to the STACC faculty at Pasadena City College, especially Elsie B. Rivas Gomez, Kirsten Ogden, Moremi Ogbara, and Carolynn M. Rosales. Their willingness to host this study, and willingness take time from their incredibly busy schedules to provide feedback on various aspects of this study is humbling.

The transition from master's to Ph.D. was seamless because of the wonderful mentorship of my master's advisor, Jodie Ullman. I can't thank her enough for her rigor, her standard of excellence, and for teaching me how to be an independent learner. She mentored me with an eye towards the future, believing from the very beginning that I would continue on to a Ph.D. program. Without her encouragement, I probably would not have continued my graduate career.

Though I have had much support during my graduate career, the support I received as a lost and unsure community college student cannot go unnoticed. A special thanks to Rosa Preciado, one of the first individuals that encourage me think; she reassured me that my pursuit of a higher education would be worthwhile. Michelle G. Hillman was also an alley and provided opportunities at Mt. SAC that would allow me to discover my passion for research.

Finally, none of this would have been possible without the unconditional love and support of my parents. They never questioned my choices and allowed me to go where I needed to go and do what I needed to do to make this possible. My accomplishments are their accomplishments and they deserve to be recognized and congratulated just as much as I have.

VITA

EDUCATION

2009 M.A. Psychology
California State University, San Bernardino, CA

2004 B.A. Psychology
California State University, Fullerton, CA

PROFESSIONAL EXPERIENCE

2014 – 2015 Teaching Assistant
Graduate School of Education and Information Studies
UCLA

2012 – 2014 Graduate Student Researcher – Data Analyst for MHSA Evaluation
Graduate School of Education and Information Studies
UCLA

2010 – 2015 Evaluator/Researcher
Pasadena City College

2010 – 2014 Consultant - Data Analyst
Various projects
Los Angeles, CA

2008 – 2009 Research assistant
Department of Psychology
Cal State San Bernardino

2008 - 2008 Lecturer – Psychology 101
Mt. San Antonio College
Walnut, CA

2007 – 2009 Teaching Assistant
Department of Psychology
Cal State San Bernardino

SELECTED PRESENTATIONS

2015 Quinones, P. Validation study of the community college student engagement scale. Paper was presented at the annual Research and Inquiry conference of the Department of Graduate and Information Studies, UCLA.

2013 Ho, T. Quinones, P., Priede, A. Strategies for improving the quality of quantitative data. Paper was presented at the annual meeting of the American Evaluation Association, Washington, D.C.

2011 Quiñones, P., Vo, A. Using structural equation modeling to test program theory. Paper was presented at the annual meeting of the American Evaluation Association, Anaheim, CA.

2008 Quiñones, P., Thomas, E., Ullman, J. B., (May 2008). The assumptions underlying statistical analyses – Is anyone checking? Poster was presented at the annual meeting of the Association for Psychological Science, Chicago, IL.

2008 Thomas, E., Quiñones, P., Ullman, J. B. (April 2008). What was the research question? A review of the literature. Poster was presented at the annual meeting of the Western Psychological Association, Irvine, CA.

SELECTED HONORS AND AWARDS

2015 Outstanding Presentation Award for Paper titled ""Validation Study of the Community College Engagement Scale" at the Research and Inquiry Conference, UCLA.

2014 Dissertation Fellowship Award

2014 Mary Jenson Fellowship Award

2013 Graduate Summer Research Mentorship

2010 – 2014 Eugene V. Cota-Robles Fellowship

CHAPTER 1

BACKGROUND AND STATEMENT OF THE PROBLEM

The use of student survey data in higher education is vast. In 2009, the National Institute for Learning Outcomes Assessment (NILOA) surveyed administrators at all institutions of higher education in the United States and found that 76% collected and used student survey data (Kuh & Ikenberry, 2009). These data are gathered for a variety of purposes, including accountability, assessment of learning outcomes, and, increasingly, to measure student engagement (Pike, 2012; Porter, 2011). The crucial role student feedback plays in higher education begs the question, just how valid are college student surveys? More precisely, are they being validated in a way that is consistent with current validity theory and with the criteria agreed upon and put forth by professional organizations?

A little over 50 years ago, Ebel (1961) stated that validity is "one of the major deities in the pantheon of the psychometrician" (p. 640). This sentiment still holds true today. According to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999), validity is "the most fundamental consideration in developing and evaluating tests." (p. 9). The tenets of validity extend to all performance assessments, tests, surveys of any and all phenomena, and observational guides, and this is why it is such an important issue.

Ebel (1961) also asserted that despite the importance of validity, there were remarkably few examples in which it had been properly assessed. Unfortunately, this also still seems to be the case. Recent efforts to validate instruments appear to be minimal and, where they are

conducted, the approach typically reflects a lack of understanding of current validity theory. Whereas theorists tend to view validity as consisting of one general form of construct validity, practitioners continue to perceive different types of validity. The result is a gap between validity theory and how validity is assessed in practice.

For this research, I assessed the validity of the newly developed Community College Student Engagement Scale that measures engagement in four different contexts: class participation, relationships with faculty, relationships with peers, and participation in campus activities. Although I assessed the entire scale, I focused particularly on class participation by collecting student, faculty, and observer ratings of this domain. The concept of student engagement has garnered much attention in recent years and has been linked to student retention, motivation, and academic achievement (Fredricks, Blumfield, & Paris, 2004). Given the effects attributed to student engagement on these important educational issues, the need for a reliable and accurate measure of it is great.

Using Messick's (1988, 1989, 1995) unified framework of construct validity, which has largely been adopted by the joint committee that developed the widely used *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), this research aimed to accomplish two things: (a) study validity as a unified framework and document the barriers, if any, to assessing validity in a concerted way that could potentially speak to the existing gap between validity theory and validity practice, and (b) assess the validity of student engagement scores in a way that is consistent with established guidelines and theory.

When conclusions and recommendations are made based on assessment scores, lackluster validity efforts can have serious implications. On the other hand, properly conducted validation studies—i.e., those that take into account current validity theory and standards—can advance the

field of measurement by demonstrating what validity principles look like in practice, documenting obstacles and gaps, and providing examples of sound methods that practitioners can reference. Hence, concerted validity efforts have the potential to advance both validity theory and practice.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a review of the literature in three areas. I first address how validity theory has evolved from including three distinct "types" of validity to the current view that there is just one general type, commonly referred to as the unified concept of construct validity. Questions regarding what it means for a test to be valid and what evidence is necessary to establish validity are also addressed. Second, I describe how validity methods have been used in published studies across many disciplines. Lastly, I discuss the importance of the student engagement construct, how it has been operationalized, and the methods used to assess the validity of specific student engagement measures.

Evolution of Validity Theory

According to Messick (1995), "validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p. 447). This somewhat recent definition is based on the theory that validity consists of one general form, construct validity, as it indicates just one overall evaluative judgment, rather than many. This singular view evolved from an earlier notion that validity consists of three independent types: criterion, content, and construct validities.

Criterion validity refers to how well test scores correlate with real world criteria (e.g., how well a job placement test correlates with actual job tasks), and predictive and concurrent validities are extensions of this broader concept. Typically, criterion validity relies on

correlational analyses, and is used to validate selection and placement tests (Kane, 2001). *Content validity* assesses the degree to which a test covers and represents a domain of interest. This type of validity typically relies on expert judgment and has been used to justify the use of various achievement tests (Kane, 2001). *Construct validity* is typically defined as the extent to which a test or assessment accurately measures the construct of interest. Assessment of construct validity is not as straightforward as assessment of criterion or content validity.

Six decades ago, Cronbach and Meehl (1955) stated that construct validation "takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are means of confirming or disconfirming the claim" (p. 290). As can be noted with this definition, the process of assessing construct validity requires in-depth analysis. To address this extensive process, in their seminal piece, *Construct Validity in Psychological Tests,* Cronbach and Meehl (1955) laid out a framework, called a nomological net, for assessing the strength of a construct. A nomological net is essentially a representation of related constructs. Within a nomological net is a nomological network that defines a construct by illustrating the interrelationships between the constructs in the net and observable or behavioral indicators implicit to a theory. Thus, the network generates testable hypotheses regarding how test scores relate to other constructs and variables.

Although this framework was theoretically useful, Cronbach and Meehl (1955) did not provide an explicit methodology for testing a nomological network. In 1959, however, Campbell and Fiske developed a specific methodological framework that could assess construct validity by testing the interrelationships between constructs or traits, called the multitrait-multimethod matrix. They also introduced two new types of validity, *convergent* and *discriminant*, which

could be assessed with the multitrait-multimethod matrix. Within this framework, to claim that a test had construct validity, both convergent and discriminant validity had to be established.

A multitrait-multimethod matrix (MTMM) is essentially a table or matrix of correlations between theoretically similar and dissimilar constructs or traits that have been measured in different ways. It provides evidence for how a measure compares to other measures. To demonstrate convergent validity, measurement of the same construct/trait conducted in different ways (e.g., survey and observation) should correlate highly. Conversely, theoretically dissimilar constructs/traits measured the same way (e.g., both via a survey) should not correlate highly, thus providing evidence of discriminant validity. Correlation coefficients also provide estimates of trait variance and method variance.

Table 2.1 depicts a typical MTMM matrix in which three traits have been measured using three different methods. Differences between methods can be instrument-based (e.g., different scales), occasion-based (e.g., time 1, time 2, time 3, etc.), or information-based (e.g., student, teacher, parent). The diagonals, shaded in black, are reliability estimates because they correlate the same trait using the same method; this is called "monotrait-monomethod." For example, suppose that Trait A_1 is depression that is being measured by a self-reported scale - Method 1. The uppermost black cell represents the correlation of the self-reported depression scale with itself.

The validity diagonals are correlations between different measures of the same trait. Returning to the example above, Trait A_1 is depression that is being measured by a self-reported scale, Method 1. Suppose that Trait A_2 is also depression that is measured by a clinician using a behavioral depression scale (Method 2). High correlations between the self-reported depression scores and the clinician reported depression scores provide evidence of *convergent* validity,

because different data sources concerning the same trait are converging. Measuring the same trait using different methods is called “monotrait-heteromethod.”

The blue cells are correlations between theoretically different traits that have been measured in the same way. For example, Trait A_1 – Method 1 is a self-reported depression scale. Suppose Trait B_1 – Method 1 is a self-reported sports viewing scale. High correlations between self-reported depression scores and self-reported sports viewing scores indicate method variance—i.e., these different traits are correlated because they are measured in the same way. Measuring different traits using the same method is called “heterotrait-monomethod.” The beige cells indicate correlations between theoretically different traits that are measured using different methods. For example, the beige A_2 – B_1 cell is a correlation between therapist-reported depression scores and self-reported sports viewing scores. The correlations in the beige cells are expected to be the lowest in the matrix because they share neither trait nor method variance. Measuring different constructs or traits using different methods is called “heterotrait-heteromethod.”

Table 1. *Example of an MTMM Matrix*

		Method 1			**Method 2**			**Method 3**		
	Traits	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1									
	B_1									
	C_1									
Method 2	A_2	Validity								
	B_2		Validity							
	C_2			Validity						
Method 3	A_3	Validity			Validity					
	B_3		Validity			Validity				
	C_3			Validity			Validity			

The view that validity consists of three independent types—criterion, content, and construct—persisted until the late 1970s (Kane, 2001). As early as 1957, however, Loevinger argued that content, concurrent, and predictive validities were *ad hoc*, and that these types of validity were probably supporting evidence for construct validity. Loevinger further argued that the different types of validities were not equal, and that construct validity was more important because it provided a scientific basis for assessing validity. And Ebel (1961) had also noted that there was something not quite right with how validity was conceptualized—specifically, the lack of a common definition. For example, Lindquist (1942) defined validity as "the accuracy with which [a test] measures that which it is intended to measure, or as the degree to which it approaches infallibility in measuring what it purports to measure" (p. 213); Edgerton (1949) stated, "By 'validity' we refer to the extent to which the measuring device is useful for a given purpose" (p. 52); and Gulliksen (1950) and Cureton (1951) defined validity as the extent to which test scores correlated with a true criterion. As Ebel (1961) rightly noted, using the same term for different concepts was not only confusing but could also lead to difficulties in assessing validity.

In short, understandings of validity were fragmented and incomplete. Previously, if a test was found to have criterion evidence, the test was deemed "valid". However, this view of validity was incomplete because it is not enough to correlate a test score with a criterion and claim that a test is valid, with no consideration of the content or the construct of the test. What good would it do to demonstrate criterion validity if the test itself did not include a representative sample of the domain of interest? As a consequence of these issues, the field began to gradually move away from conceptualizing validity as comprising different types, and towards a *unified concept of construct validity*. The shift to a unified framework occurred, in part, because it was

not enough to assess for only one type of validity without consideration of or attempting to gather evidence for other types of validity. The shift was not in the way criterion, content, construct, and other types of validity were being defined or assessed, rather the shift was about bringing together the different types of validity into one unified framework. Under this framework, these different types of validity are now evidences for construct validity, which was elevated as the unifying concept because of the potential for testing hypotheses.

As noted earlier, assessment of construct validity requires a more in-depth analysis of inter-correlations between theoretically-relevant constructs and other variables—an analysis in which hypotheses can be explicitly stated and empirically tested. This was one of the reasons that construct validity was elevated as the unifying concept that would subsume all other types of validity. As Loevning (1957) noted, content, predictive, and concurrent validities were all evidence of construct validity. Kane (2001) stated that, "taking construct validity as the unifying principle for validity puts validation squarely in the long scientific tradition of stating a proposed interpretation (or theory) clearly and subjecting it to empirical and conceptual challenge" (p. 325). Thus, the inception of the unified framework started with the work of Cronbach and Meehl (1955) and Loevinger (1957), and it gathered steam over the course of two decades. As Moss (2007) indicates, Cronbach "gave construct validity far more centrality in his general concept of validity" (p. 472) when he wrote the validity for the Educational Measurement book, a publication sponsored by NCME and the American Council on Education (ACE) in 1971. In 1989, Samuel Messick wrote the proceeding validity chapter in Educational Measurement, and solidified the unified concept of construct validity as the dominant framework. Both Cronbach and Messick addressed another important validity issue, the view that validity is a property of a test.

Validity of Test Scores and Interpretations

Previously, validity was thought to be a property of a test—a test was valid if either content, criterion, or construct validity was established. However, as the notion of validity as being comprised of one general kind of construct validity evolved, so too did the view that validity was a property of a test. In the early 1970s, Cronbach (1971) stated:

> The phrase validation of a test is a source of much misunderstanding. One validates, not a test, but an interpretation of data arising from a specified procedure. A single instrument is used in many different ways—Smith's reading test may be used to screen applicants for professional training, to plan remedial instruction in reading, to measure the effectiveness of an instructional program, etc. Since each application is based on a different interpretation, the evidence that justifies one application may have little relevance to the next. Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test is "valid" (p. 447).

More than 20 years later, Messick (1989) reiterated this stance by stating that what needs to be validated are the *interpretations* attached to test scores, because test scores are not only a function of test items, but also a function of the individuals taking the test, the context in which the test is being administered, and the specified purpose of the test. Therefore, validity is a continual process and should be assessed whenever a test is used. So what does it mean for a test to be valid? It means there is enough theoretical and empirical evidence for the interpretation and use of the test scores.

Messick (1988) identified six facets that can be used to assess construct validity: content, substantive, structural, generalizability, external, and consequential. The *content* facet requires identifying the boundaries of the construct that is being studied, including the skills, behavior,

and knowledge that should be represented, given the boundaries of that construct. The *substantive* facet refers to the collection of empirical evidence in which respondents actually engage during sampled processes or tasks; it deals with *how* people answer test items. The *structural* facet refers to reliability and an acceptable factor structure of a test (i.e., if it matches what one would expect, given the theory underlying the construct in question). The *generalizability* facet refers to the extent to which test properties are generalizable to different groups, populations, or settings. The *external* facet of construct validity assesses the relationships and predictability of test scores to other measures or behaviors that are implicit to the theory of the construct being measured. And the *consequential* facet deals with potential consequences of test use, including sources of invalidity, and issues such as fairness and test biases.

These six facets of construct validity were largely adopted by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, and NCME) in the jointly-sponsored *Standards for Educational and Psychological Testing* (1999). Commonly referred to as the *Standards* or *Testing Standards*, this document provides guidelines for researchers and test developers about "the criteria for the evaluation of tests, testing practices, and the effects of test use" (p. 2). The *Standards* also provides consensus regarding best practices. In particular, consistent with Messick's (1988) unified framework and criteria for assessing construct validity, the *Standards* identifies five categories that researchers and test developers can use as evidence to establish construct validity: test content, response processes, internal structure, relation to other variables, and consequences of testing.

