Spring 2015

The effect of examinee motivation on value-added estimates

Laura M. Williams
James Madison University

The Effect of Examinee Motivation on Value-Added Estimates

Laura M. Williams

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Assessment and Measurement

May 2015

Dedication

This dissertation is dedicated to the two most important men in my life. First, to my husband, Pete. You've been through this entire journey with me, always supported me, and put up with a lot of stress and postponing of fun. Thank you for making it possible to be a full-time student again. Now, on with our lives!

Second, to my late father, Paul. He viewed education as the most important thing he could provide for his children. Based on the way we all turned out (an engineer, a lawyer, a military pilot and now a Ph.D.), I think he succeeded. I miss him every day and wish he could be here to share this accomplishment with me.

Acknowledgements

I would like to acknowledge several people who made the work I present here possible. First, I would be remiss if I did not begin by thanking my advisor and dissertation chair, Dr. Donna Sundre. I could not imagine a better advisor than Donna. She has been my biggest cheerleader throughout my doctoral studies, and never wavered in her support as I conceptualized and wrote this dissertation. She knew when to let me marinate on my thoughts, and when to push me to take action. How she discerned the right time to do each, I will probably never figure out but will always be grateful for. Whether I needed to talk dissertation, cooking, the new tea I just tried, or simply discuss what was on my mind, Donna was there. I am forever indebted to Donna for her wisdom, mentorship, and guidance.

Second, I would like to thank the remaining members of my committee, Dr. Jeanne Horst and Dr. Monica Erbacher. Ask anyone in the Assessment and Measurement program and CARS about Jeanne and Monica and you are likely to hear how nice, helpful, and wickedly smart they both are. I appreciate all of the time you both spent with me as I grappled with some of the more technical issues in this dissertation. In particular, Monica's course on Longitudinal Data Analysis provided much of the foundation for the analyses conducted here. I can't thank either of you enough.

Third, the MOTly Crew. I was fortunate that Donna and Dr. Sara Finney invited me to be a part of this research team along with Matthew Swain and Devon Hopkins-Whetstone. We did some great research over the past three years, and I can honestly say that without the experience I gained on this research team, I don't think I would have

been inspired to take on value-added estimates as the topic of my dissertation. Thank you to the entire MOTly crew for allowing me to contribute and serving as a springboard for this dissertation.

Fourth, I would like to thank the people who I am fortunate enough to call colleagues and friends in my program. There is something about having people around who are going through an experience together. Somehow, these individuals always understood exactly how I felt and were always happy to listen to my ideas and help me work through issues. We have spent many hours together brainstorming, commiserating, and encouraging one another. I hope we continue to do so post-JMU.

Finally, I need to thank my husband and family. There were a lot of times that I had to say "no" over the last few years, because I had coursework, research, or this dissertation to work on. I am grateful for their understanding and look forward to a future filled with many more "yes's".

Table of Contents

List of Tables

List of Figures

Abstract

Questions regarding the quality of education, both in K-12 systems and higher education, are common. Methods for measuring quality in education have been developed in the past decades, with value-added estimates emerging as one of the most well-known methods. Value-added methods purport to indicate how much students learn over time as a result of their attendance at a particular school. Controversy has surrounded the algorithms used to generate value-added estimates as well as the uses of the estimates to make decisions about school and teacher quality. In higher education, most institutions used cross-sectional rather than longitudinal data to estimate value-added. In addition, much of the data used to generate value-added estimates in higher education were gathered in low-stakes testing sessions. In low-stakes contexts, examinee motivation has been shown to impact test performance. Additionally, recent empirical evidence indicated that the change in test-taking motivation between pre-and post-test was a predictor of change in performance. Because of this, researchers have suggested that test-taking motivation may bias value-added estimates. Further, if interest truly lies in measuring student learning over time, the use of cross-sectional data is problematic, since the pre- and post-test data is gathered from two different groups of students, not the same students at two time points. The current study investigated two overarching questions related to value-added estimation in higher education: 1) are different methods of value-added estimation comparable?; and 2) how does test-taking motivation impact value-added estimates? In this study, first the results from value-added estimates calculated with cross-sectional and longitudinal data were compared. Next, estimates

generated from two value-added models were compared: raw difference scores and a longitudinal hierarchical linear model. Finally, estimates were compared when motivation variables were included. Results indicated that at the institution under study, cross-sectional and longitudinal data and analyses yielded similar results and that changes in test-taking motivation between pre- and post-test did impact value-added estimates. Suggestions to combat the effect of motivation on value-added estimates included behavioral as well as statistical interventions.

CHAPTER ONE

Introduction

In past decades, the call for accountability in K-12 and higher education has increased in volume and urgency. In 2006, a U. S. Department of Education report made the call for accountability abundantly clear, making recommendations for higher education in the United States in terms of both measuring and reporting student learning (U.S. Department of Education, 2006). Specifically, the report recommended that institutions define student achievement in terms of learning and report evidence of learning in a value-added framework. Value-added refers to the change in student competency over time as a result of attending an educational institution (Astin, 1982). The value-added perspective represents a departure from traditional methods of resource and reputational measures to evaluate institutions. However, the idea of evaluating educational quality by measuring how much students learn has been around for decades (Astin, 1982; Astin & antonio, 2012).

Value-added analyses have gained traction in K-12 due to mandatory testing and reporting. In particular, the No Child Left Behind Act of 2002 set specific benchmarks as well as progress goals that schools must achieve (Braun, 2004). Although benchmarks, also known as measures of student attainment, had merit in regards to measuring student learning, some debate existed about whether benchmarks were sufficient measures of educational quality. Measures of student growth were seen as potentially better, since they not only measured students' competencies at a particular point in time, but how much students had learned over time. As a result of the shift in focus to student growth

rather than attainment as a measure of educational quality, schools turned to measures of improvement in student learning to measure quality, rather than only measures of student attainment, or competency. This led to the prevalence of value-added models in K-12 education as a way to measure school quality.

Not long after value-added models were introduced as a way to measure school quality, the results were also used to evaluate teacher quality. Although the quality of a school is due in some part to the quality of its staff, the use of value-added estimates to evaluate teacher quality is dubious. The appropriateness and validity of using value-added in such a manner has been debated in literature (Amrein-Beardsley, 2014; Braun, 2004, 2005). Value-added estimates certainly provide useful information about how much students learn over time and which students may not be progressing as much as teachers and administrators would like. However, value-added estimates did not necessarily provide information regarding school or teacher quality. To make such a claim, causal inferences regarding the effect of specific schools/teachers/etc. were implied, but the methodology surrounding data collection for value-added did not support causal inferences (Braun, 2005). To this point, the American Statistical Association (ASA) made recommendations regarding the use of value-added estimates, cautioning users to interpret estimates in the context of other measures of quality, consider the limitations of the designs, and obtain measures of precision (ASA, 2014). Along these same lines, the *Standards for Educational and Psychological Testing* (the *Standards*; AERA, APA, & NCME, 2014) also cautioned test users to take care when interpreting value-added estimates, and to be aware of "what questions each growth model can (and cannot) answer" (p. 185). The fact that value-added estimates have been used as the sole

measure of teacher quality, which in some cases was directly tied to teacher promotion and pay, is in direct contradiction to these recommendations. Such uses have caused problems, including the Chicago teacher's union strike in 2012 (Payne, 2012). Teachers were not the only ones who objected to the use of value-added estimates to make high-stakes decisions about quality. Cautions regarding using value-added estimates as the sole indicator of district, school, or teacher quality have emerged, and urged educators to use value-added estimates as only one of multiple measures of quality (American Statistical Association, 2014).