To summarize, validity is no longer viewed as consisting of three independent types, but is now seen as the unified form of construct validity. Criterion, content, predictive and other

forms of validity are now viewed as evidences for construct validity. Messick and the *Standards* have identified different sources of evidence for the assessment of construct validity. But what do these validity principles look like in practice? Are practitioners and researchers assessing validity in ways that reflect current validity theory and standards? I turn to these questions next.

Validity Studies

The question of whether practice reflects theory when it comes to validity was explored by Cizek, Rosenberg, and Koons (2008). These scholars assessed the extent to which validity reports conformed to modern validity theory, what specific sources of evidence were used to assess validity, and whether these sources of validity differed depending on the type of measure. They established a test information database by using the *Mental Measurements Yearbook* (*MMY*), which provides a guide to hundreds of testing instruments across many disciplines, including psychology, education, business, and leadership. Each *MMY* entry contains descriptive information regarding the test, including its purpose and external reviews provided by content experts. To be included in the *MMY*, technical test information and data regarding reliability and validity must be submitted for review by *MMY* staff. Staff assess the quality of the data and provide this information in the *MMY* entry.

Cizek and colleagues reviewed approximately 283 *MMY* entries, and found that only 2.5% (7 tests) used language that conveyed a unitary concept of validity. The most common perspective was that validity comprises different types. It is not surprising, then, that only 9.5% (27) of the entries made reference to Messick's unified concept of construct validity or to the *Standards*. Approximately 30% of the entries referred to validity as being a property of a test, whereas 25% referred to validity as a characteristic of a test score or interpretation. The most common sources of validity evidence reported were content (48.5%), construct (58%), and

concurrent (50.9%). The authors contended that researchers and practitioners have rejected or ignored current validity theory. Given that assessments included in the *MMY* are arguably of higher quality because of rigorous review by experts and *MMY* staff, the results of this study are even grimmer. What can be expected of assessments not published in the *MMY*?

Hogan and Agnello (2004) explored a sample of articles in the *Directory of Unpublished Experimental Mental Measures*, a publication of the American Psychological Association (APA), and one of the most widely cited sources of psychological assessments. They assessed validity and reliability reporting practices, and also reported on specific methodologies used to assess validity. Of the 696 entries they analyzed, 45.4% (316) included no validity information. Approximately 54.6% included at least one type of validity evidence. Only 2.3% included two types of validity evidence, and no entries included more than two types. Of the 54.6% that reported some type of validity evidence, 67% relied on correlations with other variables.

Barry, Chaney, Piazza-Gardner, and Chavarria (2013) reviewed 967 articles published in health education and behavioral journals to determine the frequency with which articles that used scales or subscales reported reliability and validity evidence. Of the 967 articles, 31% reported having measured validity for the surveys used to collect data, but only 26% actually provided validity statistics or methods for assessing validity. It could be that validity for the scales and subscales used in these articles was assessed but not reported, and this may have contributed to the low percentages of validity assessment. It is also possible that the journals selected for this study did not focus on assessments; articles published in assessment journals would likely fare better in this type of analysis.

Slaney, Tkatchouk, Gabriel, and Maraun (2009) explored reporting practices concerning reliability and validity in four assessment focused journals: *Educational and Psychological*

Measurement; *Psychological Assessment*; *Journal of Personality and Individual Differences*; and *Journal of Personality Assessment*. The authors also sought to determine whether researchers were employing psychometrically sound practices. Results from this study were much more encouraging with respect to reporting practices: Approximately 92% of the 368 articles reviewed actually provided some kind of validity evidence—primarily correlational evidence. When the authors assessed the extent to which researchers employed sound methods for assessing validity, however, results were less positive.

Slaney and colleagues (2009) defined psychometrically sound practices as assessment of the following evidences in the following order: (a) internal score validity, which assesses the extent to which items correlate with each other, as indicated by the theory of a construct being measured (i.e., the internal structure of a measure); (b) score precision/reliability, which refers to the reliability of a measure if it is unidimensional, or to the reliability of subscale scores if it is multidimensional; and (c) external test validity, which assesses the extent to which scale or subscale scores predict some kind of outcome, as dictated by the theory of the construct being measured or the extent to which scale or subscale scores correlate with other test scores.

Results showed that only 32% of the articles assessed all three evidences in the specified order; 53% of the articles assessed reliability first and then the factor structure of the measure. These results are problematic because it appears that researchers are not cognizant that it is not useful to assess the reliability of a score if that score is made up of items that measure different qualities or distinctions. The authors also noted that very few researchers identified the theoretical factor structure of a measure and then assessed internal score validity to see if the underlying factor structure matched the theory. This result implies that perhaps measures are not being created or are not grounded in the theory relevant to the construct that is being measured.

If that is the case, then what kind of information are researchers using to guide the development of a measure?

Collectively, these studies reveal three things. First, assessment of validity for scales or subscales used in published articles is very rare. It can be argued that validity is being assessed but not reported due to journal editing policies, although this argument does not seem likely when, for any given journal, there was at least one article that reported validity evidence. If there were a lack of validity reporting due to editing policies, it would be expected that no articles would have provided validity evidence for a given journal. It can also be argued that researchers assessed validity but simply did not report it. If one makes the effort to assess the validity of a measure, however, it seems likely that this would be reported, especially when validity evidence could be used to bolster the rationale for using a particular measure. Second, even when validity evidence is provided for published articles, sources are limited, even in articles published in measurement-focused journals. And third, the way validity has been assessed—as though it comprises different types—is not in line with current validity theory.

The studies described above assessed validity practices across many different disciplines, but not in the context of student engagement research. Before I describe the proposed study on the validity of student engagement scales, a rationale for focusing on student engagement scales is warranted.

Complexity of Student Engagement

The concept of student engagement has received increased attention as a possible means of reducing student dropout rates, increasing motivation, and raising overall academic achievement levels (Fredricks, Blumfield, & Paris, 2004). Generally, student engagement can be defined as a student's involvement in educational activities, such as attending class, completing

coursework, and participating in extracurricular activities. High levels of student engagement have been linked to higher rates of retention and higher levels of academic achievement (Kuh, Cruce, Kinzie, & Gonyea, 2008; Laird, Chen, & Kuh, 2008; Reyes, Brackett, Rivers, White, & Salovery, 2012). Given the strong effects attributed to student engagement on a range of educational issues, the need for a scale to accurately measure the construct is great. But the task of constructing a reliable and accurate student engagement scale has proven to be a difficult task, due at least in part to the complexity of the construct.

The complexity of student engagement is noted in the various student- and institutional-level domains used to measure it. Fredricks, Blumenfeld, and Paris (2004) conducted a review of the student engagement literature and indicated that student engagement consists of behavioral, affective, and cognitive domains. *Behavioral engagement* is typically defined as participation either in academic or extracurricular activities. According to the authors, it is a critical component of academic outcomes and maintaining enrollment (i.e., retention). *Affective engagement* typically consists of individuals' positive and negative emotions regarding school, teachers, and peers. It has been theorized that this is a critical component in creating ties to an academic institution (Fredricks et al., 2004). And *cognitive engagement* is typically defined as individuals' persistence, beliefs, and self-perceptions regarding learning, as well as planning, investing and self-regulating. Fredricks et al. also noted that school-level factors—e.g., teacher support, relationships with peers, classroom structure, autonomy support, and task characteristics—have been found to be associated with behavioral, affective, and cognitive engagement.

Jimerson, Campus, and Greif (2003) classified measures of student engagement into five contexts, based on a review of 45 studies. The first context, *academic performance*, consists of

items relating to grades, achievement tests, hours studying, and completion of assignments. The second, *classroom behaviors*, includes items related to asking questions, attending class, and general classroom behavior. *Extracurricular activity* consists of items related to frequency of participation in sports or other school activities. The fourth context is *interpersonal relationships*, which consists of items relating to relations with peers and teachers. And the last context is *school community*, consisting of items related to feelings and attitudes toward the school.

The complexity of the student engagement construct is not only seen in the variety of proposed components but also in the lack of consensus about a definition for the construct. Appleton, Christenson, and Furlong's (2008) review of the engagement literature identified 19 definitional variations of student engagement. Jimerson et al. (2003) also noted that terms related to student engagement have been used interchangeably, including school engagement, belonging, school community, affiliation, school membership, and motivation. Thus, the lack of consensus regarding a definition of student engagement may be related to the lack of common terminology. Nevertheless, despite conceptual and definitional differences concerning student engagement, there is strong empirical evidence to connect the concept in broad terms with academic achievement and drop-out rates at both the high school and college level, and these findings are consistent across ethnic groups (Janosz, Archambault, Morizot, & Pagoni, 2008; Kuh, Cruce, Kinzie, & Gonyea, 2008; Laird, Chen, & Kuh, 2008; Ream & Rumberger, 2008; South, Haynie, & Bose, 2007).

Student Engagement Measurement Scales

Paris et al. (2004) noted that engagement has been studied using scales, which may measure a single domain (behavioral, affect, or cognitive) or a combination of two or more domains. They noted, however, that one problem with scales containing multiple domains is that "most of the self-report measures of behavioral, emotional and cognitive engagement do not

specify subject areas." Moreover, "measures are rarely attached to specific tasks and situations, instead yielding information about engagement as a general tendency" (p. 69).

At least three college student engagement scales claim to measure engagement in multiple situations. The Community College Survey of Student Engagement (CCSSE) is administered at community colleges nationwide. This scale was developed to measure student and institutional activities that represent good educational practices. There is a debate regarding whether the CCSSE is a true measure of student engagement, however (Nora, Crisp, & Matthews, 2011). The survey consists of five benchmarks that are not grounded in theory but are empirically derived as clusters of student behavior and institutional practices. These benchmarks are: active and collaborative learning; student effort; academic challenge; student-faculty interaction; and support for learners. The number of items in the CCSSE varies because each campus has the option of adding a maximum of 15 custom items.

The National Survey of Student Engagement (NSSE) is the four-year equivalent to the CCSEE, and was developed by the same research group. The NSSE collects student information for five benchmarks: participation in educationally purposeful activities; institutional requirements and the challenging nature of coursework; perceptions of the college environment; estimates of educational and personal growth since starting college; and background and demographic information. The number of items on the NSSE varies, as each campus has the ability to customize the survey by adding topical modules, which are sets of short questions on topics such as academic advising, civic engagement, and experiences with diversity, technology, and writing.

The third engagement scale found in the literature is the Student Course Engagement Questionnaire (SCEQ), developed by Handelsman, Briggs, Sullivan, and Towler (2005). This

questionnaire consists of 23 items that comprise four factors: skills engagement (9 items), emotional engagement (5 items), participation/interaction engagement (6 items), and performance engagement (3 items). The purpose of the SCEQ is to measure engagement in specific lower division college courses. As the authors noted, "we primarily focus on the macro level—what happens in and immediately surrounding class" (Handelsman, Briggs, Sullivan, & Towler, 2006, p. 185). The authors conducted exploratory and confirmatory factor analyses, and the results indicated a four-factor model of student course engagement.

Although all three of these scales claim to measure engagement, it is unclear how, or if, the behavioral, affective, and cognitive domains, which reflect the multidimensional nature of the engagement construct, are represented in them. The CSSEE and NSSEE—although widely used by two- and four-year institutions, respectively—were not practical for the current study, given that they are not in the public domain and there is a fee to administer them, with cost varying based on the format of administration and the addition of customized topical modules or customized questions. The SCEQ was created for a specific lower-division classes and the goal was to create an engagement scale that measured engagement inside and outside of the classroom. Moreover, these existing engagement scales do not fully address the complexity of the engagement construct as described in the literature.

It was clear that a scale that was easily available and cost effective, and that encompassed behavioral, affective, and cognitive domains, as well as context/situational factors, was needed. Thus, I created a new engagement scale, named the Community College Student Engagement Scale (CCSES), based on the engagement literature rather than best instructional practices (CSSEE & NSSE) or the need to assess engagement only in the classroom (SCEQ).

The Community College Student Engagement Scale

The purpose of the CCSES is to measure engagement broadly. The CCSES is a multidimensional behavioral scale that measures engagement in four different contexts: class participation (8 items), relationship with peers (6 items), relationship with faculty (7 items), and participation in campus activities (4 items). For all items, across all four factors, students are asked to rate the frequency of their behaviors on a 5-point Likert-type scale, from 1 (never) to 5 (always). A description of the previous work done to develop the CCSES and its various iterations, including any changes made during this study is included in the Chapter 3 under the Measures section. A copy of the CCSES is included in Appendix A.

Validity of Engagement Scales

How has the validity of engagement scales been assessed? Not surprisingly, published validity studies have focused on the NSSE and CCSSE surveys. Collectively, the research has relied on the factor structure of these surveys as evidence of construct validity, and has pointed to the relationships between the NSSE and CCSSE benchmarks and academic outcomes as evidence of predictive validity.

Validity studies examining the factor structure of the CCSSE as evidence of construct validity have been mixed. One study found support for the five benchmark structure in the CCSSE (Mandarino & Mattern, 2010), while others have found support for factor structures quite different from the five benchmarks (Angell, 2009; Marti, 2004, 2008; Nora, Crisp, & Matthews, 2011). Studies assessing the predictive validity of the CCSSE have also been mixed. McClenney and Marti (2006) examined the relationship between the CCSSE benchmarks and academic outcomes—such as persistence, course completion, and credit hour accumulation—and found correlations. Nora, Crisp, and Matthews (2011) found that three of the five benchmarks

significantly predicted GPA, although in one there was a negative impact on GPA. For such a widely used instrument, studies assessing the validity of the CCSSE appear to be minimal. In fact, McClenney and Marti (2006) stated that student engagement is one of the more poorly studied concepts at the community college level.

Validation studies examining the factor structure of the NSSE as evidence of construct validity follow the same pattern as those for the CCSSE. Some show support for the five-benchmark structure (NSSE, 2010c, 2010e), while others show evidence of factor structure quite different from the five-benchmark structure of the NSSE (Esquivel, 2011; LaNasa, Cabrera, & Trangsrud, 2009). Small to moderate correlations between academic outcomes—including GPA, critical thinking, and grades—and NSSE benchmarks have been found (Carini, Kuh, & Klein, 2006; Gordon, Ludlum, & Hoey, 2008; LaNasa, Olson, & Alleman, 2007).

Although the results of CCSSE and NSSE validity studies are mixed at best, what is important to note is how validity is being conceptualized. It is evident that researchers conducting these studies still view validity as having different types, and that they use exploratory and confirmatory analysis to assess "construct" validity and correlational evidence as a way to assess "predictive" validity. This fragmented view and assessment of validity is consistent with how other measures across disciplines have been assessed.

To summarize, prominent validity theorists, as well as professional organizations, no longer view validity as possessing three distinct types, namely content, concurrent, and construct, but rather as one general form of construct validity. There appears to be a strong consensus that there is one general form of construct validity. Practitioners and researchers have, however, largely failed to take this view into account when conducting validation studies.

In applying the unified concept of construct validity framework developed by Messick (1988, 1989, 1994, 1995) and adopted by the *Standards*, the proposed study builds a validity argument for the CCSES based on theory and empirical evidence. According to Messick (1995), test validation cannot rely on just one supplementary form of validity, nor is there a single supplementary form of validity that is required for test validation. It is best to gather as many possible sources of evidence that collectively argue for the use and interpretation of scale/test scores.

Studying validity as a unified concept provided the opportunity to address the gap between validity theory and practice. Specifically, with these issues in mind, the study addressed the following research questions:

1. How does studying validity as a unified framework possibly affect future validity studies?
2. To what extent does the CCSES demonstrate evidence for the content facet of construct validity?
3. To what extent does the CCSES demonstrate evidence for the structural facet of construct validity?
4. To what extent does the CCSES demonstrate evidence for the external facet of construct validity?

The first research question sought to address a larger methodological issue—namely, what barriers exist to applying a unified framework of validity into practice. The subsequent three research questions assessed the extent to which CCSES scores exhibited different facets of construct validity: content, structural, and external.

CHAPTER 3

METHODS

Given the goals of the study, I used a sequential mixed methods design to collect qualitative and quantitative data from different sources. As I describe in this chapter, the qualitative approach allowed for the exploration of instructors' and students' thoughts regarding how they defined and identified student engagement behavior in the classroom and more broadly. Findings from interviews and focus groups were used to assess the content representativeness of the CCSES as well as to make the necessary revisions to the scale that would better reflect instructors' and students' views of student engagement. The qualitative data were also used to create five Likert-type questions to gather instructor and researcher ratings of students' engagement. Quantitative student engagement data were collected from students, instructors, and the researcher to assess the factor structure of the CCSES and to assess the convergence of all three ratings, which would ultimately provide evidence for the external facet of construct validity.

Participants

Instructors in the English department at Pasadena City College have developed a new course that is both accelerated and stretched across an academic year, rather than offered in one semester, called the Stretch Accelerated Composition Course (STACC). The purpose of this new course is to help students who place into developmental (or remedial) English complete the required course sequence faster. Currently, there are two levels of developmental English, and students who place into the lowest level have to complete two developmental English courses before they can take the transfer-level English course. With the STACC model, students who

place into the lowest level will complete both developmental English courses and transfer-level English in two semesters rather than three. The STACC curriculum also incorporates issues of social justice, multi-modal learning, affective domain (which deals with students' well being), and increased collaboration between students and faculty. The samples for this study were students enrolled in this STACC English course, and the instructors teaching these courses.

Instructors

Instructors were recruited to participate in the study during the 2014 spring STACC curriculum meetings. A detailed explanation of the purpose of this study as well as an explanation of what their participation would entail was provided. Initially, six out of ten STACC instructors indicated they would be willing to sit down for an interview and subsequently allow me to observe their classrooms during the fall 2014 quarter.

A total of five instructors eventually participated in the interview portion of the study. All five instructors were female. Three instructors were Hispanic, one instructor was White, and one instructor was Black/African American. Four out of the five instructors taught at least one section of the STACC English class, while the fifth instructor taught in the ESL department. Three instructors were part-time and two were full-time during the fall 2014 semester. All five instructors had master's degrees. The part-time instructors have been teaching for at least three years, and the full-time instructors have been teaching for at least five years. Of the five instructors that participated during the interview portion of the study, three eventually continued on to the classroom observation portion of the study.

Students

The study consisted of two different student samples. The first student sample participated in the focus group portion of the study. A total of four focus groups were conducted,

N = 22 students. Of the 22 students, ten were female and twelve were male. No age or ethnicity information was obtained from students in an attempt to keep their identity anonymous and their responses confidential. The second sample of students participated in the survey administration portion of the study. This second sample consisted of 77 students across four STACC classes. The majority of these students were Mexican/Mexican American (36.1%), followed by Other Hispanic (16.7%), Asian (16.7%), and White (15.3%). Black/African American, Native American, and Pacific Islander students accounted for the remaining 15.2% of the sample. The sample included a higher percentage of females (55.6%) compared to males (44.4%), and although students ranged from 18 to 47 years of age, the majority of students were 18 (41.7%) and 19 (31.9%) years old.

Procedures

As noted earlier, I used a sequential mixed-methods design that yielded both qualitative and quantitative data to assess different facets of construct validity. Data collection began with instructor interviews, followed by student focus groups, and classroom observations. It concluded with the survey to obtain engagement ratings from students and instructors. To supplement the data from students and instructors, I served as an external observer and also provided ratings of students' engagement. A subset of students from each class was selected, and detailed classroom observations notes were recorded for each of them. I provided engagement ratings only for this subset of students. Table 3.1 provides an overview of all the methods used to collect data for this study. The table provides the order and semester in which each method was used, as well as the sample and purpose of each method. The procedures section below will provide a more detailed narrative of each method used to collect data in the order in which is presented in the table. The development of the CCSES and additional measures used to conduct

classroom observations and collect instructor and research ratings of student engagement are presented in the Measures section below.

Table 2. *Summary of Methods Used to Collect Data*

Order of data collection	Semester	Method	Sample	Purpose
First	Spring 2014 (April)	Instructor Interviews	5 instructors, 4 were STACC English instructors, and 1 was an ESL instructor. Of the 5 instructors, 2 were full-time, 3 were part-time	-Assess content facet of construct validity for CCSES -Develop instructor/researcher engagement questionnaire -Develop class observation guide
Second	Spring 2014 (May)	Student Focus Groups	5 focus groups, 5–6 students each, for a total of 22 students.	Assess content facet of construct validity for CCSES
Third	Fall 2014 (September)	Classroom observations	Total of 4 class observations from 3 of the 5 instructors interviewed.	Researcher to observe a subset of student in each class to provide external ratings of engagement to assess the external facet of construct validity
Fourth	Fall 2014 (November)	CCSES survey administration to the students in the 4 classrooms observed	A total of 72 students across four classrooms	-Assess structural facet of construct validity -Use self-reported engagement ratings for convergent analyses with instructor and researcher ratings of engagement to assess external facet of construct validity
Fifth	Fall 2014 (November)	Instructor ratings of their students' engagement	Instructors provided ratings for a total of 72 students	Use instructor engagement ratings for convergent analyses with student and researcher ratings of engagement to provide external evidence of construct validity
Sixth	Fall 2014 (December)	Researcher ratings of students' engagement	Researcher provided ratings for a total of 27 students	Use researcher engagement ratings for convergent analyses with student and instructors ratings of engagement to provide external evidence of construct validity

Phase I: Qualitative Data

The qualitative data collection process consisted of three different methods (interviews, focus groups, and class observations) that spanned two academic semesters. Qualitative data were used to assess the content facet of construct validity as well as to create additional engagement materials for the purpose of creating an observation guide, creating a set of questions for instructors to use to rate their students' engagement, and creating a set of questions for the researcher to use to rate students' engagement. Further details regarding the development of the additional materials are presented in the Measures section below.

Instructor interviews. Ultimately, four STACC instructors and one ESL instructor followed through with the interview portion of the study. Instructors were interviewed either via phone or at a meeting space, such as an office or coffee shop. They responded to six open-ended prompts about how they defined engagement in general, and more specifically engagement in the classroom. They were also asked to identify specific classroom behaviors that would indicate that a student was or was not engaged, and to answer a general question regarding the connection between engagement and academic success. These interviews served three purposes: (a) to assess the content representativeness of the CCSES; (b) to create a classroom observation guide; and (c) to create five Likert-type questions for instructors and the researcher to use when rating students' class engagement (discussed in greater detail in the Measures section).

A copy of the instructor interview protocol is provided in Appendix B. Interviews lasted approximately 30 to 45 minutes, and were recorded with the instructors' permission. During the interviews, three out of the four instructors were asked to review the CCSES and provide feedback regarding the relevancy of the questions and domains, recommend whether any

questions needed to be revised/edited/removed, and indicate if there were any domains or questions not present that should be included.

Instructors who participated in the interview portion of the study during the spring 2014 semester were invited to participate for a second round of data collection, which included observing their classrooms, administering the CCSES to their students, and asking them to rate their own students' engagement. Of the four STACC instructors who were interviewed, three allowed me access to their classrooms.

Student focus groups. As part of the ongoing evaluation of the STACC program, a recruiting e-mail was sent to all STACC full-time and part-time instructors about having their students participate in focus groups to get feedback regarding the various elements, assignments, readings, etc., of STACC. Four instructors responded and were wiling to have their students participate in focus groups. The researcher served as the sole external evaluator for the ongoing STACC evaluation and conducted all the focus groups. With the permission of the instructors and consent of the students, five additional questions regarding student engagement were added to the focus group protocol.

One focus group per instructor class was conducted, for a total of four focus groups, (n = 22 students). The four STACC instructors who volunteered to have their students participate in focus groups were not the same four STACC instructors who participated during the interview portion of this study, with the exception of one instructor. Focus groups lasted no longer than 45 minutes and were conducted over the course of two weeks. A copy of the student engagement questions included in the student focus group protocol is provided in Appendix C.

The purpose of the focus groups was to get a sense of how students defined their own engagement, to understand if their definitions were similar to instructors' definitions of

engagement, and to determine whether their definitions were represented in the CCSES. Specifically, after students responded to questions regarding their STACC course, they were asked how they defined engagement, what engagement looked like in the classroom, verbal and non-verbal indicators of engagement, how they engaged with instructors, whether they participated in campus activities, and how they engaged with their peers.

Classroom observations. As noted above, instructors who participated during the interview portion of the study were invited to participate in classroom observations. Three faculty members allowed access to their classrooms, resulting in four classes being observed. Consent forms were provided and collected for all students in each classroom. All four classes met twice a week for 1 hour and 50 minutes. One class met on Monday/Wednesday and the three other classes met on Tuesday/Thursday.

Classroom observations lasted for 10 weeks, and I attended both weekly sessions of the Tuesday/Thursday classes when possible. Due to schedule conflicts I attended only the Monday sessions of the Monday/Wednesday class.. The first five weeks of the observation served as a testing period. This allowed me to get a sense of what was going on in each classroom, and also to determine how many students could be tracked at any give time for the purposes of providing external/researcher ratings of student engagement. I was also able to test the observation guide. During the second five weeks of observations, I began tracking and observing a subset of students in each class. I kept detailed notes for a total of 27 randomly selected students[1] across all four classes and across six different areas: posture/body language, participation in larger group discussion, instructor/student interaction, attendance, asking questions in class, and

[1] A random number generator was used to obtain a random sample of seven or eight students per class. A parameter between 1 and the number of students for each class was set, and the random number generator provided a starting point from which to include every 4th or 5th student into the observation pool, again depending on the number of students in each class.

group/pair participation. It was helpful to keep track of the students using an excel spreadsheet. Table 3.2 below depicts the structure of the spreadsheet used to record notes. I created a spreadsheet for each class.

Table 3. *Sample of Notes Obtained During Class Observations*

Student	Attendance	Prepared for Class	Group/Pair Participation	Asked Qs in Class	Posture/Body Language	Instructor/Student Interaction	Larger Group Discussion
Student A	Here – arrived on time	Has a draft of essay. Doesn't have copy of reading book.	10:15 a.m. Reading partner's essay draft, writing comments and underlining.	N/A	9:59 a.m. free-writing, sitting up straight, arms resting on desk.	10:25 a.m. instructor visiting w/her group. She is observing while another student responds to the instructor's questions. Alternates between looking up at the instructor as she speaks and looking through the textbook.	10:50 a.m. volunteered a response to instructor's question, "what is the social capital /deficit model?"
Student B	Here – arrived late (9:30 a.m.)	No draft essay. No copy of reading book.	10:19 a.m. Working with 2 other students who also do not have copies of their drafts. Talking about the weekend, classmate's 2-year-old child	N/A	10:00 a.m. free-writing, upper body leaning up against the desk.	10:35 a.m. instructor visiting her group, looking up at the instructor and presumed to be listening to her as she provides advice about what they need to do to catch up, since none of them have a draft to work on.	N/A
CLASS NOTES							
9:56 a.m. Instructor has asked students to free write in their journals while she writes the agenda up on the board. 10:06 a.m. Instructor has asked students to finish free-writing and take out their essay drafts; students have been asked to partner up with another student to review each other's drafts.							

Each observation session began as soon as the instructor started class. Typically, students started their classes with a free-writing exercise, and after noting that the class was working on their free-writing activity, I shifted focus to the students I was tracking. I kept detailed notes for each of the seven to eight students who were being tracked in any given classroom.

If I observed a student slouching while working on a free-writing assignment, then that activity was documented under the body language/posture column. When students transitioned from the free-writing exercise to group work, I took notes regarding what they were supposed to be working on before shifting focus to what the students were actually doing. During group activity, I moved closer to students who were being tracked, and recorded what they were doing (e.g., looking through the text book) and what the study was saying (e.g., summarizing the assigned reading to the group). I circulated throughout the classroom during group activity time to get detailed notes regarding what students were doing during group work. I documented these activities under the group/pair participation column. When an instructor circulated and visited with groups, I recorded how the students interacted with the instructor in the "Instructor-Student Interaction" column.

Phase II: Quantitative Student Engagement Data

Students. Students were administered the CCSES one time, during the 13th and 14th week of the fall 2014 semester. The CCSES was administered either at the beginning of the class or at the end of class; this was dependent on what worked best for the instructor. Students were provided with explicit instructions for how to respond to the four domains of the CCSES. Given that instructor and researcher ratings of class engagement were based on these four specific STACC courses, students were instructed to rate their class participation for their STACC course only. In an effort to provide the most accurate data, students were instructed to refer back to the

previous three weeks, rather than the entire semester, when answering the class participation questions. When answering the questions related to relationships with peers, relationships with instructors, and campus participation domains, students were instructed to refer to *all* of their courses. Again, they were instructed to refer to the previous three weeks of the semester. According to Porter (2011) students are more accurate at rating the frequency of their behavior when they are asked about more recent behavior, versus behavior for an entire semester.

Instructors. During the 15th week of the fall 2014 semester, the week after students were administered the CCSES, instructors were asked to provide ratings of their students' class engagement. Specifically, they were asked to indicate whether students submitted assignments on time, how often they contributed to class discussions, if they actively participated during group work, how often they interacted with students before, during, and after class, and if they came to class prepared. A copy of the five questions instructors used to rating their students' engagement is included in Appendix B. more detailed narrative about the development of these five questions is provided in the Measures section below.

The five instructor engagement questions were set up online (at SurveyMonkey.com). I sent the survey link to instructors and asked them to answer all five questions for each of their students. These five questions used a Likert-type scale from 1 (never) to 5 (always). Instructors were asked to refer to the previous three weeks when rating their students' engagement, to ensure that students and instructors provided engagement ratings that reflected roughly the same period of time.

Researcher. During the 16th week of the fall 2015 semester, I provided class engagement ratings for the 27 students who I tracked during the observation period. I used a set of five questions to rate students' engagement that was only slightly different than the five questions

instructors used to rate their students' engagement. These five questions were based on a Likert-type scale from 1 (never) to 5 (always). A copy of the researcher engagement questions is included in Appendix D. A more detailed narrative regarding the development of these five questions is provided in the Measures section below.

Data were collected and assessed for three of the six facets of construct validity. Table 3.3 below provides a definition for each of the three facets, as well as a summary of the methods and type of data collected to assess each of the three facets.

Table 4. *Three Facets of Construct Validity, Descriptions, and Proposed Strategies*

Type	Description	Strategy
Content	Identifying the boundaries of a construct that is being addressed. What skills, knowledge, behaviors, attitudes, etc., should be represented, given the boundaries of a construct.	Conducted interviews with instructors, focus groups with students, and classroom observations to assess the representativeness of the CCSES, and to identify aspects of student engagement that were not represented in the scale.
Structural	Reliability and an acceptable factor structure—one that reflects the construct in question (e.g., if student engagement is multi-dimensional, then is the factor structure of the CCSES reflecting this?).	Administered the CCSES to students to assess the underlying factor structure of the scale.
External	Relationships between scale scores and other measures/behaviors that are implicit in the theory of the construct. One way of measuring external relationships is the multitrait-multimethod matrix (MTMM).	Obtained student, instructor, and researcher ratings of student engagement for a partial MTMM framework to assess the convergence of all three ratings. Assessed the relationship and predictability of student, instructor, and researcher ratings of student engagement on academic achievement (English GPA and overall GPA for fall 2014 semester).

Secondary data. Secondary data regarding student demographics and GPA were obtained through the Internal Planning and Research Office (IPRO). The student data included age, gender, ethnicity, English GPA, and overall GPA for the fall 2014 semester. All identifying

information has been kept confidential and saved on a password-protected computer, and all names have been changed to protect student identities.

Measures

There were a total of four measures used to collect data for this study: CCSES, observation guide, a set of questions for instructors to use for rating student engagement, and a set of questions for the researcher to use for rating student engagement. The development of the CCSES occurred before embarking on this study. The other three measures however were created during the process of this study and guided by the qualitative data gathered by instructors and students. The goal was to have the class participation domain of the CCSES, the classroom observation guide, and the instructor and researcher questions align as closely as possible. A detailed narrative of the development of each measure is presented below.

Community College Student Engagement Scale (CCSES). Initial scale development consisted of developing a pool of potential scale items dictated by the behavioral, affective, and cognitive domains, and by engagement in five different contexts: class participation, relationship with faculty, relationship with peers, participation in campus activities, and utilization of campus facilities. For example, “I attend class on a regular basis” is a class participation item and “I attend faculty office hours to ask coursework related questions” is a relationship with faculty item.