In contrast to the strict requirements in K-12 education, measuring quality in higher education has been subject to far fewer explicit requirements. Although there have been calls for increased accountability, transparency, and measures of institutional quality (U.S. Department of Education, 2006), reporting on student learning was not federally mandated. Student learning reporting mechanisms were usually a result of accreditation requirements and in some cases, state mandates. Upcoming changes to accreditation policy, however, indicate that more transparency regarding student learning in higher education is expected in the future (Wheelan, 2014). For example, the Southern Association of Colleges and Schools Commission on Colleges (SACS-COC) requires that each institution identifies

> expected outcomes, assesses the extent to which it achieves these outcomes, and provides evidence of improvement based on analysis of the results in each of the following: 3.3.1.1 educational programs, to include student learning outcomes; 3.3.1.2 administrative support services; 3.3.1.3 academic and student support services; 3.3.1.4 research within its mission, if appropriate; 3.3.1.5

> community/public service within its mission, if appropriate. (SACS-COC, 2012, p. 27)

Clearly, accreditors were interested in measures of student achievement, but they did not explicitly recommend a specific approach to estimate value-added. In contrast, the 2006 report by the Commission on the Future of Higher Education recommended very specific methods for measuring quality: institutions should not only measure student learning, they should do so in a value-added framework (U.S. Department of Education, 2006). The state of Virginia has gone so far as to recommend three specific ways to estimate and report value-added estimates of student learning (State Council of Higher Education in Virginia, 2007). Specifically, they recommend: 1) longitudinal repeated-measures designs; 2) cross-sectional designs; or 3) residual analyses. All three of these designs have advantages and disadvantages which will be explored in the later chapters of this dissertation.

In response to the call for evidence of quality in higher education, the Lumina Foundation funded development of a framework for accountability in higher education (VSA, 2011). With this support, leaders in the National Association of State Universities and Land-Grant Colleges (NASULGC, now the Association of Public and Land-Grant Universities, or APLU) and the American Association of State College and Universities (AASCU) set out to establish a common set of accountability measures for institutions of higher education (McPherson & Shulenburger, 2006; Shulenburger, 2007). These measures were intended to provide accountability to three main audiences: students and the general public; faculty and staff on campuses; and higher education stakeholders, including policymakers. Specifically, the measures included consumer data, student

engagement data, and student learning outcomes data. Consumer data, intended to be "helpful to prospective students and their parents in deciding which university best fits their educational wants and needs" (McPherson & Shulenburger, 2006, p. 8), was related to the cost of college, retention rates and graduation rates. Student engagement, or campus climate, measures were intended to provide information to students and parents regarding the campus environment. Suggested measures in this area included survey data from instruments such as the National Survey of Student Engagement (NSSE) and reported data such as student satisfaction, opportunities for active learning, and experiences with diverse groups of people (Shulenburger, 2007). Finally, NASULGC and AASCU recommended that the accountability system include measures of student learning outcomes, particularly in areas thought to cut across disciplines and institutions, such as critical thinking and communication skills.

As a result of these conversations and recommendations, the Voluntary System of Accountability (VSA) formed for the purpose of providing transparency to stakeholders and measuring core educational outcomes in higher education; these core outcomes were defined as written communication, critical thinking and analytical reasoning (Liu, 2009; VSA, 2014). The VSA chose these skills in particular because they are likely common across disciplines and universities (Liu, 2009). Institutions who chose to participate in the VSA were required to provide information intended to make higher education more transparent to the general public; this information was subsequently published on the College Portrait website (http://www.collegeportraits.org/). Reporting requirements for VSA members included: net price of attendance (i.e., total cost of attendance minus scholarships and grants), a description of the campus community, campus safety

statistics, student success and progress rates, institution-specific learning outcomes data, and value-added results from one of three assessment instruments: the Collegiate Assessment of Academic Proficiency (CAAP), Collegiate Learning Assessment (CLA), ETS Proficiency Profile (VSA, 2008). Later, AAC&U Value Rubrics were added to the list of acceptable assessment instruments, thus giving institutions a choice of four instruments (VSA, 2014). Membership in the VSA is completely voluntary but has been declining since its founding in 2006. Although initially free to participating institutions, in 2010 member institutions began paying dues ranging from $500-$2500, depending on the institution's student enrollment (VSA, 2011). Due to its voluntary nature and decline in membership, the VSA did not encompass a wide representation of higher education. However, the VSA represented the only widely-implemented model of value-added in higher education, and thus its methodology warrants investigation. Concerns about value-added in K-12 have trickled into higher education applications; a premise of the proposed study is the inattention to and lack of research regarding validity threats in value-added frameworks.

As in K-12, the testing contexts used to measure quality in higher education, whether in value-added or other frameworks, were primarily low-stakes for students. That is, test results did not impact students either positively or negatively. In contrast, high-stakes test results impacted students personally, such as scores on a placement exam or the SAT. Despite the low-stakes nature of testing for students, the test results were often used by schools and institutions to make high-stakes decisions, such as those regarding teacher pay, resources, or curriculum changes. At first glance, this may not seem like an issue, but literature has established a relationship between testing

performance and test stakes (DeMars, 2000; Liu, Bridgeman & Adler, 2012; Sundre & Kitsantas, 2004; Wolf, Smith & Birnbaum, 1995). That is, in low-stakes testing, students were not as motivated as they were in high-stakes contexts. Further, motivation and performance were positively correlated in low-stakes testing: the more motivated a student was, the better he or she performed on the test (Sundre & Kitsantas, 2004). More recent research has indicated that *change* in examinee motivation over time was a predictor of change in performance (Finney, Sundre, Swain, & Williams, under review). Because value-added implied learning *change* over time, the fact that change in examinee motivation was related to change in performance suggested value-added estimates may be biased. It follows, then, that the results from scores obtained in low-stakes testing contexts at single administrations and value-added estimates may not truly reflect how much students have learned. These findings have major implications for institutions and the decisions based on scores obtained in low-stakes testing, particularly value-added scores. For example, if value-added estimates were confounded with motivation, it could be that the estimates were inaccurate. Thus, any high-stakes decisions, like passing and failing students in a class or decisions about program curricula, may also be incorrect when based on these estimates. The impact of motivation on value-added estimates must be investigated.

In addition to the concerns with the impact change in motivation over time had on value-added estimates, methodology was also a concern. Most institutions used a cross-sectional design to collect data for value-added estimation purposes. Yet, institutions used this data to make inferences regarding student growth and development—inferences that were not supported. Research indicated that in order to make inferences regarding

growth, longitudinal data is imperative (Castellano & Ho, 2013; Singer & Willet, 2003). Although cross-sectional designs can support inferences regarding *differences between two groups of students* (i.e., freshmen and seniors), it cannot support inferences regarding the *change in students over time*. Thus, when institutions made claims based on cross-sectional data regarding how much their students have learned as a result of attending the institution, those inferences were simply not accurate. Further, as discussed in the previous paragraph, the change in motivation over time has been shown to be a predictor of change in performance (Finney et al., under review). Modeling this change and its impact on value-added estimates is impossible with a cross-sectional design; only longitudinal data is sufficient for this purpose. For these reasons, research is needed to investigate the comparability of value-added estimates when cross-sectional vs. longitudinal data is used.

Based on the growing empirical evidence regarding the relationship between test stakes, examinee motivation, and performance, and concerns regarding data collection methodology, it is imperative to investigate the effects of examinee motivation within multiple value-added frameworks. The proposed study was designed to investigate comparability of cross-sectional and longitudinal data collection designs, compare two value-added methodologies used in higher education, and to explore whether value-added estimates systematically differ when examinee motivation is or is not included in the model. Specifically, this study first examined whether analyses conducted on cross-sectional and longitudinal samples yield the same results. Next, the study explored whether value-added estimates obtained from raw difference scores and a hierarchical linear model (HLM) provide similar value-added estimates. Finally, the comparability of

raw difference scores and HLM methods of value-added estimation when motivation variables are included in the HLM was investigated. In light of the findings from the study, recommendations for both assessment practice and policy based were made.