Once the scale items were written, graduate students in a measurement course were asked to review them. The students classified the items on two levels: First, they identified whether each item was behavioral, affective, or cognitive; second, they identified whether each item was participation in the classroom, relationship with peers, relationship with faculty, participation in campus activities, or utilization of campus facilities. The students provided a strong consensus

and agreed the items were written and identified appropriately. When an item was identified as belonging to a factor other than the one it was written for, it was revised for clarity. Students were also asked to provide feedback regarding the wording of items and the appropriateness of the item scales. Item revisions were made according to their feedback.

For behavioral items, students were asked to rate how often they engaged in school activities on a 5-point Likert-type scale, from 1 (never) to 5 (always). For affective and cognitive items, they were asked to rate how often they agreed with statements on a 5-point Likert-type scale, from 1 (completely disagree) to 5 (completely agree). The behavioral domain contained 16 items, the affect domain contained 16 items, and the cognitive domain contained 15 items. The class participation target factor contained 10 items, relationship with faculty and staff contained nine items, relationship with peers contained 11 items, participation in campus activities contained eight items, and utilization of campus facilities contained nine items. The initial engagement scale contained 48 items. Three models were tested: a three-factor model of behavioral, affective, and cognitive engagement, a five-factor model of class participation, relationship with faculty, relationship with peers, participation in campus activities, and utilization of campus facilities, and an eight bi-factor model that incorporated the three domains and five contextual factors.

This first version of the scale was administered to a sample of students (N = 376) at California State University, San Bernardino. After assessing the factor structure using exploratory and confirmatory methods, results indicated that the cognitive and affect domain overlapped significantly (r = .90), however neither the cognitive nor affect domain correlated more than r = .30 with the behavior domain. The cognitive and affect domains also overlapped significantly with context specific factors, and the utilization of campus facilities factor did not

hold. Based on these results, the student engagement scale was revised to include only one domain (behavior) and four contextual factors: class participation, relationship with peers, relationship with faculty, and participation in campus activities.

The second iteration of the student engagement scale contained 30 behavioral items. Due to the fact that minimal engagement scale development has been done with the community college population, I elected to administer the revised scale to a sample of community college students (N = 510) to create an engagement scale that was relevant for the population of students in two-year institutions.. Exploratory and confirmatory (second order model) analyses revealed that there were four distinct first order factors that measured student engagement in the classroom, relationship with peers, relationship with faculty, and participation in campus activities. The covariance or correlation between these factors could be explained by the fact that they all tapped into behavioral engagement. In total, five items were eliminated during analyses because their factor loadings were less than .32. The revised student engagement scale contains 25 items. The administration and analyses of the second iteration of the engagement scale was conducted prior to embarking on the current study.

Observation guide.

The development of the observation guide was created using the qualitative data obtained from instructors and students, as well as using the class participation domain of the CCSES. The goal was to have the observation guide reflect instructor and students' definitions of behavioral engagement – specifically those that could possibly be observed in a classroom, as well as have the guide reflect the items in the class participation domain. After reviewing interview and focus group notes, as well as the questions in the class participation domain, the following domains were created: posture/body language, participation in larger group discussion, instructor/student

interaction, attendance, asking questions in class, and group/pair participation. A copy of the observation guide is included in Appendix E.

Instructor questions for rating engagement. Five questions that asked about classroom engagement were developed for the instructors to use when rating their students' engagement. These questions were based on instructor interview data, the domains in the observation guide, and the questions in the class participation domain of the CCSES. These questions were designed to capture three things: (a) behavioral engagement in the classroom; (b) consistency with the class participation domain of the CCSES; and (c) consistency with how instructors defined class engagement.

Instructor interview notes, the questions in the class participation domain, and the domains in the observation guide were reviewed to find commonalities between these three sources. Instructors were asked to rate their students' engagement based on these commonalities: did students submit assignments on time, how often students contributed to class discussions, if students actively participated during group work, if students came prepared for class, and how often instructors interacted with their students. A copy of these questions is provided in Appendix B.

Researcher questions for rating engagement. Rather than developing a new set of questions, I used the same questions as the ones created for the instructors to rate students' engagement, with the exception of one question. The question that asks about students submitting assignments on time was not used, as I would not be able to answer this question by relying on classroom observations. Instead, I used a question regarding class attendance. A copy of these questions is provided in Appendix D.

Analysis

Instructor Interviews and Student Focus Group Data

Data analysis began with verbatim transcription of all interviews and focus groups. For instructor interviews, transcripts were reviewed to identify common responses to interview questions. These common responses generated a list of preliminary codes that reflected instructors' views of student engagement. This first round of preliminary coding was done by hand, as was a subsequent round of coding. The second round of coding was used to ensure that codes identified during the first round of coding captured the range of responses provided by instructors. These codes were grouped into three larger categories. This same process was repeated to generate a list of codes for student responses.

To assess if there were similarities and/or differences in instructor views of student engagement compared to student views of student engagement, a constant comparative analysis was conducted. Constant comparative analysis is an iterative process by which data are reduced through constant recoding to allow for possible categories to emerge (Fram, 2013). Instructor codes were compared with student codes to identify similarities and differences. Ultimately, there were few differences between instructor and student codes.

Student Engagement Data

To analyze the student engagement data, I first conducted a series of exploratory factor analyses (two, three, and four factor models) to assess the underlying factor structure of the CCSES. This also allowed me to assess whether the resulting factor structure contained domains identified in the literature as key elements of student engagement. Once the best factor structure was identified, reliability analyses were conducted for all factors.

Then a series of correlations were conducted to assess the convergence of student, instructor, and researcher ratings of class participation engagement. I also assessed the relationship between student and instructor ratings of class participation engagement and English GPA and overall GPA for fall 2014. The next step was to assess whether there was a predictive relationship between student, instructor, and researcher ratings and English GPA and overall fall 2014 GPA.

Lastly, I conducted a series of simple regression models to asses: (a) if all three ratings of class participation engagement were significant predictors of English GPA, and which of the three ratings was a better predictor of English GPA; (b) if all three ratings of class participation engagement were significant predictors of overall fall 2014 GPA, and which of the three ratings was a better predictor of overall fall 2014 GPA; and (c) if any of the other three factors of the CCSES (relationship with peers, relationship with faculty, and campus participation) were significant predictors of overall fall 2014 GPA. Moderate to high correlations, as well as statistically significant regression models, would collectively provide external evidence of construct validity by demonstrating convergence between all three ratings of class participation engagement, and by establishing relationships and predictions of academic success (GPA) as dictated by student engagement theory.

CHAPTER 4

RESULTS

Current theory argues for a unified framework of construct validity rather than a view of validity as consisting of at least three independent types. Assessing validity in this way requires collecting data to assess multiple facets of construct validity. Thus, for this study, qualitative and quantitative data were gathered in order to answer the research questions regarding the validity of the Community College Student Engagement Scale (CCSES). Specifically, three different facets of construct validity were assessed: content, structural, and external.

Assessing the content facet of construct validity meant identifying the boundaries of the student engagement construct by gathering data from content experts and then assessing if the CCSES encompassed those boundaries. For this study, instructors and students were used as content experts and their definitions of student engagement were used to assess the content of the CCSES. Assessment of the structural facet of construct validity required analyses of the structure of the CCSES, which is typically done with exploratory factor analyses and required that the CCSES be administered to the target population to gather quantitative data. For this study, the CCSES was administered to community college students and their responses were used were exploratory factor analyses. Assessing for the external facet required analyses that assessed relationships between other measures and/or behaviors implicit to the student engagement construct. For this study, the relationship between student, instructor and researcher ratings of engagement were analyzed, as well as the relationship between the engagement ratings and GPA

to assess the external facet of construct validity. Results for each facet of construct validity are presented below.

Content Facet of Construct Validity

In their interviews, community college instructors put forth a multi-layered and nuanced definition of classroom engagement. Collectively, they indicated that engagement in the classroom consisted of three elements: engagement with the material (e.g., assigned readings, assignments), engagement with other students, and engagement with the instructor. Rather than relying solely on verbal and behavioral cues, instructors often stated that non-verbal behavior (e.g., eye contact) was also a good indicator of whether students were engaged. In this section, I explore these issues in greater depth, and describe how they relate to the CCSES.

Engagement with Course Material

According to instructors, engagement with course material was the most significant aspect of class participation engagement. Specifically, students were not only reading the books, articles, and class handouts, and watching films, but they were also in one way or another reflecting, thinking about, and trying to understand the issues and ideas that were embedded in these course materials. When asked to specify what engagement with course material looked like, one instructor said, “Are they turning in work, are they able to relate to it in someway, whether that is answering a certain question, asking a question, making a product in response to it.…So they are making a poster or they are making conversation, or they are writing. They are doing something with the content.” Another instructor indicated, “Engagement with the material has different components. It’s like, did you [student] do the reading? Did you [student] write the essay? It’s also, what are you [student] doing with the feedback I’m giving you?”

These quotes illustrate that instructors used a wide range of behaviors in class to assess whether students had read the various course materials, as well as to assess whether they understood and were synthesizing the material. According to instructors, students who were engaged with course material would keep up with course readings, take notes during class lectures, and were able to integrate readings into their writing, whether in essays, free-writes, or in-class reflective writing assignments. Students who were engaged with course material turned assignments in on time, and contributed to larger class discussions by making comments, asking questions, or providing opinions regarding the broader themes/issues addressed in course readings.

In regards to asking questions in class, one instructor indicated, “I pay attention to the kinds of questions they are asking.” Indeed, instructors made it a point to differentiate between the types of questions asked in class, noting they were a good indicator of how students were engaged with the material. As one instructor explained, “Questions can be basic, when students ask about assignment due dates and page requirements.” In contrast, questions that included some element of trying to understand, apply, or synthesize course material were identified as “advanced” and “multilayered.” In addition, instructors indicated that students who were engaged with course materials were better able to actively participate and contribute during smaller group work.

For instructors, engagement with course material could lay the foundation for students’ engagement with peers and instructors. Without engagement with the material, class discussions may become stagnant, there is less productivity and collaboration during group work, and assignments reflect a general lack of understanding of reading assignments and any concepts

discussed during class. As this observation of a class session makes clear, a lack of engagement with the material can bring a class to a halt.

> On October 7th, class started as usual, students were instructed to take out their writing journals and answer the prompts the instructor had written up on the whiteboard. After the free-writing exercise, students moved their desks to form a circle and began discussing their responses to the free-writing assignment. The discussion was stagnant. After about 30 minutes of forced conversation, the instructor asked students to take out their reading assignment, an article titled "Mi Voz, Mi Vida." Upon noticing some students partnering up to share a copy of the article, the instructor asked, "How many of you have a copy of the article with you?" Slightly less than half the students raised their hands. The instructor then asked, "How many of you actually read the article, whether you have a copy of it in front of you or not?" About one third of the class raised their hands. "This is a problem, we can not continue with what I had planned because it required that you guys did the reading." The instructor then stood up from the desk and asked, "How do you think we should handle this?" She walked over to the whiteboard, picked up a marker, put her hand up on the whiteboard and asked, "I want to know the reasons why you guys didn't do the readings." She held her position, waiting for students to speak up, so she could document their responses on the whiteboard.

Instead of having the planned class discussion on the topic of resilience, and eventual group work centered on identifying key themes of the article, the class turned into a discussion centered on personal responsibility. This day was a perfect illustration of how a course can go awry when students do not engage with course material. It is important to note, however, that

this situation was not the norm, and classes proceeded as planned for the majority of the observation time frame.

Engagement with Peers

Another element of classroom engagement that instructors mentioned was interaction between students. Instructors were asked, “When I say engagement what do you think of?” One instructor stated, “I think of participation, so are the students relating and connecting and interacting with each other.” Another instructor indicated that she assigned participation points, and part of the participation points had to do with how students were interacting with each other, she assessed if, “ students are providing meaning comments to each other.”

Additional comments from instructors indicated that peer interaction consisted of students contributing equally during group work, willingness to help each other with assignments during and outside of class time, and generally asking and responding to each other’s questions during larger class discussions and smaller group discussions. Bringing essays drafts to class and partnering up with other students to provide each other constructive feedback was also considered an important element of peer engagement.

Engagement with Instructors

Instructor perspectives. Lastly, instructors indicated they considered interaction between students and instructors to be an important element of classroom engagement. Classroom observations quickly revealed that there was plenty of opportunity for instructors and students to interact during class time. This type of engagement can take different forms: students seeking one-on-one help when the opportunity presents during class; students interacting with the instructor when he/she circulates and visits each group during in-class group work; and

students' incorporation of instructor feedback on their writing assignments or their follow-through on suggestions to visit the writing center to get help with their essays.

In addition to providing behavioral and verbal elements of classroom engagement, instructors also stated that there were non-verbal indicators of classroom engagement, including body language or posture. Often, they said that engaged students would sit upright or lean forward in their desks, and they had a tendency to maintain eye contact with the instructor when he/she was speaking. Engaged students would typically give some indication that they were paying attention, either by nodding their heads, tilting their heads to the side, or making some kind of facial gesture like raising eyebrows, laughing, or possibly even frowning. Instructors also indicated that engaged students came prepared to class, meaning they had their reading books, MLA handbook, articles, writing journals, essay drafts, and other necessary materials. These cues were so important that one instructor actually said she relied primarily on non-verbal behaviors to gauge whether a student was engaged.

Reliance on non-verbal behaviors to define classroom engagement was due to at least two factors. First, instructors acknowledged that there were personality and cultural differences between their students, and as a result not all showed their engagement by asking questions or participating in larger class discussions. In other words, just because a student was not very verbal during class time did not necessarily mean that the student was *not* engaged. Second, and conversely, just because a student *did* voluntarily speak, share, or answer questions during class did not always mean that the student *was* engaged. A prime example of the second point was noted during a preliminary classroom observation during the spring 2014 semester.

> On this day, the class was having a discussion on the use of language and words to convey an emotion or set a tone for the reader. The instructor projected a paragraph from

an unknown article up on the screen. She asked students to read the paragraph and then describe the mood the author was trying to convey. As an exercise, the instructor asked students to alter whatever words necessary that would change the mood or tone of the paragraph into whatever mood or tone they wanted. After a few minutes, one student raised his hand, eager to share what he had created. Originally, the paragraph conveyed a mood of loneliness as the author used words to describe the dreadful, damp, and empty bar he was sitting in. This particular student changed the story to that of a lonely guy sitting at a bar, meeting a woman who walked into the bar, and who ultimately turned out to be a man. The class roared with laughter. As the class proceeded on to do some group work, the instructor sat next to me, pointed out three students, including the one who shared his story, and said, "These guys, they like talking. They are quick to raise their hand and say something, but I can't get them to turn anything in!" I asked, "So they do not turn in assignments or essays?" "No, never on time, I have to pull it out of them…[I'm] always on them."

These students voluntarily participated during class exercises, yet they were not doing their coursework and only turning in (late) assignments after much prodding from the instructor. In speaking with other instructors about this situation, most of them indicated that they had encountered students like these before. They were quick to say these students were probably seeking attention rather than wanting to have an honest discussion about the course material. One instructor said, "Sometimes, it will come down to, they are not really engaging with the material. They are big talkers, they have personalities, but they are not revising their drafts or they don't have a draft ready. I recommend tutoring, but they are not following up on that, so they are not attending to some of the stuff they need to attend to, and can have performance issues."

Given their experiences in the classroom, instructors were able to provide depth to their definitions of classroom engagement. They consistently looked at multiple indicators to make judgments about whether students were engaged, knowing that simply talking or turning in assignments did not *always* constitute class engagement. Students, on the other hand, had a harder time articulating their thoughts regarding engagement. When pressed, they were able to provide definitions of engagement that mirrored those provided by instructors. I discuss these definitions next.

Student perspectives. During the focus groups, the majority of students indicated that engagement in the classroom meant being active, asking questions or speaking up during class, and interacting with classmates and the instructor. They initially tended to define engagement in terms of activities they could see, as one student noted, "As long as you are doing something physically to put in effort to keep up with the class." After some thought, however, many said it was possible to not be active, but still be engaged in the class. Some relayed stories about classmates who were shy and would not say much in class, but who took a lot of notes, came to class prepared, turned in their assignments, and were doing well. In their minds, if a student was doing well, then they must be engaged.

Many students had a hard time articulating non-verbal indicators of engagement, and were better able to identify non-verbal behaviors of students who were *not* engaged. They indicated that they knew, almost intuitively, if a fellow student was not engaged. One student said, "Sometimes you can be looking, but there is nothing there," presumably attempting to explain that some students may be looking up, appearing to pay attention, but have blank stares, which indicates a lack of engagement. Another student said, "They are staring off into space, they are just there." Yet another student said, "I can look around the classroom and say 'he's in his world.'"