CHAPTER TWO

Literature Review

Turn on the news or open a newspaper, and chances are there will be a story or two that discusses the quality of education in the United States. The last decade has seen increased calls for accountability at all levels of education, with the Commission on the Future of Higher Education report (U.S. Department of Education, 2006) making this explicitly clear, stating:

> students increasingly care little about the distinctions that sometimes preoccupy the academic establishment, from whether a college has for-profit or nonprofit status to whether its classes are offered online or in brick-and-mortar buildings. Instead, they care—as we do—about results. (p. xi).

In other words, the definition of quality was shifting from what assets and resources a school might have to how much students actually learned as a result of attending that school. Yet even as students, parents, and other stakeholders insisted that institutions provide evidence of results, institutions continued to entice students by building not just state-of-the-art educational facilities, but also sexy campus amenities: modern recreation facilities, luxury on-campus housing, and gourmet food service. In an age of increasing accountability and calls for evidence of quality, should schools focus their energy on such niceties?

This chapter begins with a short definition and discussion of the term "value-added" and its application in educational contexts. Although I will briefly discuss application in K-12, the majority of this literature review focuses on value-added as it is

currently conceptualized in higher education. The tension between value-added for assessment purposes and value-added for accountability purposes will be explored, as well as benefits and drawbacks of each purpose. Then, I will turn toward value-added models currently used in higher education, including a discussion of advantages and shortcomings of current methodologies. A consistent shortcoming (regardless of methodology) in the current research on value-added in higher education involves the lack of information regarding the potential impact of examinee motivation on value-added estimates. To this end, current research regarding examinee motivation and its impact on test performance in low-stakes contexts will follow. Finally, I will discuss the purpose of this study and the research questions it addresses.

What is Value-Added?

Although the idea of measuring how much students learn as a measure of quality rather than reputation or resources may sound revolutionary to some, it was not a new idea. Indeed, the idea of value-added as a metric for institutional quality has been discussed for over 30 years (Astin, 1982; Astin & antonio, 2012). In the most general sense, value-added was defined as the impact that a particular school had on student outcomes. Astin (1982) argued that it would not make sense "to conclude that a given manufacturing company was an 'excellent' business just because it had higher-paid employees than its competitors or because it spent more money than its competitors," (p. 9) and it makes little sense to treat measures of educational quality in this way either. An alternative to this resource-based approach for evaluating quality could be outcomes-based: what students know, think, or can do upon completion of their time at an institution (Astin, 1982). At first glance, the idea of an outcomes-based approach may

seem to be a useful way to measure the quality of education. Although the outcomes-based approach was certainly an improvement over simply considering reputation or resources of an institution, it lacked a key component: that of measuring growth. In other words, the outcome view did not take into account *how much* a student learned, only that he or she could perform at a particular level upon graduation. Depending on a student's prior knowledge, he or she could potentially not learn anything at all, yet still perform at acceptable levels for an outcomes-based approach to measuring quality.

Rather than the outcomes approach to measuring quality, Astin (1982) argued that the quality of education should be measured by what students learn while they are at an institution. In other words, quality should be defined in terms of what the value added is as a result of a student's attendance at an institution. In contrast to the outcomes-based approach, the metric of quality was not about performance at the end of an educational experience but about everything that happens in between—what Astin (1982) called educational impact, or value-added. Using this definition, Astin (1982) asserted that the highest-quality institution were those with the greatest educational impact.

There are two key concepts worth highlighting regarding Astin's definition of value-added: 1) measures of quality were about the institution as a whole; and 2) longitudinal data were necessary in order to measure value-added. Regarding the first key concept, although the unit of analysis is the institution, Astin was very clear that value-added measures should primarily be used for internal improvement purposes. That is, "the value-added approach permits institutions to attain high levels of excellence without regard to what other institutions accomplish" (Astin, 1982, p. 15). Although value-added measures could be used for institutional comparison purposes, comparisons

should focus on the "degree of improvement in student performance that occurs at individual schools and colleges" (Astin, 1982, p. 15).

The second key concept, that longitudinal data was necessary for value-added, was critical to Astin's definition of value-added. If the purpose of value-added was to measure improvement of student performance, a baseline against which to measure knowledge at completion was crucial (Astin, 1982). Although cross-sectional data could provide information regarding the differences between two independent groups of students, it did not provide information regarding change in time, or growth, in those students. Thus, longitudinal data were the only data that could suffice for the purpose of measuring growth. Further, if value-added estimates were intended to also be used for institutional improvement purposes, collection of longitudinal data allowed institutions to also examine the impact of curricular changes. For example, an institution may have noticed that students' writing skills were not improving as much as the institution would like. As a result, the institution may have modified its writing curriculum in hopes of improving students' writing skills more than what was accomplished under the previous curriculum. Through longitudinal measurement, the institution could then compare whether the writing proficiency of students experiencing the new curriculum improved more than the writing proficiency of students participating in previous curricula. In contrast, cross-sectional data could not capture these types of changes because the same students are not measured at pre- and post-test. At best, cross-sectional data could provide information about whether students' writing proficiency was better at completion of the new curriculum compared to students who had not experienced any writing instruction, but cross-sectional data could not provide information about how much

students' writing proficiencies actually *improved*. For most accountability reporting purposes, the difference in performance between freshmen and seniors was sufficient, but cross-sectional methods of measuring student learning did not provide stakeholders with information regarding the *change* in student performance over time. In fact, results from cross-sectional and longitudinal data may not be comparable, particularly if the interest lies in examining student growth (Liu, 2011a, 2011b; Castellano & Ho, 2013).

Although value-added was originally intended to refer to learning gain, growth, or change in students as a result of attending school, the meaning of value-added evolved over the past decades. Of exceptional note was that in K-12 education, value-added has often been translated into a high-stakes public judgment of teacher and/or school effectiveness (Amrein-Beardsley, 2008, 2014). Value-added as a means to compare institutional quality was an appropriate application; indeed, Astin (1982) mentioned this as a secondary use of value-added estimates. However, value-added estimates as a measure of teacher effectiveness was less tenable. Value-added measures were best used for purposes of programmatic or institutional improvement of student learning: "Differences in 'pretest' and 'postttest' performance would provide students, teachers, and school officials with critical feedback on the nature and extent of student growth and development" (Astin, 1982, p. 16). Value-added can and should be used to *inform teaching practice*, but not *measure teacher quality*. Making statements about teacher quality based on value-added estimates implied a causal statement, which the methodology of K-12 value-added models did not support (Braun, 2005).

Even though value-added in K-12 may have gotten off-track in terms of how scores are used, K-12 systems have done a good job of collecting longitudinal data to

measure student growth. Most value-added methods in K-12 (e.g., growth curve models and multivariate models, among others) required longitudinal data as an assumption of the statistical model; districts tended to use standardized test scores for this purpose (Kim & Lalancette, 2013). In this respect, at least, K-12 systems are following Astin's advice: longitudinal data collection allowed schools to use value-added information about student learning to improve curriculum and teaching practices.