CCSES Capture of Student Engagement

Interviews with faculty, and to a certain extent the student focus groups, provided a sense of whether the CCSES encompassed this multi-layered definition of student engagement, particularly in the classroom. More precisely, for the most part, the class participation domain of the CCSES contains items that tap into the instructors' and students' definitions of class engagement. It contains items that relate to engagement with course material, such as questions about using course lectures/information to complete assignments, completing homework assignments on time, note taking, and actively participating in small group work, keeping up with assigned readings, and being prepared for class. Table 4.1 below demonstrates how the indicators of class engagement, as identified by instructors and students, correspond to the items of the class participation domain of the CCSES. The quotes included in the table are responses that instructors or students provided when asked to define student engagement.

Table 5. *Evidence of Content Validity for Class Participation on the CCSES*

Source	Evidence	CCSES Class Participation Domain Questions
Instructor interviews	"Do they come to class prepared with the readings?"	I kept up with assigned readings.
Instructor Interviews	"Taking notes is an indication that a student is engaged."	I took notes during class lectures.
Instructor Interviews	"I check if students have the materials, if there [are] relevant materials to be had."	I came prepared to class (ex: paper, pen/pencil, books, drafts, free-writing journals).
Student Focus Groups	"Taking the right materials."	
Instructor Interviews	"I check to see if students have incorporated feedback, lecture into their assignments."	I used information from class lectures to complete homework assignments.
Instructor Interviews &	"Engagement with course material means, 'Are students turning in their work on time?'"	I completed homework assignments on time.
Student Focus Groups	"Continually keeping up with your work."	

Source	Evidence	CCSES Class Participation Domain Questions
Instructor Interviews	"Attendance is important too."	I attended class meetings on a regular basis.
Student Focus Groups	"Going to class and being on time."	
Instructor Interviews	"Collaborating with other classmates." "I also have to see how they engage in group work."	I actively participated in group activities during class (ex: contributed ideas, listened/responded to group members, completed my work on time).
Student Focus Groups	"Communicating with other students, especially when you have a group project."	
Instructor Interviews	"Asking questions and providing responses during class."	I contributed to class discussions (ex: by asking questions, responding to questions, sharing opinions.
Student Focus Groups	"Participating in class discussions."	
Student Focus Groups	"She would walk around during class, asking us if we needed help."	I communicated with my professor during class (ex: asked questions right before, right after class, asked questions during group work).

Ultimately, the interviews and focus groups provided evidence for the content facet of construct validity for the class participation domain of the CCSES. In addition, interviews with instructors also provided an opportunity to have them review the entire CCSES to assess the domains related to relationship with peers, relationship with faculty, and campus participation. The instructors were asked to assess the domains and the questions within each domain. They asked if there were any questions that required revision or deletion, or if there were any questions that should be added.

Three of the five participating instructors reviewed the scale, and all indicated that the appropriate domains and questions appeared to be included in the CCSES. Their review of the entire CCSES provided preliminary evidence for the content facet of construct validity. Minor changes were made to the CCSES based on instructor interview feedback, including adding two questions to the class participation domain that reflect student-instructor interaction in the classroom.

Structural Facet of Construct Validity

Exploratory factor analyses were conducted to assess the factor structure of the CCSES, despite the small sample size (n = 72). Previous exploratory and confirmatory efforts on the CCSES revealed a four-factor model of student engagement; previous results are included in Appendix F. Although the sample size is not ideal, I elected to proceed with the EFA anyway, in order to assess whether the current four-factor model would hold with a new and much smaller sample.

EFAs were conducted using multiple factor solutions. Principal axis factor extraction and varimax rotation were used for all solutions. Questions that loaded less than .32 were suppressed. The analyses did confirm that items were functioning well in the context of a factor structure. The four-factor solution was deemed the better factor structure, as it was the most interpretable solution. It also contained the four-factors that were consistent with the hypothesized class participation, relationship with faculty, relationship with peers, and campus participation factors.

There were two questions related to relationship with peers that ultimately ended up loading on the campus participation factor. “I got together with my classmates to socialize” and “I use my classmates as a source of information regarding school resources” both loaded onto the campus participation factor. Both questions differed from the rest of the relationship with peers questions in that they asked about peer engagement outside of the classroom. All questions that loaded onto the relationship with peers factor pertained to engagement between peers in academic activities (i.e., helping each other with homework, sharing notes, talking about what courses to take). In seeing this distinction between the relationship with peers questions, it made sense that the two questions referenced above would load onto the campus participation factor. There were two items that did not load onto the participation in campus activities domain. Table 4.2 below identifies which items loaded onto each factor, as well as factor loading indices.

Overall, it can be said that there is sufficient evidence of a factor structure that is consistent with student engagement literature. The factor structure also reflects the definitions and descriptions of student engagement provided by instructors and students.

Table 6. *CCSES Items and Loading for Each Factor*

	Class Participation	Relationship w/Faculty	Relationship w/Peers	Campus Participation
I actively participated in group activities during class meetings.	.758			
I contributed to class discussions.	.718			
I kept up with assigned class readings.	.644			
I completed homework assignments on time.	.597			
I communicated w/instructor during class.	.592			
I came prepared to class.	.420			
I took notes during lecture.	.400			
I attended class meetings on regular basis.	.396			
I used information from lecture to complete homework.	.386			
I talked to my professor about my academic plans.		.814		
I attended faculty office hours to discuss my assignment grades.		.770		
I attended faculty office hours to ask coursework-related questions.		.754		
I attended faculty office hours to discuss exams/essays.		.733		
I talked w/professor about what other classes to take in the future.		.704		
I talked to professor about personal issues.		.667		
I communicated w/professor online.		.510		
I shared my class notes with my classmates.			.820	
My classmates shared their notes with me.			.820	
I asked classmates help with homework assignments.			.726	
I got together with my classmates to study or complete homework assignments.			.575	

	Class Participation	Relationship w/Faculty	Relationship w/Peers	Campus Participation
I talked to classmates about what classes to take next semester.			.369	
I attended workshops offered on campus.				.669
I attended campus events and/or activities.				.638
I attended information sessions offered on campus.				.608
I got together w/classmates to socialize.				.766
I used my classmates as a source of information regarding school resources.				.646

Once the four-factor model was identified as the most interpretable, reliability analyses were conducted to assess the internal consistency of the four factors. A series of Cronbach's alphas were calculated and provided the following reliability estimates: class participation = .79, relationship with faculty = .88, relationship with peers = .78, and participation in campus activities = .81. According to Streiner (2003), coefficients between 0.7 and 0.8 are deemed acceptable, and coefficients between 0.8 and 0.9 are deemed good. Reliability estimates for all four factors are within the acceptable and good range.

External Facet of Construct Validity

A total of 77 students completed the CCSES, however a final sample size of 72 was used for all subsequent analyses. Five cases were dropped for the following reasons: three students did not provide their names, and as a result I could not include them on the student list provided to instructors so they could rate students' engagement; one student completed the CCSES but was enrolled in a different section of the class; and one student only completed about half of the CCSES. Missing data were imputed using the EM Algorithm method. There was minimal missing data, approximately four cases.

For each factor, as well as for instructor and researcher ratings, summing across items created engagement scores. Table 4.3 below provides the possible minimum and maximum values for each factor in the CCSES, instructor class engagement ratings, and researcher class engagement ratings, as well as the mean and standard deviation for each.

Table 7. *Descriptive Statistics for Each Factor of the CCSES and Engagement Ratings*

Domain	N	No. of Items	Min	Max	Mean	SD
Relationship w/instructors	72	7	7	35	14.64	6.44
Relationship w/peers	72	5	5	25	15.36	4.79
Campus activities	72	5	5	25	12.70	4.94
Student class engagement ratings	72	9	9	45	35.49	5.36
Instructor class engagement ratings	72	5	5	25	17.30	2.83
Researcher class engagement ratings	27	5	5	25	17.55	3.75

To provide external evidence of construct validity, a series of correlations was conducted to assess if the student, instructor, and researcher ratings of class engagement would converge, and to assess if these three ratings correlated English GPA and overall GPA. I then conducted a series of regression models to assess if student, instructor, and/or researcher ratings would predict English GPA and/or overall GPA.

Correlations

I conducted Pearson correlations to assess relationships between student, instructor, and researcher ratings of class engagement. Student, instructor, and researcher ratings were all normally distributed and linearly related. Using the MTMM framework, correlations between measures of the same trait (class engagement) using different methods (student, instructor, and researcher) should be moderate to high to demonstrate convergence.

Given that sample sizes for student, instructor, and researcher ratings of engagement were different, two different techniques for dealing with missing data were used to calculate correlations between the three sources of engagement ratings. Correlations were conducted using pairwise and listwise deletion methods. Pairwise deletion uses available cases for any given variable, whereas listwise deletion uses complete cases only and removes data for a case that is missing one or more values. Regardless of deletion methods, correlations indicated that ratings between the three different sources of class engagement ranged from moderate to high, and were statistically significant (refer to Table 4.4 and Table 4.5 below).

The correlations using a pairwise deletion method (Table 4.4) indicate that the correlations between were statistically significant at the $p = .01$ significance level. The relationship between student and instructor class engagement ratings was $r = .64$, indicating that as student class engagement ratings increased, so did instructor ratings. The lowest observed correlation was between student class engagement ratings and researcher class engagement ratings, $r = .58$, indicating that as student class engagement ratings increased, so did researcher ratings. The highest observed correlation was between instructor and researcher ratings, $r = .65$, indicating that as instructor ratings increased, researcher ratings also increased.

Correlations between student and instructor ratings and instructor and researcher ratings were essentially the same, however the correlation between researcher and student ratings differed. The rating questionnaire used by the instructor and the researcher overlapped quite a bit, rating students on the same four questions out of five. Thus, the difference in the one question may account for this difference in the correlation between the student–instructor correlations and the student–researcher correlations. Another possible explanation for this difference is that I was the only observer in the classrooms, and some information may have

been missed during the observation process. However, because I was observing for the very specific behaviors included in both the instructor and researcher questionnaire, it is possible that I was able to observe more than the instructor, thus accounting for the difference in the correlations noted below.

Table 8. *Correlation Matrix for All Three Engagement Ratings (Pair-wise deletion)*

Source	Student	Instructor	Researcher
Student	1.0	---	---
Instructor	.64**	1.0	---
Researcher	.58**	.65**	1.0

***p<.01*

The correlations in Table 4.5 are the same as the ones above, with the exception of the student – instructor correlation. Using the listwise deletion method reduced the sample size for student, instructor, and researcher to n = 27. The decrease in the correlation between instructors and students is accounted for by the decrease in the sample size. Using the pairwise deletion method, the sample size for the instructor-student correlation was n = 72. For subsequent analyses, it was elected to use the pairwise deletion method to maximize sample size.

Table 9. *Correlation Matrix for All Three Engagement Ratings (List-wise deletion)*

Source	Student	Instructor	Researcher
Student	1.0	---	---
Instructor	.39*	1.0	---
Researcher	.58**	.65**	1.0

***p<.01*

** p <.05*

I conducted an in-depth look at the ratings of the 27 observed students to identify if there were any cases in which student, teacher, and researcher ratings diverged. Table 4.6 below displays data for four cases in which my own ratings of students were at least twice as high or twice as low as the ratings provided by the instructors. Student, instructor, and researcher ratings were converted into Z-scores to allow for comparisons across all three ratings. There were no instances in which instructor and researcher ratings diverged (i.e., the instructor rated a student above the mean where I rated that same student above the mean and vice versa).

Table 10. *Z-score Class Engagement Ratings*

	Instructor Ratings	Researcher Ratings	Student Ratings
Student A	-.15	-1.52	-.84
Student B	.65	1.31	.655
Student C	.39	1.32	1.22
Student D	-.15	-1.52	-.66

I relied on observation notes more than memory when rating students' class engagement, and this may offer one possible explanation for the gap between researcher and instructor ratings. Additionally, observations revealed that some students tended to be more on-task when the instructor was nearby or when the instructor was visiting their group. As a result, it is possible that the instructor observed students when they were on their best behavior, so to speak.

In the content section above, instructors made it a point to say that being verbal and active in the class did not necessarily mean a student was engaged with the course material. Case in point, two students were rated above average in class engagement but they ended up doing poorly in the class. Table 4.7 below reports these cases. In both instances, the students' own ratings were the highest, followed by the researcher and instructor ratings. These students either

did very poorly on all of their assignments or did not turn in assignments, but continued to attend class. Perhaps there were extraordinary circumstances that interfered with their ability to complete their assignments, or perhaps they believed that being verbal and active in class was enough to earn a passing grade, regardless of the grades they received on their assignments.

Table 11. *Z-score Class Engagement Ratings and Grades*

	Instructor Ratings	Researcher Ratings	Student Ratings	Final Grade
Student A	.12	1.31	1.40	D
Student B	.12	.96	1.22	F

I conducted a second series of correlations to assess if student, instructor, and researcher ratings of class engagement were related to final grades in English class (referred to from this point on as English GPA) and overall GPA for the fall 2014 semester. I also correlated each of the five instructor rating questions to English GPA and overall fall GPA, to assess if any one item correlated higher to the GPA outcomes. English GPA and overall fall 2014 GPA were calculated in the way that the Office of Institutional Effectiveness at Pasadena City College calculates them—by dividing quality GPA points by the number of GPA points earned. Both English GPA and overall GPA ranged from 0.0 to 4.0. Given the ranked nature of English GPA (overall GPA is on an interval scale), I conducted Pearson and Spearman correlations. Both tests produced the same pattern of results, and all correlation coefficients were within 0.1 of each other. As a result, I elected to present the results of the Pearson correlations in Table 4.8, and to treat the data as parametric for the remainder of the analyses.

Table 12. *Correlation between Engagement Ratings and GPA*

	English GPA	Overall GPA
English GPA	---	---
Overall GPA	.83**	---
Student Ratings	.36**	.35**
Instructor Ratings	.71**	.64**
Researcher Ratings	.35*	.42*

***p<.01, *p<.05*

All three ratings of class participation engagement were statistically significant.

Not surprisingly, instructor ratings of class participation engagement correlated the highest with English GPA, r = .71, followed by student ratings of class participation engagement, r = .36, and researcher ratings of class participation engagement, r = .35. All three correlations were positive, indicating that as instructor, student, and researcher ratings of class participation engagement increased, there was also an associated increase in English GPA. Instructors' rating of students' engagement consistently reflected the final grade that students earned in class. The high correlation between instructor ratings and English GPA might imply that how students were performing in the class influenced instructors' ratings of their engagement. It should be noted that instructors were not asked to rate the quality of their students work or how they were currently doing in the class.

When looking at the relationship between overall GPA and class participation engagement ratings, results show that student, instructor, and researcher ratings were all statistically significant. Again, not surprisingly, instructor ratings correlated the highest with overall GPA for the fall 2014 semester, r = .64, followed by researcher ratings, r = .42, and student ratings, r = .35. All correlations were positive, indicating that as instructor, researcher,

and student ratings of class participation engagement increased, there was an associated increase in overall fall 2014 GPA.

If and how class participation engagement ratings in English class would correlate with overall GPA for the 2014 fall semester is unknown, partly because it is reasonable to assume that engagement levels can fluctuate from class to class. For example, a student who feels more competent in math than in English may be more engaged in math class than in English class. That overall GPA is correlated with student, instructor, and researcher ratings could mean at least two things. First, it is possible that some students only took English during the fall 2014 semester, so their overall GPA would be the same as their English GPA. Second, it is possible that students' levels of class participation engagement did not fluctuate significantly from class to class.

The last series of correlations attempted to assess the external facet of construct validity for the other three domains of the CCSES (relationship with peers, relationship with faculty, and campus participation). Specifically, it assessed the relationship between these domains and English GPA and overall fall 2014 GPA. Table 4.9 below displays all possible pairwise correlations between CCSES subdomains and GPAs.

Table 13. *Correlation Matrix between Other Factors of the CCSES and GPA*

	English GPA	**Overall GPA**
English GPA	---	---
Overall GPA	.83**	---
Relationship w/peers	-.10	-.13
Relationship w/faculty	-.05	-.06
Campus participation	.57	.08

***p<.01, *p<.05*

Results indicate that relationship with peers, relationship with faculty, and campus participation did not correlate with English GPA or overall GPA. In fact, the correlations between relationship with peers and relationship with faculty, although not statistically significant, were negative. On the surface, these results may appear to be undesirable, however the items in these domains ask about student behaviors *outside* of class. Additionally, engagement literature indicates that establishing interpersonal relationships and participating in extracurricular activities at school help students build a sense of community, and, as a consequence, increases the likelihood of students persisting from one year to the next (Fredericks, et al., 2004). The extent to which these three domains correlated with persistence data would have been a better indicator to assess for the external facet of construct validity compared to English GPA and GPA for the fall 2014 semester.