In contrast, higher education had very little structure for measuring value-added. Whereas K-12 had very clear policies and procedures for collecting, analyzing, and using value-added estimates, higher education had very little. Few institutions gathered longitudinal data and measured student gains over time; the most widely-used value-added models in higher education employed cross-sectional designs (Kim & Lalancette, 2013; Liu, 2011a, 2011b; Steedle, 2012). Although individual *programs* (i.e., academic majors, student affairs programs, etc.) at an institution may have collected longitudinal value-added data, institutions themselves generally collected data regarding student outcomes. Outcomes data refers to data collected that reflects what students know, think, or can do at the conclusion of their academic career rather than data that illustrated how students have changed between entry and exit. Even when outcomes data was collected, it was usually in response to accreditation demands and rarely used for institutional improvement. In this era of increased accountability, higher education would do well to gather quality student learning data, reconsider the use of cross-sectional models, and investigate methods to implement longitudinal measures of value-added. Doing so would provide institutions with actual measures of student improvement, which could be used for both institutional improvement purposes as well as external accountability purposes.

There is a wealth of literature debating value-added models and their applications; however, this literature review will specifically focus on value-added models used in higher education. Current research on value-added has mostly been conducted in the K-12 environment with very little focus on higher education. Further, research regarding higher-education value-added models has relied on cross-sectional data indicating a clear need for more intentionality and research in value-added models in higher education.

Value-Added in Higher Education

Assessment for improvement or accountability? As discussed earlier, Astin (1982) proposed value-added as a way to measure quality in terms of student growth and development. Using Astin's definition, assessment data should be used for improvement purposes and results should inform program changes that hopefully enhanced the learning of future student cohorts. In other words, the primary purpose of assessment was not to demonstrate to external stakeholders how well a particular institution was doing, but the purpose was to continually improve student learning. When assessing student learning, institutions were often attempting to answer the question: how are we doing? However, if any given institution were to try to answer that question, inevitably they might answer with a question of their own: compared to what? (Miller, 2012). The answer to this question depends on whether the question was being asked by institutions themselves, or external stakeholders. Assessment for improvement purposes was typically driven by an internal desire to improve—it could be compared to athletes who tried to run a faster mile to achieve their own goals (improvement) versus athletes who tred to run a faster mile because their coaches told them they had to (accountability). In assessment for improvement, the answer to "Compared to what?" was "Ourselves over time." If

"ourselves over time" was the answer, the question was best answered with longitudinal data collection. In contrast, the answer to the same question when conducting assessment for accountability purpose was "Other institutions" or "A particular benchmark". Just as faculty (and coaches) would like students to learn for learning's sake, institutions should conduct assessment for their own formative purposes. If done well, however, the data gathered when conducting assessment for improvement could also be used for accountability purposes.

Nevertheless, increased calls for evidence of institutional quality were making it necessary for institutions to provide proof of student performance for accountability purposes. The Commission on the Future of Higher Education report (U.S. Department of Education, 2006) made this explicitly clear in its recommendations regarding a change "from a system primarily based on reputation to one based on performance" (p. 21):

> ***Postsecondary education institutions should measure and report meaningful student learning outcomes.***
>
> - Higher education institutions should measure student learning using quality assessment data from instruments such as, for example, the Collegiate Learning Assessment, which measures the growth of student learning taking place in colleges, and the Measure of Academic Proficiency and Progress, which is designed to assess general education outcomes for undergraduates in order to improve the quality of instruction and learning.…
> - Faculty must be at the forefront of defining educational objectives for students and developing meaningful, evidence-based measures of their progress toward those goals.

- The results of student learning assessments, including value-added measurements that indicate how students' skills have improved over time, should be made available to students and reported in the aggregate publicly. Higher education institutions should make aggregate summary results of all postsecondary learning measures, e.g., test scores, certification and licensure attainment, time to degree, graduation rates, and other relevant measures, publicly available in a consumer-friendly form as a condition of accreditation.
- The collection of data from public institutions allowing meaningful interstate comparison of student learning should be encouraged and implemented in all states. By using assessments of adult literacy, licensure, graduate and professional school exams, and specially administered tests of general intellectual skills, state policymakers can make valid interstate comparisons of student learning and identify shortcomings as well as best practices. The federal government should provide financial support for this initiative. (U.S. Department of Education, 2006, p. 24)

At first glance, these recommendations appeared to be a positive shift in terms of measuring educational quality—the preference was for measuring learning, not resources or reputations. It is clear from these recommendations that the call for accountability was moving toward value-added frameworks, and examining the change in student learning over time. This was excellent news, as institutions could conceivably conduct assessment for both improvement and accountability purposes simultaneously. However, reading the recommendations closely, this was not necessarily the case. Institutions and faculty were

encouraged to advance learning objectives that were meaningful at their own institutions, for their own students. In addition, institutions were asked to develop direct measures of those learning objectives. Yet the report also recommended specific instruments that institutions should use to measure student growth, an apparent contradiction to developing institution-specific measures to capture student learning. Further, the report called for institutions to publicly report aggregate assessment results so that states and the general public could make "meaningful interstate comparison(s) of student learning" (U.S. Department of Education, 2006, p. 24). In other words, higher education has been prompted to adopt value-added interpretations similar to those currently used in K-12: comparison of institutional quality. As a result, value-added models have been applied to higher education in response to this call for accountability. The next section of this literature review will discuss commonly used value-added approaches in higher education.

Models for higher education. Although many methods of estimating value-added existed across the entire educational system, this literature review will concentrate on those used in higher education: difference scores and residual gain scores. Although some would debate whether difference scores are indeed a way of estimating value-added, Astin's (1982; Astin & antonio, 2012) definition implied that the difference in student performance upon entry and completion is an indicator of value-added institutional impact. In fact, the State Council of Higher Education in Virginia (SCHEV) advocated longitudinal repeated-measures designs (i.e., raw difference scores) as one of the acceptable methods of reporting value-added for institutional assessment and

accountability purposes (SCHEV, 2007). Residual gain scores were another method approved by SCHEV; both of these will be discussed in detail next.

Difference scores. Difference scores are known by several names, including gain scores, simple gain scores, or growth relative to self. These scores are exceptionally straightforward in their definition: the difference in performance over time. The most basic value-added model, and arguably the foundation for most other types of value-added models, was the raw difference score (i.e. gain scores; Castellano & Ho, 2013). Indeed, difference scores are what Astin (1982) and Astin and antonio (2012) seem to be describing when they discussed value-added. Difference scores answered the question "How much has a student (or group of students) learned on an absolute scale?" (Castellano & Ho, 2013). As the reader might guess, the calculation of a difference score was represented as (Castellano & Ho, 2013, p. 36):

Difference score = current status – initial status

In other words, subtract the scores at the first time point from the scores at the last time point. In order to calculate meaningful difference scores, at least two time points were needed for each student, and those scores must be on the same scale. Vertical scaling methods were acceptable for this purpose (Castellano & Ho, 2013). Vertical scaling links scores on two or more tests that measured the same construct across academic levels; research has shown that "while the specific choice of linking approach may not lead to dramatically different value-added effect estimates, the choice not to link the tests at all almost certainly will" (Briggs & Weeks, 2009, p. 408).

In any discussion of the use of difference scores to describe change over time, it would be remiss to overlook the decades-long psychometric debate regarding the

reliability of difference scores. Cronbach and Furby (1970) are perhaps the most well-known opponents of using difference scores to measure change, stating that "gain scores are rarely useful, no matter how they may be adjusted or refined" (p. 68). In short, difference scores were unreliable due to their correlation with error (Cronbach & Furby, 1970). Classical test theory assumed that measurement error was uncorrelated with true score, yet difference scores were related to measurement error because at each time point, the measures were completed by the same person. Further, it was assumed that variances in true scores at each testing occasion were equivalent; if this were true, the reliability of difference scores would be less than the reliability of scores on either the pre- or post-test. For these reasons, using difference scores to describe change was discouraged, and alternative approaches were suggested. Of the suggested alternatives, only residual gain scores have been received well and used to estimate change. Residual gain scores were a method of estimating change by comparing actual performance to predicted performance, with predicted performance estimated via regression-based techniques. More thorough discussion of residual gain scores will emerge later in this literature review as one of the current methods to estimate value-added in higher education.

In the years since the criticism of raw difference scores as a method to measure change was published, other researchers have examined the argument further and concluded that only under a very strict set of circumstances were difference scores unreliable (Fulcher & Willse, 2007; Nesselroade, Stigler & Baltes, 1980; Williams & Zimmerman, 1996). Difference scores were *un*reliable only if *all* of the following conditions were met:

1) The correlation between observed scores at time 1 and observed scores at time 2 was large and positive;
2) The variance in observed scores at time 1 and time 2 was equal;
3) The variance in true scores at time 1 and time 2 was equal;
4) The reliability of the scores at each occasion was equal;
5) The correlation between true scores and true change scores was negative.

Review of these conditions makes it clear that in any realistic testing condition, at least one, if not multiple, of these conditions will not be met. Consider a situation where students were tested prior to and after a particular curriculum intervention—let us assume we were testing scientific reasoning skills. At pre-test, students presumably did not have much knowledge and/or skill in scientific reasoning. As a result, scores were likely on the low end, and there may be a good deal of variability in scores (e.g., large variance) due to guessing on items to which students do not know the answer, varying levels of student knowledge, and differing educational experiences related to scientific reasoning. Students then experienced the intended curriculum to gain scientific reasoning knowledge and skills, and were tested again. Assuming the curriculum was effective, student scores likely increased. Moreover, the variability of scores was likely be less than at pre-test, because presumably students guessed less than at pre-test, were more similar in their knowledge, and had more skill in scientific reasoning. In cases such as this (which are common when measuring student learning), it is easy to not meet condition 2 (variance in observed scores at time 1 and time 2 are equal), resulting in a case where difference scores would be reliable. Remember, *all five* conditions listed above must be met for

difference scores to be unreliable; in this example, only four of the conditions would be met, meaning that difference scores would be reliable. Further, if variances of pre- and post-test scores were unequal, reliability was likely unequal at both testing conditions as well, in which case condition 4 (reliability of the scores at each occasion is equal) is untenable. In short, not only is it unlikely that all five conditions for unreliable difference scores will be met, but it is straightforward to determine whether or not the conditions were met. As a result, current thinking is that difference scores are an acceptable way to measure student learning over time; in other words, they are an acceptable value-added model.

Advantages of differences scores: There were three main advantages to using difference scores as a value-added model. First, they were straightforward in their calculations. As long as one had scores and unique identifiers for each student on at least two testing occasions, difference scores could be calculated. Further, if each student had taken the same test and had a school identifier, difference scores for each institution could be computed as well and compared to one another. However, given that few institutions use the same measures of student learning, using difference scores for institutional comparison is a very unlikely scenario.

Second, difference scores were easy to understand. As Castellano and Ho (2013) pointed out, "the gain score model aligns closely with intuitive notions of growth" (p. 42); a difference score can easily communicate to a stakeholder whether or not performance has changed over time and by what magnitude. Further, because the difference scores were expressed on the same scale as the test itself, a stakeholder could also make a judgment about whether the difference was meaningful or not.

Finally, difference scores can be calculated using both longitudinal and cross-sectional data, making the data collection design more flexible. As described earlier, longitudinal difference scores were calculated by averaging all of the individual difference scores in the sample. A cross-sectional difference score is calculated by subtracting the mean "pre-test" score (e.g., mean test score for freshmen) from the mean "post-test" score (e.g., mean test score for seniors). It is often costly and time-consuming to collect longitudinal data, thus cross-sectional data can be an appealing alternative. If cross-sectional sampling is chosen, however, inferences about student growth cannot be made since the difference scores are calculated as the difference between the average performances of the two *groups*, not the average difference score of *individual students*. When difference scores are calculated using a cross-sectional sample of students, only statements about the difference in performances of two groups are supported. If the question of interest is about student growth, then longitudinal data is required.

Disadvantages of difference scores. Although difference scores have the advantage of simplicity and parsimony, there are two main disadvantages to using difference scores as a measure of value-added: 1) difference scores cannot model non-linear growth; and 2) difference scores cannot control for other variables. First, because difference scores are calculated by simply comparing current status to initial status, difference scores do not give any information about the pattern of growth (linear, discontinuous, etc.; Anderman, Gimbert, O'Connell & Riegel, 2014; Singer & Willett, 2003; Rogosa, Brandt, & Zimowski, 1982). Growth in general is not assumed to be linear (Singer & Willett, 2003), and it is likely that student learning follows a nonlinear pattern as well.

Second, difference scores cannot control for other variables that may affect change in student learning over time. This was especially problematic given that previous research has shown both student- and institutional-level variables can influence value-added estimates (Cunha & Miller, 2014; Liu, 2011a, 2011b; Steedle, 2012). Indeed, in K-12 contexts in particular, the call to control for variables known to influence student learning estimates, such as socioeconomic status, was especially loud. Alternative methods of calculating value-added were developed as a response to this drawback of difference scores. Particular to higher education, residual gain scores have been commonly used for this purpose.

Residual gain scores. When using the residual gain score method of calculating value-added, the basic purpose was to compare actual and expected performance; linear regression was usually applied to calculate expected performance based on past performance and/or predictors of current performance (Castellano & Ho, 2013). As the name implies, residual gain scores are the residuals that result from the regression of post-test scores on pre-test scores, controlling for other variables in the model. The residual gain score method of quantifying growth came about in response to concerns about the reliability of difference scores. Although difference scores were usually reliable (Fulcher & Willse, 2007; Nesselroade et al, 1980; Williams & Zimmerman, 1996), using residual gain scores allowed researchers to include predictors in models of value-added estimates. Further, because predicted scores can be generated either through predicting current performance based on previous performance or other predictors, residual gain scores can be used with either longitudinal or cross-sectional data (Castellano & Ho, 2013). When using longitudinal data, gain scores can serve as the

dependent variable, and expected change in performance is predicted by variables thought to influence change in performance. When using cross-sectional data, however, the institution was the unit of analysis and thus mean performance of the group of students who were included at the second time point was usually the dependent variable. Performance at the first time point was then used to predict expected performance for the second time point. Other variables thought to influence performance could also be included in the model to calculate expected performance.

In higher education, two types of residual gain score calculations have been commonly used: Ordinary Least Squares (OLS) regression and hierarchical linear modeling (HLM). More detail on the mechanics of residual gain score calculations in both OLS and HLM frameworks follow.

Ordinary least squares (OLS) regression residual gain scores. In the OLS framework for calculating value-added, OLS methods are used to predict current performance. Perhaps one of the more well-known OLS residual gain score value-added methods is that used by the Voluntary System of Accountability (VSA). As described in Chapter 1 of this dissertation, the VSA formed to evaluate core educational outcomes common in public institutions. The OLS method used by the VSA provided value-added estimates to institutions that allowed them to compare their performance to other VSA institutions.

To estimate value-added for VSA institutions, OLS regression was used to analyze data from cross-sectional student samples from each institution: samples of freshmen and seniors (Liu, 2009). Residual scores were calculated for both freshmen and seniors; a residual score is the difference between the score predicted by the OLS

regression equation and the observed score for a particular school. The difference between the residuals of freshmen and seniors was the value-added estimate for a school. Specifically, the following formula was used to calculate the predicted performance for each group (freshmen and seniors) in a school, controlling for the average SAT score at any given school:

$$Y = \beta_0 + \beta_1(\overline{SAT}) + e \quad (1)$$

where:

Y: mean ETS Proficiency Profile/CLA/CAAP score for a school

β_0: mean of ETS Proficiency Profile/CLA/CAAP mean scores across all schools

β_1: slope for the predictor of students' mean SAT scores for a school

e: the residual for a school. It can be calculated by subtracting the predicted score for each school from the observed score for that school.