Some of the questions in the relationship with peers domain asked students about how often they socialize with their classmates, how often they talk to their classmates about courses to take, if they use their classmates as a source of information regarding school resources, and if they ask classmates for help with homework. If students rated themselves highly on these questions, then it indicates that they were spending a lot of time socializing with classmates, and it is conceivable that spending too much time with classmates/friends could interfere with coursework, hence the negative relationship between the domains.

Some of the relationship with faculty domain questions asked students if they would attend office hours for a variety of reasons, such as for questions about homework or exams/essays, personal issues, academic plans, or coursework related questions. Although not statistically significant, the relationship with faculty domain was negatively correlated with English GPA and overall fall 2014 GPA. Higher scores on the relationship with faculty domain

were associated with lower GPAs. One possible explanation for this relationship is that if students are spending a significant amount of time in office hours, they may not be doing well in class. As such they may have a lot of questions regarding their essays or their coursework in general, or they may have personal issues that are getting in the way of their work.

Regression analyses. In an effort to provide additional external evidence of construct validity, multiple regression analyses were conducted to assess whether student, instructor, and researcher ratings of class participation engagement were significant predictors of English GPA. This analysis would also provide information as to which of the three ratings—student, instructor, or researcher—was the better predictor. Given that instructor ratings had the highest correlation with English GPA, it was hypothesized that instructors' ratings would be a significant predictor of English GPA.

I also conducted analyses to determine whether student, instructor, and researcher ratings of class participation engagement were significant predictors of overall fall 2014 GPA. Again, given that instructor ratings correlated highest with overall fall 2014 GPA, it was hypothesized that, at minimum, instructor ratings would be a significant predictor. Given the low and non-significant correlations between GPAs and relationship with peers, relationship with faculty, and campus participation, I elected not to run a regression assessing the predictability of these three subdomains of the CCSES.

Prior to any analyses assumptions of normality, linearity, and homoscedasticity were assessed. P-P plots, histograms, and scatter plots of the residuals indicated that all assumptions were met. Additionally, no outliers were observed in the data. As indicated above, there were four cases that were dropped, resulting in a final sample size of n = 72 for student and instructor ratings, and n = 27 for researcher ratings.

Model one examined whether student, instructor, and/or researcher ratings predicted English GPA. This model was statistically significant, R = .727, R^2 = .53, Adjusted R^2 = .47, F_{change} (3, 23) = 8.59, p <.01. A summary of regression analyses is presented in Table 4.9. Student, instructor, and researcher ratings of class engagement accounted for approximately 47% of the variance in English GPA. Although the overall model was statistically significant, only instructor rating was a significant predictor of English GPA, unstandardized b = .28, t = 4.24, p <.01. For every one-unit increase in instructor engagement rating, there is an expected .28 unit increase in English GPA, holding constant student and researcher class engagement ratings.

Table 14. *Results of Multiple Regression Analysis, English GPA by Engagement Ratings*

	t	p	B	F	df	p	Adj R2
Engagement Ratings				8.57	3, 23	.001	.467
Instructor	4.24	.001	.282				
Student	-.571	.574	-.111				
Researcher	-.804	.430	-.158				

Model two examined whether student, instructor, and/or researcher ratings predicted overall fall 2014 GPA. This model was statistically significant, R = .646, R^2 = .42, Adjusted R^2 = .34, F_{change} (3, 23) = 5.502, p <.01. A summary of regression results is presented in Table 4.10. Student, instructor, and researcher ratings of class engagement accounted for approximately 34% of the variance in overall fall 2014 GPA. Although the overall model was statistically significant, again only instructor rating was a significant predictor of English GPA, unstandardized b = .181, t = 2.96, p <.01. For every one-unit increase in instructor class participation engagement rating, there is an expected .18 unit increase in overall fall 2014 GPA, holding constant student and researcher class engagement ratings.

Table 15. *Results of Multiple Regression Analysis, Overall GPA by Engagement Ratings*

	t	p	B	F	df	p	Adj R2
Engagement Ratings				5.502	3, 23	.005	.342
Instructor	2.96	.007	.181				
Student	-.489	.623	-.020				
Researcher	.198	.844	.015				

Instructor rating of class participation engagement was the only predictor that was statistically significant across both models. This was not surprising, for a few reasons. First, it would be expected that instructors' ratings would correlate the highest and also predict English GPA. They have a more complete profile of who their students are, they observe them during class, they read and grade their work, and they interact with them during class and during office hours. When instructors were asked to rate their students' class engagement, it appears likely that how students were performing influenced their ratings. Second, researcher ratings were based strictly on what could be observed in the classroom; these ratings did not include some of the other indicators related to English GPA and overall GPA, such as the quality of work or assignment grades and study habits. Third, students' class participation engagement ratings were expected to be a significant predictor of their English GPAs, however that was not the case. It is possible that students were not the best raters of their own behavior. Another (statistical) issue, however, is that the ratings between students, instructors, and the researcher correlated at least moderately, and perhaps there was too much overlapping variance for more than one of these predictors to be statistically significant. Also, the small sample size for researcher ratings (n = 27) may have been one of the reasons this predictor was not statistically significant.

Overall, results indicate that there is evidence for the external facet of construct validity. Student, instructor, and researcher ratings of class participation engagement converged.

Additionally, student, instructor, and researcher ratings correlated with English GPA as well as overall GPA for the fall 2014 semester, providing further evidence of the external facet of construct validity. Ideally, all three engagement ratings would have been statistically significant predictors of English GPA, and overall fall 2014 GPA, however only instructor rating was a significant predictor of GPA.

There is evidence to suggest that instructor ratings of students' engagement tended to take into account how students were performing. The rubric that instructors were provided to rate students' engagement did not ask them to rate students on academic performance, but rather strictly classroom behaviors of engagement. The fact that instructors indicated that, for the most part, doing well in class and being engaged go hand in hand supports the assertion that they used that information when rating students' engagement.

Additional evidence to support this claim comes from Z-scores of instructor and researcher ratings of engagement. There were instances in which researcher ratings, based strictly on what students were doing in class, would have led to the conclusion that the student earned a "D" or "F" in the class, because Z-scores were at least one standard deviation below average. In these instances, however, students actually earned "C"s, and instructor ratings were only slightly below average. The opposite situation was never true. That is, there were no instances in which instructor ratings deviated significantly from students' final grades in the class. What does this mean in regards to measuring student engagement? It illustrates the importance of adding an academic performance domain to the CCSES, or perhaps adding academic performance questions to the class participation domain. This domain and/or questions should address students' perceptions about how they are doing in class, about the quality of their class work, and time spent on homework during class and outside of class, to list a few.

Findings Related to Student Engagement Patterns

Classroom observations were conducted to gather external ratings of student engagement for purpose of assessing the external facet of construct validity. However, classroom observations and engagement ratings also revealed an interesting finding regarding student engagement patterns over the course of ten weeks. These results, although not entirely having to do with validity of the CCSES, are presented below.

Not surprisingly, classroom observation data revealed that students varied in the degree to which they were engaged in the classroom. Based on classroom observations, and student, instructor, and researcher engagement ratings, students could be categorized into three general levels of engagement: high, low, and somewhere in between. Their levels of engagement were fairly constant throughout the ten-week observation period. It seems reasonable to assume that students would be more or less engaged depending on what they were doing in class, or depending on whether there was an impending deliverable (i.e., exam or final essay). However, observing these students proved otherwise.

Highly engaged students. Highly engaged students missed one class period at most, were always on time, and came prepared to class (i.e., had essay drafts ready for peer review, and always had their MLA handbooks, reading books, articles, and free-writing journals). There was a bit of variability in their tendency to participate in larger group conversations. Some tended to speak out more than others, but those who were quieter exhibited important non-verbal behaviors, which were identified by the instructors during their interviews. Specifically, they tended to take a lot of notes, appeared to be paying attention (e.g., always turned to face whomever was speaking), and would sporadically use their electronic devices. They were also observed actively participating during group activities—they would often be the spokesperson

for their group during presentations, helped coordinate schedules for meetings outside of class to get group work done, and were typically the ones to interact with the instructor when the group had questions.

One example of a "highly" engaged student is Ralph. Ralph was an 18-year-old Hispanic male student enrolled in the Tuesday/Thursday morning STACC class. He was identified as highly engaged based on classroom observations and engagement ratings. I rated Ralph well above the mean in class participation engagement (Z = 1.31), and the instructor also rated Ralph above average in classroom engagement (Z = .40). Ralph rated himself well above average in engagement (Z = 1.22). He ultimately earned a B in the class and had an overall GPA of 3.0 for the fall 2014 semester. A class observation from the first week of the study highlights Ralph's typical behavior in the classroom:

> On this day, students spent most of the class time working in pairs. Students were asked to bring in a draft of their essays and then work with another student on identifying thesis statement, topic sentences, evidence, etc. While students worked on their drafts, the instructor invited anyone to visit with her if they wanted feedback on their drafts. When the instructor asked the class questions about "What is a topic sentence…thesis statement?," Ralph volunteered responses on at least three occasions. Ralph then quickly pulled out his draft essay and partnered up with another male student sitting in front of him. They pushed their desks together to sit side by side. While his partner read aloud his essay, Ralph was noted to be following along on his copy, underling and making comments. When it was Ralph's turn to read aloud his essay, he chose to visit with the instructor instead. For 15 minutes, Ralph visited with the instructor, read parts of his essay, asked questions, and wrote comments on his essay as the instructor spoke to him.

When he was finished, he returned to his partner and they continued working on their drafts for the remainder of the class time. During his time, Ralph and his partner were observed to be looking down at each other's drafts, underlining, crossing out, and making comments on their drafts.

Another observation, this one from the third week of data collection, strengthens the interpretation of Ralph as highly engaged:

> Students spent most of their time working in groups on this day. They briefly engaged in a free-writing exercise and then were placed into groups. Students were to create outlines of the first three chapters of *Critical Race Counter Stories Along the Chicano/Chicana Educational Pipeline*. Ralph was in a group with three other students; once groups were formed, he immediately took out his laptop and could be overheard asking all group members for their Gmail accounts, as he was creating a Google document so they could all work on the assignment. Over the course of 1.5 hours, Ralph was the one in charge of gathering all their answers and typing them on the Google document. He would periodically walk over to the instructor and ask her questions about the assignment. When the instructor would visit his group to check on their progress, Ralph would respond to her questions, show the instructor the progress in their outline, and read back what he had written to confirm they were all on the same page. The last 15 minutes of class consisted of a larger group discussion centered on race. Ralph was an active participant in this discussion, raising his hand to provide his opinion on "what is race?" and "what is majoritarian?" As class was ending, Ralph was strategizing with his group members about who was going to do what, and he volunteered to be responsible for putting all the parts together in one Google document.

Low engagement students. A second group of students could be categorized as "low engagement." These students were noted to be absent between two and four times over the course of the observation period. They generally did not come to class prepared (i.e., they always had to borrow or share classmates' books or articles), they did not bring in draft essays for peer review, and they rarely contributed to larger class discussions. On occasion, a few students were observed dropping off their essays prior to class starting and then leaving. They were on their electronic devices frequently during class lectures and discussions, and in some instances they were sleeping during class. These students usually did not contribute much during group work. A prime example of a low engagement student is Mark.

Mark was a 19-year-old Hispanic male enrolled in the Monday/Wednesday evening STACC class. Mark was identified as a low engagement student, as noted by class observations and engagement ratings. I rated Mark more than two standard deviations below the mean (Z = -2.23), the instructor rated him almost one and a half standard deviations below the mean (Z = -1.48), and Mark rated himself over one standard deviation below the mean (Z = -1.21). Mark ultimately earned an F in the class, and finished with an overall GPA of .36 for the fall 2014 semester.

The following observation from the second week of the study illustrates Mark's typical level of engagement in the class:

> On this day, students began class with the customary free-writing exercise, which gave way to a larger class discussion regarding the assigned reading, *The Joy Luck Club*, followed by students working in pairs. While in pairs, students were supposed to write responses to the questions the instructor put up on the whiteboard. Mark arrived on time, as he generally did, and worked on his free-writing assignment briefly. During the class

> discussion, Mark did not have a copy of the reading with him, did not contribute to the class discussion, and when it was time for group work, immediately partnered up with the female student sitting beside him. During the time in which pairs were supposed to be writing responses to the questions up on the board, Mark was observed to be on his phone for most of the time. He would periodically look over to see what his partner was writing, and then would go back to his phone. When the instructor asked each pair to visit with her briefly to check their responses, his partner went up alone. He remained seated, looking around the room, was fidgeting with a piece of paper, and stared straight ahead at the whiteboard, while his partner got feedback from the instructor. When his partner joined him again, his same general pattern of behavior continued until the end of class.

Likewise, in the fourth week of observations, it was clear that Mark's low level of engagement had continued:

> On this day, rather than the free-writing exercise to begin the class, students were handed a quiz to work on for the first 30 minutes. The quiz tested students' knowledge of MLA citation rules. After students completed the quiz, the instructor handed each student someone else's quiz, and they were going to grade them as a class. The second half of the class consisted of group work—creating arguments for and against a "college freshmen taking ethnic studies courses" prompt. When handed a quiz, Mark was observed sitting at his desk, staring blankly up at the board, while his classmates started working on the quiz. He continued to sit, look around the room, fidget with his pencil, and lay his head on the desk for the next 17 minutes before he started working on his quiz. When asked to pass up his quiz, he continued working until the instructor walked over and took it from him. The class then reviewed answers and each student was responsible for grading

someone else's quiz. During this time, Mark was observed periodically checking the quiz he was supposed to have been grading, writing on the quiz twice, and then eventually falling asleep during this part of the class. After approximately 10 minutes, presumably in an effort to wake him up, the instructor called on him to provide the answer for the next question. His lack of engagement continued during the second half of class. During group work he could be overheard talking about various different sports he enjoyed watching, boxing in particular, and at one point got up and left the classroom for approximately 15 minutes. He told a fellow group member that he stepped outside to, "wake up…it's cold outside." When it was time for each group to select a spokesperson to share their arguments for or against the prompt, all group members immediately pointed to Mark. Two group members said to Mark that he should be responsible for presenting their responses since he had not contributed anything to the group. Mark quickly and repeatedly said "No." Ultimately, someone else from his group volunteered to share the group's responses.

Average engagement students. The third group of students could be categorized as "middle" or "average" engagement students. These students were not as engaged as those in the "high engagement" category, but they were definitely more engaged than students in the "low engagement" category. They more often than not came prepared to class (i.e., had their books, articles, essay drafts), and did not have more than two absences. They were noted to occasionally contribute to class discussions when prompted by the instructor. They were also seen actively participating during group work, but they could also get distracted and veer off topic.

As noted earlier, it was surprising that engagement levels appeared to remain fairly stable throughout the ten-week observation period. This was especially true for students who could be

categorized as “high” and “low” engagement. For students who were in this middle category, there were more instances in which engagement levels fluctuated, especially during group work. This fluctuation depended on the composition of the group. Specifically, if these students had group members who were highly engaged, they tended to focus more on class work. However, if they had group members who were not engaged, group work tended to taper off, and the group would begin to engage in conversations unrelated to their assignments.

A few students noted this dynamic, as well. One student said, “Who you surround yourself with, where you sit…if they are engaged, it makes me even more engaged.” Another student agreed, saying, “Yeah, definitely, who sits in front and who sits behind you.” This dynamic of influence was evident during the first week of observations, when two students I was observing closely were in the same group. Alisha was identified as being somewhere in the middle in terms of engagement; I rated her engagement close to average (Z = -.10). She was consistently distracted by Pam, whom I identified as a low engagement student (Z = -1.52):

> Alisha and her group were huddled towards the back of the classroom, each with their critical race counter stories book in hand. All three members were highlighting/underlining, taking notes, and placing Post-It notes on various pages as they flipped through the pages of chapter two. Sometimes they would talk with each other and then eventually return to flipping through the chapter. Pam arrived to class approximately 45 minutes late. She walked towards the back of the classroom and joined her group, and a group member quickly filled her in on the task at hand. Pam then sat and looked around, did not have a copy of the book, and watched her classmates flip through the book and would check Facebook on her phone. Minutes later, however, Pam was overheard talking to Alisha and one other group member about her daughter and showed

them pictures of her daughter. The conversation then veered into other topics, such as the availability of a summer class schedule and a discussion of the summer classes they took last year. These conversations lasted until students were asked to reconvene as a class. Pam was then observed positioning her body in such a way that would allow her to text on her phone without being seen by the instructor. She would alternate between looking up at the instructor and texting on her phone for the remainder of the class time.