Once the residuals were calculated for the group of freshmen and the group of seniors in each school, the residual gain score was calculated by subtracting the senior residual from the first-year residual for each school. If a school had a positive residual gain, it meant they were performing better than expected; alternatively, a negative residual indicated that a school was performing worse than expected. These residual gain scores were then used to categorize schools into performance groups, reflected in Table 1 (ETS, 2008; Liu, 2011a). Such results can certainly be considered high stakes for institutions.

Table 1
Value-Added Performance Categories

Category	Performance Indicator
Well Above Expectations	> 2SD above mean
Above Expectations	1SD - 2 SD above mean
At Expectations	±1 SD of the mean
Below Expectations	1SD -2 SD below mean
Well Below Expectations	> 2SD below mean

The OLS method of calculating residual gain scores, which was used by the VSA, was fairly straightforward, but it did have two major shortcomings: 1) OLS methods only used school-level data and ignored student-level information; and 2) some assumptions of OLS methods may be violated due to the nested data structure. First, the OLS method only analyzed school-level data, and did not take student-level variables into account. Using the school as the unit of analysis amounted to essentially ignoring information at the individual student level (Liu, 2011b).

Second, some assumptions of OLS estimation methods may be violated when using data from multiple institutions. OLS regression makes several data assumptions, some of which are independence of residuals, normality of residuals, and homoscedasticity of residual variance (Cohen, Cohen, West & Aiken, 2003). If violated, these assumptions can lead to biased standard errors, thus leading to biased significance tests. One common reason for violation of these assumptions is nesting of data. That is, data observations were related to one another because of a factor they have in common; this data feature is often referred to as "nesting". It is reasonable to think that students who attend the same university may have similar responses to a test due to similarities in their education and backgrounds, thus students are nested within universities. If these similarities were exhibited, independence and normality of residuals may be violated,

thus leading to biased standard errors and inaccurate significance tests. For example, a particular variable in a model might appear to significantly predict an outcome not because it actually does predict the outcome, but because significance tests were biased due to ignoring nesting in the data. Techniques have been developed to address the issues brought about by nested data, one of which was hierarchical linear modeling (HLM). For this reason, the VSA abandoned OLS residual gain scores in favor of HLM residual gain scores in 2009, as described in the next section (Steedle, 2012).

Hierarchical linear modeling (HLM) residual gain scores. Hierarchical linear modeling (also known as multilevel modeling, mixed-effects models, random-coefficient models, and covariance components models) was developed to accommodate nested data, such as data generated when studying organizational effects or growth (Raudenbush & Bryk, 2002). As mentioned earlier, using OLS to analyze nested data can create issues in the form of biased standard errors and significance tests. By using HLM methods, variance at the individual level was separated from variance at the group level, thus allowing analysis of nested data without concern for biased standard errors or significance tests due to dependency in the data. Because HLM was an extension of linear regression, it could be used in the residual gain score framework to estimate value-added and was proposed as an alternative to OLS value-added estimation. Prior to adoption of HLM as the preferred method for value-added calculation, consideration of OLS and HLM method comparability needed to be addressed. Fortunately, several researchers explored the comparability of methods; these studies and their conclusions are described next.

To compare whether OLS and HLM methods of value-added resulted in similar estimates, Liu (2011b) compared results from both OLS and HLM methods using a sample of 23 higher-education institutions; the outcome measure in the study was the ETS Proficiency Profile. Both the OLS and the HLM methods used the basic framework described above in the discussion of OLS residual gain scores, which was used by the VSA. A residual for freshmen and a residual for seniors was calculated using each method, and the difference between the two residuals was the value-added estimate for each school. The OLS equation used mean SAT to predict ETS Proficiency Profile scores; see Equation 1.

In contrast to the OLS equation, the HLM method used to calculate value-added was represented by the following equation (Liu, 2011b):

$$Y_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij}) + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(selectivity)_j + \gamma_{02}(DGI)_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} \quad (2)$$

$$Y_{ij} = \gamma_{00} + \gamma_{01}(selectivity)_j + \gamma_{02}(DGI)_j + \gamma_{10}(SAT_{ij}) + r_{ij} + u_{0j}$$

where:

Y_{ij}: ETS Proficiency Profile score for student *i* in school *j*

β_{0j} is the mean ETS Proficiency Profile score for all students in school *j*

β_{1j}: slope for the predictor of student SAT score (constrained to be the same across all schools)

γ_{00}: average ETS Proficiency Profile score across all schools

γ_{01}: change in school mean score for one unit change in schools' selectivity (indicated by a percentage of students admitted among all applicants)

γ_{02}: difference in school means for degree-granting status (0 = undergraduate only; 1 = graduate programs)

γ_{10}: slope for the predictor of student SAT score (constrained to be the same across all schools)

r_{ij}: student-level residual for student *i* in school *j*

u_{0j}: difference between overall mean score for school *j* and grand mean, after controlling for selectivity and degree-granting status. This is the value-added estimate for each school.

Once value-added estimates were generated for both the OLS and HLM methods, they were used to assign decile groups to each school. Institutions were then ranked according to their decile groups. The rankings obtained from the OLS method were compared to the HLM rankings by correlating the two sets of rankings to determine the similarity of the results. The correlation between the two sets of results was .76 for critical thinking and .84 for writing. Although the correlations suggest similarity across the two methods, the results were far from identical; some schools' decile group rankings changed by as much as 5 groups! When discussing study limitations, it was mentioned that only one student-level predictor was included in the HLM model: SAT scores (Liu, 2011b). The authors suggested that other viable student-level predictors could be field of study, gender, and student motivation. Motivation is of special interest to the current study as it has consistently predicted performance in low-stakes testing. This topic will reemerge in a later section of this literature review.

One shortcoming of this research was that there were no definite conclusions as to whether OLS or HLM value-added estimates were more accurate, since real rather than

simulated data was used in the study. However, given the nested nature of the data, it is likely that the HLM results were more appropriate and trustworthy. The intraclass correlation coefficient (ICC) is an indicator of the variance in the dependent variable due to nesting of data. Intraclass correlation coefficients (ICC's) in the unconditional HLM model for both writing and critical thinking ranged from .09- .15 for both freshmen and seniors, indicating that anywhere from 9-15% of variance in scores was due to nesting (Liu, 2011b). This amount of variability does not seem trivial; in fact, depending on the size of the sample in each school, the Type I error rate could be inflated up to .43 or more with ICC's ranging from .09 - .15 (Kreft & DeLeeuw, 1998; Musca, Kamiejski, Nugier, Mèot, Er-Rafiy & Brauer, 2011). Therefore, HLM was preferred over OLS when estimating value-added using residual gain scores; recall that HLM methods for calculating value-added in the VSA framework were implemented in 2009 (Steedle, 2012).