Pam was more absent than not during the observation period. When she did attend class and join her group, however, she tended to distract Alisha. These fluctuations in engagement were not drastic; I never observed an instance in which a student went from being highly engaged to not at all engaged (or vice versa) depending on the composition of the group. When a group contained both high and low engagement students, they did not significantly influence each other's behavior. For example, Rosa was identified as a high engagement student based on class observations notes and my rating of her engagement (Z-score = .98). She also earned an A in the class and had a fall 2014 overall GPA of 3.76.

During one observation, Rosa's class was working on a group research paper. She was grouped with two other students, Maria and Scott. Rosa and Maria worked diligently during designated group time. I observed them searching for articles on Rosa's laptop, alternating between flipping through their textbooks and taking notes on Rosa's laptop, asking the instructor questions, and responding to the instructor's questions when she visited their group. Scott, although not one of the 27 students tracked in the study, was identified as a low engagement student. During group work, he spent most of the time on his phone. Occasionally he would get up and speak to another male student. During one class period, he was observed not saying a single word to his group members.

Rosa worked diligently, regardless of who was in her group. Scott and his lack of contribution, interest, or engagement did not distract her from her work. In the same sense, Rosa did not influence Scott's behavior during their group time. They existed as two separate entities within the same group space. This was not the only occasion or the only group in which this dynamic was observed. In these instances, or in these groups, highly engaged students worked alone or with one other student who was equally engaged, while the remainder of the group had conversations on topics not related to class assignments (e.g., gaming consoles, Halloween party plans, tattoos).

It is possible that how students feel about particular subject matter—in this case, English—can influence their engagement in class. If this is the case, then it has to be assumed that students who are highly engaged have more positive attitudes towards English and those with low engagement have negative views toward English. This is arguably an overly simplistic view of what makes students more or less engaged in class, however. It should also be noted that although classes tended to have a general structure, students' engagement levels stayed the same, even when they engaged in out-of-the-ordinary activities (i.e., Jeopardy games, movies, gallery walks). Although fluctuations in engagement were not observed for the subset of 27 students over the course of the observation period, it does not mean that fluctuations in engagement (i.e., from low to high) cannot happen.

Classroom Observations and Validity of CCSES

How did classroom observations help with validity efforts? Classroom observations provided knowledge about what student engagement looks like in the classroom. As such, items on the CCSES could be assessed for relevancy. Items in the class participation factor were deemed relevant, as the factor contained items applicable to students' daily experiences in the

classroom, such as group work, class discussions, keeping up with assigned readings, and note taking during lecture. A few items were worded slightly differently to make them more applicable to students, such as providing specific examples regarding what it meant to be "actively" participating in group work, and what it meant to participate in class discussions. Classroom observations made these revisions possible. In addition, classroom observations identified one other area of class participation that was not anticipated or included in an earlier version of the CCSES: engagement with the instructor during class time. Observations revealed that students had plenty of opportunities to interact with the instructor on a one-on-one basis or within a smaller group setting, so an additional item regarding how often students interacted with instructors was added.

Conclusion

Just how valid is the CCSES in measuring students' engagement levels? Data collected from instructor interviews, student focus groups, and classroom observations indicate that there is sufficient evidence of the content facet of construct validity for the class participation domain of the CCSES. Based on instructors' review and feedback of the three other domains of the CCSES, there is preliminary evidence of the content facet of construct validity. However, results for the external facet of construct validity indicate that perhaps the inclusion of an academic performance domain or the inclusion of academic performance questions is warranted, and the inclusion of such a domain/questions would better align with the way instructors assess student engagement.

Results also indicate that, structurally, the CCSES corresponds with the multi-dimensional nature of the student engagement construct, as defined by the engagement literature. External results show that instructors' ratings of their students' class participation engagement

was the only significant predictor of English GPA and overall fall 2014 GPA. With the inclusion of the academic performance domain/questions in the CCSES scale, it is hypothesized that students' self-reported ratings of engagement will be a predictor of their GPA.

CHAPTER 5

DISCUSSION

Current validity theory no longer views validity as consisting of three independent types. Rather, theorists now argue for a unified framework that identifies various different evidences or facets of construct validity. Studies looking at current validity practices for measurement yearbook-type publications indicate, however, that practitioners and researchers have not been assessing validity in this way (Cizek, Rosenberg, & Koons, 2008). Additionally, only approximately 2% of measurement entries have reported more than two types of validity evidence, with most replying on only one type of validity evidence (Hogan & Agnello, 2004). Research on validity practices in published articles indicates that a third report indicate having measured validity for the surveys used to collect data for the study, but only approximately 26% actually report some kind of validity evidence (Barry, Chaney, Piazza-Gardner, & Chavarria, 2013).

The current study was an attempt to address the gap between validity theory and current validity practice by using the unified framework of construct validity to assess a particular measurement tool, the CCSES. Rather than relying on just one type of validity evidence, data were collected to assess the content, structural, and external facets of construct validity. A concerted effort to study validity as a unified framework provided an opportunity not only to assess the validity of the CCSES, but also to document and provide feedback on the process of using a unified framework. Implications for using a unified framework and for encouraging and measuring student engagement are discussed after a brief summary of the results below.

Summary of Findings

The CCSES incorporated feedback from instructors and students as well as domains identified in the literature as necessary for measuring student engagement. For example, as Fredricks et al. (2004) noted, engagement has been identified as consisting of either behavioral, affective, or cognitive domains. Jimerson et al. (2003) identified contextual factors of student engagement, such as classroom behaviors/activities, interpersonal relationships, extracurricular activities, and academic performance. In keeping with these assertions, the CCSES was constructed as a *behavioral* student engagement scale with *four contextual factors*: class participation, relationship with peers, relationship with faculty, and participation in campus activities.

Results indicated that the CCSES is indeed a four-factor scale with class participation, relationship with peers, relationship with instructor, and participation in campus activities contextual factors. There was sufficient evidence to establish the structural facet of construct validity. The underlying factor structure of the CCSES was multi-dimensional with class participation, relationship with peers, relationship with faculty, and participation in campus activities contextual factors that reflect the student engagement literature. Results also indicate that the inclusion of one more domain, or additional questions regarding academic performance, might better align with how instructors assess engagement.

Conclusions regarding evidence for the external facet of construct validity were mixed. Results indicated that student, instructor, and researcher ratings of class participation engagement were moderately to highly correlated with each other, demonstrating convergence of all three ratings, as well as offering partial evidence for the external facet of construct validity. Regression analyses were conducted to gather additional evidence for the external facet. These

analyses showed that instructor rating of class participation engagement was the only significant predictor of English GPA and overall fall 2014 GPA.

Implications for Measuring Student Engagement

That instructor ratings were the only significant predictor of English GPA and overall GPA highlights the importance of consulting with and including instructor feedback regarding student engagement measures. Instructors' views of student engagement may not necessarily match what student engagement theory has to say about operationalizing the construct, but they are the ones ultimately evaluating students based on academic performance, participation, engagement, and a host of other metrics. Incorporating instructor views into student engagement scales is a good way to strengthen the link between student engagement and academic success.

Results from the external facet of construct validity provide some evidence that perhaps the content of the CCSES was missing a domain or questions related to academic performance. There is evidence to suggest that instructors' ratings of their students' engagement were influenced by how the students were performing in the classroom. I hypothesize that by including questions and/or a domain regarding academic performance/activity, the CCSES will better align with instructors' views of class participation engagement, as well as improve the possible predictability of course GPA and overall GPA.

Implications for Student Engagement Practices

Student engagement levels remained fairly consistent throughout the observation period. It is reasonable to assume that students would be more engaged as deadlines approached, or would be more or less engaged depending on what they were doing in class for any given day. This was not case, however, even when writing assignment due dates were approaching, or even when instructors introduced different activities during class time. This finding raises questions

about the factors that may inhibit or facilitate student engagement, and whether these factors are instructor-based, curriculum-based, or student-based—or possibly a combination of all three. If student engagement levels do indeed remain fairly stable across an entire semester, if not longer, then the most important question to address would be how to intervene and help students who are low engagers.

Perhaps an obvious recommendation would be for instructors to create a more dynamic learning environment. Importantly, at the community college level, instructors are content experts, and do not necessarily have the pedagogical training that K–12 teachers do. As noted earlier, these instructors did, from time to time, introduce different activities in the classroom that were not part of the students' day-to-day class routines. Even these activities did not appear to move the engagement needle for low engagement students. In fact, at the end of one classroom observation in which the instructor had students play a game of Jeopardy to teach them MLA formatting rules, she approached me and asked, "What else do you think I can do to have some students participate more?" This is an indication that these instructors assumed some responsibility for students' engagement.

If some students were not responding to instructors' efforts to make the classroom a more dynamic place, at what point does the responsibility of being engaged in the classroom fall on the students themselves? During the focus groups, students acknowledged as much. One said, "We listen to things we want to listen to, so it's not on the teacher, it's on the student to be engaged." Another commented that engagement "depends on the student." One instructor spoke a lot about a particular student (not one of the 27 students observed during classroom observations) who was not at all engaged in class, was failing, and had multiple individual meetings with him. What made this student memorable to the instructor was that he seemingly

went from not being engaged to being engaged overnight. She had not done anything different to bring about a change in his behavior. When asked about it, she said, “Hopefully he will be in one of your focus groups so you can ask him.” When I spoke with this student, he indicated that he was disillusioned with college, felt like other students were immature and disruptive, and felt like college was just an extension of his high school experience. He said he made the decision to change his behavior when he realized that he could not change anything about his environment except for himself.

Although students did acknowledge that they should assume some responsibility for their own engagement in class, this does not mean that instructors should not continue to try to make the classroom a more active learning environment. Nevertheless, it would serve students well to know and understand that not all of their college courses will be inherently interesting to them; they will have to take courses that may be challenging, and they are ultimately responsible for how they respond to these challenges.

Colleges would do well to incorporate into their orientations some discussion of what it means to be a college student and how college is different from high school, rather than bombard students with information about services, programs, deadlines, etc.—information that may not be relevant to the students at that point in time. Perhaps having shorter, multiple orientations would be more beneficial, as discussions about support services may be more relevant during the first few weeks, once they are actually enrolled in courses and have had a bit of college experience. Discussions about where to get tutoring, help with financial and/or personal problems, or any other services that could potentially help students early in the semester could address some of the possible issues that may interfere with students’ abilities to become and/or remain engaged in courses.

The research revealed a great deal about student engagement at the community college level, and much of this information would have been missed, had it not been for a concerted effort to gather as much data and evidence of validity as possible. Simply remaining at the theoretical level about what it means for students to be engaged, without sitting in a classroom for ten weeks, would not have stimulated questions regarding what factors inhibit or facilitate student engagement or about the stability of engagement. In the same fashion, theoretical discussions about the difficulties of conducting validity studies with a unified framework without attempting to actually do so does not help the field. Certainty, there were difficulties in conducting the validity study and also revelations about the validity process. I turn to these issues next.

Implications for Assessing Validity as a Unified Framework

Using the unified framework for gathering various evidence of construct validity was difficult. First, time and resource constrains made it impossible to gather evidence for all six facets of construct validity. Instead, I collected data for three of the six facets, and even this limited approach had its difficulties.

Unpredictability

Some circumstances were out of my control. Specifically, the students who were targeted for observation did not always show up to class, and some eventually dropped out, reducing the sample size. In addition, some students enrolled in the targeted courses did not complete the CCSES. Moreover, I found it necessary to delay survey administration until a day when instructors could allot a few minutes for the procedure. On occasion, class meetings were cancelled. Because of scheduling issues, I was limited in the classes that I could observe. The only way to handle these circumstances was to adjust data collection strategies as they arose.

This included extending the observation period to account for cancelled classes and high absenteeism. When students were outside of the classroom attending workshops, I also attended these workshops and took notes whenever possible.

Conflicting Evidence

Another difficulty with collecting multiple evidences of construct validity is determining what to do when one or more evidences provide conflicting information. In the current study, there were significant correlations, which was a great first step in establishing external validity. However, regression analyses indicated that only instructor rating of classroom participation was a significant predictor of English GPA and overall GPA. This is somewhat conflicting evidence, considering the expectation that, at minimum, instructor and student ratings would be significant predictors of academic success.

There are three possible reasons for the conflicting external evidence in this study: (a) there was a construct issue—student engagement theory suggests that engagement is a predictor of academic success. However, it is possible that the link between student engagement and academic success did not hold for this population of students, which is not far-fetched, considering there is minimal student engagement research being done at the community college level; (b) there was a measurement issue—although there was sufficient evidence of the content facet of construct validity, it is possible that the CCSES is still missing key items or elements of engagement, and this is likely the case here, as detailed above; and (c) there was a statistical issue—the correlations between the three ratings were moderate to high, especially between instructor and researcher ratings. It is possible that researcher ratings were not statistically significant because of a small sample size (n = 27) compared to instructor and student ratings (n = 72). Moderate to high correlations between the three ratings also indicate a fair amount of

shared variance between the predictors. Instructor rating was the only significant predictor because it correlated highest with English and overall GPA.

When there is conflicting evidence, it is useful to assess all data that are collected, as this can provide clues as to whether the issue is construct-based, measurement-based, or statistical. One possible approach is to make the easiest possible revisions to the scale/measure first and then reassess the validity evidence. This iterative process can make validity seem to be an endless cycle, although validity theory does state that validity is an iterative process that should be conducted when a measure/scale/test is being used in a different context.

Non-linear Process

Engaging in this study also showed that validity is not a lock-step process, one in which a facet is addressed, crossed off the list, and not revisited. Instead, this process has illustrated that data or findings for a particular facet may provide insights or information about another facet. As noted above, although there was sufficient evidence to establish the content facet of construct validity, findings from the external facet of construct validity highlighted a need to include an additional domain or at least additional questions about student engagement, due to the fact that instructor rating was the only significant predictor of English and overall GPA, possibly because the instructors incorporated academic performance into their ratings.

Lack of Explicit Guidelines

Another issue that makes it difficult to study validity using this framework is that there are no guidelines for how many different evidences are needed to establish construct validity, or for whether one facet of construct validity is more important than another. As Messick (1988) stated, validation cannot rely on just one type of evidence, nor is there a particular type of

evidence required for establishing construct validity. In fact, there has been pushback on the current view of validity being a unified framework for this very reason.

Brennan (1998) stated that the unified framework provides no guidelines for test validity. Likewise, Fremer (2000) indicated that construct validity has been elevated to the point where the unified framework seems impractical. Lissitz and Samuelsen (2007) went so far as to suggest a reclassification of validity, from a unified framework to an "internal" and "external" validity framework. Their reasoning is again that the unified framework for construct validity is too ambiguous and does not provide guidance for assessing test validity. Moss (2007) has pushed back on this validity reconstruction, however, because it "appears to move away from a generative understanding of validation as a scientific inquiry in the unitary approach, (back) toward a representation of validity in terms of general methodological prescriptions that the unitary approach was intended, in part, to overcome" (p. 470).

To address the lack of explicit guidelines, a researcher must take the time to plan. Assessing the amount of time and resources that can be devoted to this process is a good starting point for deciding how many possible evidences can be addressed. Kane (1992, 2006) has suggested that researchers/practitioners create a comprehensive research plan for studying validity, and use this plan to guide decisions about which research activities seem feasible. Kane called this the "interpretive argument" approach.

Another factor that should drive this process is the purpose of the scale/test/measure itself. Is a measure being created to gather participant data to make decisions about the effectiveness of a program? Is a test being used to make decisions about who gets placed into developmental math courses? If the stakes of a test or measure are high, and the consequences of making a wrong decision are great, then the investment of time and resources into establishing

evidences of construct validity should also be high. As Moss (2007) indicated, "Although low-stakes tests may require only evidence gathered during the development stage, high-stakes tests appropriately require a more extensive evaluation of the fully developed test in use" (p. 475).

Lack of Sequential Process

Just as there are no explicit guidelines for how many evidences to address, there are also no explicit guidelines regarding the order in which facets of construct validity should be addressed. There is, however, something to be said about where to begin with this process of establishing construct validity. What became clear during this process was that it was important to address the content facet of construct validity before attempting to address the external and structural facets. It would have been irrelevant to establish structural evidence for the CCSES if it did not cover the boundaries of the student engagement construct. In essence, it would mean establishing structural evidence for a scale that only partially measures student engagement.