Other researchers also investigated the comparability of OLS and HLM value-added estimates, paying particular attention to issues of reliability, consistency, and precision of value-added estimates (Steedle, 2012). Steedle used Liu's (2011b) HLM as a starting point, but rather than controlling for SAT scores at the individual level only, SAT scores were controlled for at the group level as well, shown in the following equation (Steedle, 2012):

$$
\begin{aligned}
&Y_{ij} = \beta_{0j} + \beta_{1j}(EAA_{ij} - \overline{EAA}_j) + r_{ij} \\
&\beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{EAA}_j) + \gamma_{02}(\overline{Test}_{fr,j}) + u_{0j} \\
&\beta_{1j} = \gamma_{10}
\end{aligned}
\qquad (3)
$$

$$
Y_{ij} = \gamma_{00} + \gamma_{01}(\overline{EAA}_j) + \gamma_{02}(Test_{fr,j}) + \gamma_{10}(EAA_{ij} - \overline{EAA}_j) + u_{0j} + r_{ij}
$$

Where:

Y_{ij}: the predicted score for seniors

EAA_{ij}: entering academic ability of student *i* at school *j* (measured with SAT scores)

$\overline{EAA}_j$: average entering academic ability of school *j*

$\beta_0 j$: student-level intercept (equal to mean test score at school *j*)

$\beta_1 j$: student-level slope coefficient for EAA at school *j* (assumed to be the same across all schools)

γ_{00}: school-level value-added equation intercept

γ_{01}: school-level value-added equation slope coefficient for senior mean EAA

γ_{02}: school-level value-added equation slope coefficient for freshman mean test score

γ_{10}: student-level value-added equation slope coefficient for EAA (assumed to be same across schools)

r_{ij}: residual for student *i* at school *j*. σ^2 is the variance of the residuals, which is the pooled within-school variance of test scores after controlling for entering academic ability

u_{0j}: value-added equation residual for school *j*. This is the value-added score, which is the difference between the mean score for all schools and school *j*'s mean score. The more positive the residual, the higher the value-added score.

Notice that slightly different school-level control variables were used, as graduate degree-granting status was not a significant predictor of scores at either the freshman or senior level, and selectivity was only significant at the senior level (Liu, 2011b). Further,

the model in Equation 3 explicitly included freshman scores as an institutional-level predictor of senior-level scores. By using only one equation that included information on both freshmen and seniors, the model in Equation 3 (Steedle, 2012) was more parsimonious and straightforward than the previous version of the HLM model for residual gain scores (Liu, 2011b). Specifically, the previous iteration required three steps: the first two steps entailed estimating the model (once to generate residuals for freshmen and once for seniors), and then a third step to subtract the freshman from senior residuals to produce the value-added estimate. In contrast, the updated HLM in Equation 3 required only one estimation of the model, and the u_{0j} parameter provided the value-added estimate—thus calculating value-added estimates in one step rather than three (Steedle, 2012).

The correlation of HLM and OLS value-added estimates was approximately .72, similar to previous findings. Further, year-to-year estimates were correlated more highly using the HLM (.55) than when using the OLS method (.32; Steedle, 2012). Reliability of the OLS-based approach ranged from .62-.64 across samples, whereas the HLM approach estimated reliability at .74-.75—a clear improvement. Entering academic ability was controlled for at both the student and institutional level, which may explain why estimates were more stable in the HLM methods (Steedle, 2012).

Unlike the OLS method, the HLM method produced a standard error for each school's value-added estimate, which could then be used to create a 95% confidence interval as a measure of precision. The value-added estimates, along with their 95% CI's, could then be plotted, providing a quick visual as to whether schools are performing

above or below expected. For these reasons, HLM is now the preferred method of value-added calculation, and is employed for all VSA value-added calculations (Steedle, 2012).

Advantages of Residual Gain Scores: On the surface, the concept of expected performance compared to actual performance was easy to understand. Further, residual gain scores were on the same metric as the test itself, making it easier for stakeholders to determine whether the differences in the actual and expected performances were important. The average person would immediately have a sense of whether, for example, a 5-point difference in expected and actual performance was meaningful, assuming familiarity with the original test's scale.

Disadvantages of Residual Gain Scores: Although most people can conceptualize the difference between observed and expected performance, the residual parameter used to actually define this difference was easily (and often) misunderstood by those unfamiliar with statistical methods. Although it is true that a residual is the difference between the observed and predicted performance of the unit of analysis (here, institutions), statistically speaking a residual is simply unexplained variance in the dependent variable. In the case of OLS regression, the residual is often represented by *e* and referred to as "error" (Cohen et al, 2003). Error can refer to either measurement error, or error in model specification (e.g., a significant predictor has been left out of the model; a relationship modeled as linear is actually curvilinear, etc.). Under this definition, then, the value-added estimates produced by residual gain score methods were simply error, whether from measurement or model misspecification. In other words, rather than interpreting value-added estimates as the differences in student learning due to attendance at a particular institution, they should be interpreted as differences in learning

due to factors other than those represented in the model, which could include systematic error (measurement or model specification).

Another disadvantage of residual gain scores was that they were inherently norm-referenced. Because residual gain score calculations were based on the sample mean, half of the observed residual scores would be below the mean (Pike, 1992). In other words, an institution could be contributing quite a bit to student learning and growth as well as meeting criterion-referenced cutoffs for outcomes-based measures, but still be categorized as "below expectations" or "well below expectations", depending on the sample of schools included in the analysis. In contrast, an institution may not be contributing much to student learning, but if none of the schools in the sample are contributing much, a school could be classified as above expectations for no particular reason at all. Because interpretations like these were often made in an absolute rather than relative sense, a clear flaw to residual gain scores was the fact that the performance categorizations were norm- rather than criterion-referenced. This was a major flaw with the residual gain score value-added model.

Shortcomings of Current Value-Added Models in Higher Education.

The current applications of value-added methods have serious shortcomings. Namely, 1) reliance on cross-sectional rather than longitudinal data; 2) frequent causal interpretation of value-added estimates; 3) identified value-added institutional differences had no clear meaning; and 4) potential model misspecification due to neglect of test-taking motivation effects on value-added estimates. Each of these issues will be briefly summarized.

First, value-added models in higher education regularly employed cross-sectional data when calculating value-added (Kim & Lalancette, 2013). In the purest sense of value-added, measuring student growth was of interest, not simply comparing the performance of a group of incoming students to a group of graduating students (Astin, 1982). If growth, defined as the change in individuals over time, was what we truly wish to measure, "cross-sectional data-so easy to collect and so widely available-will not suffice" (Singer & Willett, 2003, p. 3). Cross-sectional data cannot capture within-individual changes, because *the same people are not measured across time*. Indeed, even though cross-sectional value-added methods were the most common in higher education, researchers acknowledged that there is no current research indicating that cross-sectional and longitudinal value-added estimates are comparable (Liu, 2011b). This is a shortcoming in value-added research that needs to be investigated (Liu, 2011a, 2011b).

Second, value-added estimates from current models implied that the institution (or program) in question was the cause of differences in performance. To think that no other factor than the institution itself could impact change in student performance was unrealistic (Klein, Benjamin, Shavelson, & Bolus, 2007). Similar to Braun's (2005) assertion that definitive statements about teacher quality in K-12 could be made based on value-added estimates, statements regarding institutional quality cannot be made based on cross-sectional value-added estimates in higher education. Causal statements imply randomization—and the admissions process at colleges and universities completely negates any assumption that students were randomly chosen to attend institutions. At best, we could say that "a… portion of the improvement is likely to be a function of the learning opportunities provided by a college education" (Klein et al, 2007), but by no

means should causal statements about institutional quality be made (Pike, 2006). In fact, both the ASA (2014) and the *Standards* (AERA, APA & NCME, 2014) cautioned users against making undue interpretations based on value-added modeling. Specifically, the *Standards* stated that "it is important to clearly understand which questions each growth model can (and cannot) answer, what assumptions each growth model is based on, and what appropriate inferences can be derived from each growth model's results" (AERA, APA, & NCME, 2014).