In the same vein, establishing external evidence for construct validity using a scale that only partially measures student engagement could lead to the wrong conclusions. Establishing a relationship or a predictive relationship between the CCSES and GPA, for example, would not necessarily mean that there is a relationship between student engagement and GPA, as the CCSES would not be measuring the full range of student engagement behavior. Assessing content evidence seems like a good place to start, since including the range of appropriate content is the first step in developing tests/measures, and there has likely (hopefully) been some work done to assess the content of the construct in question. Additionally, assessing for the external facet of construct validity before assessing the underlying factor structure (structural facet) would seem to be out of order.

For this study, there was an inherent order to assess the facets of construct validity—first content, second structural, and lastly external. It can be argued that this order would hold for any other validation study that was assessing the same three facets of construct validity. Overall, there appears to be some natural progression to evaluating these different facets. This progression does not mean that assessing for validity is strictly a linear process, however. A validity study may start off assessing for content evidence, then move on to structural evidence, and then external evidence. There might be instances, however, in which results from the structural or external facets warrant modifications to the content of the measure.

Limitations

This study was not without its limitations. Ideally, with additional time and resources I would have done a few things differently. First, it would have been ideal to have an additional external rater, and this was part of the original study proposal. A second external rater would have allowed for the assessment of reliability of the external ratings. The classroom observations turned out to be too time consuming for a second researcher to take on the responsibility, however.

Second, it might have been helpful to video record the classroom sessions, rather than conduct only classroom observations and document classroom activities by extensive note taking. Videotaping would have captured more classroom behaviors for a larger group of students. I did attempt to videotape all four classes, however the three instructors indicated that they were not comfortable with it, and some students also expressed concerns. In addition, it would have been difficult for one person to manage multiple video cameras, always having to move them around to follow the students who were being tracked. The advantage that more traditional observations provided was the flexibility to move around the classroom and get a

close-up account of how students were engaging. Ultimately, both classroom observations and videotaping would have been ideal.

Lastly, the plan was to have administered the CCSES twice during the observation period, to obtain additional supporting evidence regarding the stability of student engagement across time. How often and when I was able to administer the CCSES was left to the discretion of the instructors. CCSES data collection was pushed back because classes got cancelled due to the instructor not being there, or because students were spending class time at the library attending workshops, or students were attending a First Year Experience conference. In three instances, instructors forgot that they allowed for data collection on a specific day, and the agenda for that day was so full that I was asked to postpone to the following class period or week. Although certain processes could have been conducted differently, they were out of my control, and the best possible adjustments were made to ensure that the best possible data were collected, given the circumstances.

Future Research

This study is an improvement upon many other validation studies for three reasons: (a) this study used the most current validity framework, rather than relying on an outdated framework; (b) the unified framework of construct validity provided a guide to the different possible facets that could be investigated and, as a consequence, multiple sources of data were collected and different facets of construct validity were assessed, rather than simply relying on exploratory or confirmatory analyses results and claiming that the CCSES is valid, as many studies currently do; and (c) by making a concerted effort to use the unified framework of construct validity, feedback regarding the validity process and lessons learned could be shared.

There are several different ways in which the current research can be expanded. For example, longitudinal studies of student engagement can deepen our understanding of whether engagement levels remain constant across time (e.g., during a semester or academic year) and across different classes. Additionally, being able to identify shifts in engagement, for example students going from low engagement to high engagement and vice versa, would provide a great opportunity to study the factors that facilitate change in engagement levels. This would also be a great opportunity to address a gap in student engagement research, as studies measuring engagement across time are largely missing from the student engagement literature.

Future validity studies should move away from the fragmented and outdated framework that they continually rely on, one that states that there are three separate types of validity. Continued attempts to assess validity using a unified framework can potentially improve validity theory and practice, because they allow for the sharing of experiences, strategies, and lessons learned. Providing concrete examples of validity practice in different contexts will help researchers and practitioners.

The validation process is time consuming, resource intensive, and at times difficult. There will be times where data collection strategies do not go according to plan. However, these are not sufficient justifications for short cutting the validity process. As Humbly and Zumbo (1996) stated, “It is no longer sufficient to simply demonstrate the correlation between a measure and a gold standard (a correlation that at one time was referred to as a validity coefficient) or simply perform a factor analysis and label the factors” (p. 214). When measures/tests are used to make high stakes decisions about the effectiveness of programs, about whether or not particular students will graduate from high school, or whether certain students will have to start their college careers in remedial education courses, the only fair and justifiable thing to do is to take as

much time and as many resources as possible to establish that the interpretation of these tests/measures is accurate.

Concluding Remarks

The goal of this research was, in part, to provide an example of what a sound validity study looks like in the student engagement context. As noted in Chapter 2, many validity studies of student engagement scales have relied on factor analyses as a validation process. This study was an attempt at a comprehensive process using current theory that can speak to the validity of a particular instrument and provide a concrete example of validity research in a particular setting. This study was a successful first step in establishing this goal.

APPENDIX A

Copy of the CCSES that was administered during this study

Class Participation Domain

1. I kept up with assigned readings.

2. I took notes during class lectures.

3. I came prepared to class (ex: paper, pen/pencil, free-writing journal, draft essays).

4. I used information from class lectures to complete homework assignments.

5. I completed homework assignments on time.

6. I attended class on a regular basis.

7. I actively participated in group activities during class meetings (ex: contributed ideas, listened and responded to group members, completed my part on time).

8. I contributed to class discussions (ex: asking questions, responding to questions, sharing reactions/opinions w/ the class).

9. I communicated with my professor during class (ex: asked questions right before or after class, asked questions during group work.

Relationship with Faculty

10. I attended faculty office hours to discuss my assignment grades.

11. I attended faculty office hours to discuss my exam/essay grades.

12. I talked to my professor(s) about my academic plans.

13. I attended faculty office hours to ask coursework related questions.

14. I talked to my professor(s) about personal issues.

15. I talked to my professor(s) about what other classes I should take in the future.

16. I communicated with my professor(s) online (ex: e-mail, canvass).

Relationship with Peers

17. My classmates shared their class notes with me.

18. I shared my class notes with my classmates.

19. I asked my classmates for help with my homework assignments.

20. I got together with my classmates to study or complete homework assignments.

21. I used my classmates as a source of information regarding school resources (ex: registration dates, financial aid, etc.).

22. I talked to my classmates about what classes to take next semester.

23. I got together with my classmates outside of class to socialize (ex: lunch, movies, sporting events).

Participation in Campus Activities

24. I am involved in a campus club or organization such as the associated body or a sports team.

25. I am involved in organizing events/activities on campus (ex: club meeting).

26. I attended campus events and/or activities.

27. I attended workshops offered on campus.

28. I attended information sessions offered on campus (ex: transfer, financial aid, library).

APPENDIX B

Instructor interview protocol

1. When I say engagement what do you think of?
2. To you what does it mean to be engaged in the classroom?
3. What does a student do when they are engaged in class? (Probe for specifics or examples)
4. Conversely, what does a student do when they are not engaged in class? (Probe for specifics or examples)
5. Do you give participation points? If what indicators do you use to assign these points?
6. Is a good student necessarily an engaged student? Or can a student be engaged in the classroom and not be doing well or not receiving the best grades?

Instructor questions for rating their students' engagement

Refer to the last 3 weeks of your STACC 100 class (Monday/Wednesday [time] or Tues/Thurs [time]) to answer the following:

1. Student submitted assignments on time (ex: free writing, essays, draft essays, annotated bibliographies, etc.)
2. Student came to class prepared (ex: had appropriate textbooks, writing materials, free writing journals, essay draft for peer review, completed assigned readings)
3. Actively participated when working in a group (ex: contributed ideas, listened to and responded to group members, completed his/her part of the group project)
4. How often did the student interact with you during & outside of class time? (ex: office hours, e-mail, canvass, right before or after class)?

5. Student contributed to class discussions (asked questions, responded to questions, shared thoughts, opinions, experiences, etc)

APPENDIX C

Student focus group protocol

1. When I say “engagement” what do you think of?

2. Can you tell if a fellow classmate is engaged in class? (Probe for specific examples)

B. If a student does not speak much in class, can they still be engaged?

3. What are some non-verbal indicators of engagement?

4. How do you engage with your peers? Do you study/school work? If so where?

B. Do you hang out with classmates outside of class time? Outside of school?

5. How do you engage with your instructors? Attend office hours?

APPENDIX D

Researcher questions for rating student engagement

1. Student attended class on a regular basis
2. Student came to class prepared (ex: had appropriate textbooks, writing materials, free writing journals, essay draft for peer review, completed assigned readings)
3. Actively participated when working in a group (ex: contributed ideas, listened to and responded to group members, completed his/her part of the group project)
4. How often did the student interact with instructor during class
5. Student contributed to class discussions (asked questions, responded to questions, shared thoughts, opinions, experiences, etc)

APPENDIX E

Observation guide – Domains

	Attendance (on time/late)	**Prepared for Class (eg. Has draft essay, reading materials, etc)**	**Asked questions in class (record the Qs being asked)**	**Group/pair participation (responding to Qs, contributing to work/discussion etc.)**	**Student-Instructor interaction**	**Larger Group Discussion**
Student A						
Student B						
Student C						
Student D						
Student E						
Student F						

APPENDIX F

Previous exploratory factor analyses results of the CCSES – Spring 2013

Item	Relation w/Faculty α =.864	Class Participation α=.733	Relation w/Classmates α=.831	Campus Activity α=.821
I attend faculty office hours to discuss my assignment grades	.784			
I attend faculty office hours to discuss my exam grades	.781			
I talk to my professor(s) about my academic plans	.730			
I attend faculty office hours to ask coursework related questions	.707			
I talk to my professor(s) about what other classes I should take in the future	.590			
I talk to my professor(s) about personal issues	.536			
I ask questions during class	.464			
I e-mail my professor(s) with any questions (ex: coursework, exams, etc.)	.430			
I keep up with assigned class readings		.644		
I take notes during class lectures		.585		
I come prepared to class (ex: paper, pen/pencil, text books)		.574		
I use information from class lectures to complete homework assignments		.560		
I complete homework assignments on time		.483		
I attend class meetings on a regular basis		.422		
I participate in group activities during class meetings		.374		
My classmates share class notes with me			.761	
I share my class notes with my classmates			.701	
I ask my classmates for help on homework assignments			.581	
I use my classmates as a source of information regarding school resources			.566	
I talk to my classmates about what classes to take in the future			.522	
I get together with classmates outside of class to socialize (ex: go to lunch, etc.)			.427	

Item	Relation w/Faculty α =.864	Class Participation α=.733	Relation w/Classmates α=.831	Campus Activity α=.821
I am involved in a campus club or organization such as the associated study body or a sports team				.856
I am involved in organizing events/activities on campus (ex: club meeting)				.829
I attend campus events and/or activities				.688
I attend workshops offered on campus				.366

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Angell, L. R. (2009). Construct validity of the community college survey of student engagement (CCSSE). *Community College Journal of Research and Practice*, *33*(7), 564-570.

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical, conceptual, and methodological issues of the construct. *Psychology in the Schools*, *45*, 369-386.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2013). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior A Review of Seven Journals. *Health Education & Behavior*, *41* (1), 12 - 18.

Brennan, R.L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17 (1),* 5 – 9.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, *47*(1), 1–32.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement*, *68*(3), 397–412.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cronbach, L. J. (1971). Test validation. *Educational Measurement*, *2*, 443–507.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, Pp. 621 - 694.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, *16*(10), 640–647.

Edgerton, H. A. (1949). The place of measuring instruments in guidance. In Willma T. Donahue, C. H. Coombs, & R. M. W. Tavers (Eds.), *The measurement of student adjustment and achievement*. Ann Arbor, Mich: University Michigan Press, Pp. 51 - 58

Esquivel, S. L. (2011). The Factorial Validity of the National Survey of Student Engagement. *Doctoral Dissertations*. Retrieved from http://trace.tennessee.edu/utk_graddiss/965

Fram, S.M. (2013). The constant comparative analysis method outside of grounded theory. *The Qualitative Report, 18(1),* 1-25.

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

Fremer, J. (2000). Promoting high standards and the "problem" with construct validation. *NCME Newsletter, 8(3),* 1.

Gordon, J., Ludlum, J., & Hoey, J. J. (2008). Validating NSSE Against Student Outcomes: Are They Related? *Research in Higher Education*, *49*(1), 19–39.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, 5, 511 - 517.

Handelsman, M., Briggs, W., Sullivan, N. & Towler, A. (2005). A measure of college student course engagement. *Journal of Educational Research, 98,* 184.

Hogan, T. P., & Agnello, J. (2004). An Empirical Study of Reporting Practices Concerning Measurement Validity. *Educational and Psychological Measurement*, *64*(5), 802–812.

Hubley, A.M., & Zumbo, B.D (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123(3),* 207-215.

Janosz, M., Archambault, I., Morizot, J., & Pagani, L. (2008). School engagement trajectories and their predictive relations to dropout. *Journal of Social Issues, 64,* 21-40.

Jimerson, S.R., Campos, E., & Greif, J. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *The California School Psychologist, 8,* 7-27.

Kane, M.T., (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M.T., (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kuh, G. D., & Ikenberry, S. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. National Institute for Learning Outcomes Assessment.

Kuh G.D., Cruce, T. M., Kinzie, J., & Gonyea, R. M. (2008) Unmasking the effects of student engagement on first year college grades and persistence. *The Journal of Higher Education, 79,* 540-563.

Laird, T., Chen, D., & Kuh, G. (2008) Classroom practices at institutions with higher-than-expected persistence rates: What student engagement data tell us. *New Directions for Teaching & Learning, 115,* 85-99.

LaNasa, S. M., Olson, E., & Alleman, N. (2007). The Impact of On-campus Student Growth on First-year Student Engagement and Success. *Research in Higher Education*, *48*(8), 941–966.

LaNasa, S. M., Cabrera, A. F., & Trangsrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education*, *50*(4), 315–332.

Lindquist, E. F. (1942). *A first course in statistics*. (Rev. ed), Boston: Houghton, Millin.

Lissitz, R., W & Samuelsen, K (2007). A suggested change in the terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437-448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory:

Monograph Supplement 9. *Psychological Reports*, *3*(3), 635–694.

Mandarino, C., & Mattern, M. Y. (2010). *Assessing the Validity of CCSSE in an Ontario College*. Higher Education Quality Council of Ontario.

Marti, C. N. (2004). Overview of the CCSSE instrument and psychometric properties. *Retrieved September*, *14*, 2004.

Marti, C. N. (2008). Dimensions of student engagement in American community colleges: Using the Community College Student Report in research and practice. *Community College Journal of Research and Practice*, *33*(1), 1–24.

McClenney, K. M., & Marti, C. N. (2006). Exploring relationships between student engagement and student outcomes in community colleges: Report on validation research. *Austin, TX: The University of Texas. Retrieved August, 19*, 2007.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955–966.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. *Test Validity*, *33*, 45.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, *23*(2), 13–23.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470-476.

National Survey of Student Engagement (NSSE, 2010c). *Benchmarks of educational practice.* Retrieved from http://nsse.iub.edu/pdf/nsse_benchmarks.pdf.

National Survey of Student Engagement (NSSE, 2010e). *Construction of NSSE benchmarks.* Retrieved from http://nsse.iub.edu/html/PsychometricPortfolio_SurveyDevelopment.cfm.

Nora, A., Crisp, G., & Matthews, C. (2011). A reconceptualization of CCSSE's benchmarks of student engagement. *The Review of Higher Education*, *35*(1), 105–130.

Pike, G. R. (2012). NSSE benchmarks and institutional outcomes: A note on the importance of considering the intended uses of a measure in validity studies. *Research in Higher Education*, 1–22.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education*, *35*(1), 45–76.

Ream, R. & Rumberger, R. (2008) Student engagement, peer social capital, and school dropout among Mexican-American and non-Latino White students. *Sociology of Education, 81,* 109 – 139.

Reyes, M.R., Brackett, M.A., Rivers, S.E., White M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 10,* 1 – 13.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric Assessment and Reporting Practices Incongruence Between Theory and Practice. *Journal of Psychoeducational Assessment*, *27*(6), 465–476.

South, S.J., Haynie, D.L., & Bose, S. (2007). Student mobility and school dropout. *Social Science Research, 36,* 68-94.