Third, value-added estimates provide no information regarding the meaning of institutional differences or diagnostic information for improving practice (Steedle, 2012). As noted earlier, value-added estimates were used to place institutions into performance categories that indicate whether they were performing at, above, or below expectations. However, that is all—there was no information about what institutions who performed above expectations may be doing right, and what institutions who performed below expectations may need to improve (Klein, et al, 2007; Steedle, 2012). If an institution wanted to use value-added estimates obtained from methods described in this literature review, there would be no way for them to identify where they should focus attention and resources in order to improve student learning.

Finally, current value-added methodology does not model examinee motivation. Researchers have acknowledged the established positive relationship between examinee motivation and test performance in low-stakes contexts, and have called for research that investigates the impact of motivation on value-added estimates (Liu, 2011a, 2011b; Liu et al, 2012). In addition, the *Standards* (AERA, APA, & NCME, 2014) advised institutions to consider information regarding examinee motivation when interpreting test scores.

This last shortcoming of current value-added models provides a major impetus for the current study and brings us to the next portion of this literature review, an overview of examinee motivation and its recognized impact on test performance.

Low-Stakes Testing and Test-Taking Motivation.

Previous researchers expressed concern regarding examinee motivation and recommended that the influence of test-taking motivation on value-added estimates should be investigated (Liu, 2011a, 2011b). The relationship between examinee motivation and test performance is well established, particularly in low-stakes assessment contexts (DeMars, 2000; Sundre & Kitsantas, 2004; Wolf et al., 1995). In the following section, I will first distinguish between low- and high-stakes testing, and then discuss the implications of low examinee motivation in low-stakes testing conditions. Finally, I will review interventions that have been suggested to combat the effects of low examinee motivation.

Low- vs. high-stakes testing. When students take a test, they usually complete it under one of two conditions: either the results of the test have personal consequences to examinees, or there are no consequences. In the former situation, commonly referred to as high-stakes tests, students' performances on the test have personal implications: scores on the SAT are used for college admissions; a final exam may determine whether a student passes or fails a course. Because high-stakes tests have personal consequences to the examinee, it is assumed that students try their hardest in these situations. In contrast, because students do not experience any personal consequences when taking low-stakes tests, the assumption that students put forth their best effort is questionable. Further, if students do not try on a low-stakes test, lack of motivation could "introduce construct-

irrelevant variance in the test score, and the test will be a measure not only of actual knowledge but also of motivation" (Knekta & Eklöf, 2014, p. 1). The following sections will first describe test-taking motivation, and then discuss the impacts that the construct-irrelevant variance it introduces has on test scores and their validity in low-stakes contexts.

Test-taking motivation. Test-taking motivation is defined as "a student's engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (Wise & DeMars, 2005, p. 2). Expectancy-value theory (Wigfield & Eccles, 2000) provided a useful framework for thinking about test-taking motivation. In expectancy-value theory, motivation to perform was comprised of two components: 1) the expectancy a person has for task success; and 2) the value a person places on that task. Expectancy referred to the person's perceived ability to capably complete the task at hand, whereas value referred to the value a person ascribed to that activity. In other words, a person's motivation to perform a particular task was directly related to not only whether they thought they could accomplish the task, but also whether they perceived the task as important and if they were willing to give up a valued alternative to complete the task. If either expectancy or value was low, motivation suffered.

Value in particular could be thought of in many ways, and either positively or negatively impact motivation. There were three categories of value thought to positively influence motivation: intrinsic value, utility value, and attainment value (Eccles, Barber, Updegraff, & O'Brien, 1998). First, intrinsic value related to the interest in and enjoyment for completing a particular task. In other words, intrinsic value lies in whether a person will feel good about completing a particular task. For test-taking tasks to have

intrinsic value, the test-taker would have to enjoy actually completing the test. Next, utility value, or relevance, relates to whether a person feels that completing the task provides some sort of useful outcome. In a testing context, this may be something like completing an SAT because the scores provided one of the components for a successful college application. As another example, for many nursing students, a Chemistry course final exam signaled readiness to begin courses they perceived as more directly related to nursing practice. Finally, attainment value, or importance, relates to the value placed on doing well on a particular task. In a testing context, attainment value could be related to whether students think doing well on a test was important in the context of their everyday lives, such as passing a licensure or certification exam.

In contrast to the three types of value that could positively influence motivation, cost tends to have a negative impact on motivation. Cost is typically seen as what a person must give up in order to complete a task, such as not engaging in another valued activity (Eccles, et al, 1998). In a mixed-methods study that investigated student motivation in low-stakes testing (Williams & Swanson, 2014), one student illustrated the cost of completing a low-stakes test when he said that he did not think the tests were important because "I have other homework and tests that I want to worry about more than this that are for grades and stuff." For this student, the cost of doing well on the low-stakes tests paled in comparison to the value of completing other assignments that would impact his grades. Other forms of cost could be the arduousness of a test, items requiring a good deal of mental effort to answer correctly (mentally taxing items; Wolf et al., 1995), or a large amount of time required to complete the test. In a high-stakes environment, these issues may not be barriers to student motivation, as the cost icould be

outweighed by the personal benefit to the student. In low-stakes contexts, however, a difficult test, mentally taxing items, or the time given up may be seen as not worthwhile to the student, given that the results had no personal impact on students. Thus, motivation suffered.

Given the motivational framework described above, "a student's engagement in and expenditure of energy toward taking a test" (Wise & DeMars, 2005, p. 2) was related to both expectancy and value with value being operationalized as attainment value, or importance. Cost also factors in, as students were often giving up valued alternatives to complete the tests, such as time that could be used to work on homework for classwork, as articulated by the student above. In other words, motivation to take a test was related to not just the student's perceived ability to complete the test, but also the importance the student placed on taking a test and the personal cost of taking the test. Not surprisingly, literature has referred to test-taking motivation as comprised of two distinct constructs: perceived importance and test-taking effort. (Sundre, 1997; Thelk, Sundre, Horst, & Finney, 2009). Perceived importance of a test related to how strongly a student feels doing well on the test is of concern to them as an individual, whereas effort referred to how hard a student tries during a test. As mentioned earlier, these are two distinct factors in test-taking motivation but they are related.

One instrument used to measure student motivation in testing contexts, the Student Opinion Scale (SOS; Sundre & Thelk, 2007), was grounded in expectancy-value theory. There are two subscales on the SOS: perceived importance and test-taking effort. In a study that investigated the use of the SOS, the correlation between perceived importance and test-taking effort was consistently in the .6 - .7 range (Hopkins-

Whetstone, Swain, Williams, Finney & Sundre, 2013). In low-stakes testing, though, not only were effort and importance related to one another, they were related to performance, which will be discussed next.

Effect of motivation on test performance. It is not realistic to think that examinees have the same level of motivation for taking a test in low-stakes testing contexts as they might in a high-stakes testing context. Indeed, literature consistently showed that test-taking motivation was related to test performance in low-stakes contexts, such that students consistently performed more poorly when taking a test in low-stakes vs high-stakes context (DeMars, 2000; Sundre & Kitsantas, 2004; Wolf et al., 1995). Wolf et al. (1995) studied the effect of what they termed "mentally taxing items" (p. 342) on test performance; mentally taxing items were described as items requiring a good deal of mental effort to answer correctly. In other words, mentally taxing items had high cost to students. Results of the study indicated that test stakes influenced performance, and that item type also impacted performance. Mentally taxing items consistently resulted in lower performance in the low-stakes testing condition than in the high-stakes context. Although students may have performed more poorly in the low-stakes context because they were not as motivated as students in the high-stakes context, no *direct* measure of motivation was collected in this study. Thus, it was not possible to empirically describe the relationship between motivation, test stakes, and performance.

Later research investigated the effect of test stakes and item types (constructed vs. selected response) on performance for students taking a high-school diploma test (DeMars, 2000). Not only did students perform lower in the low-stakes context, but there was an interaction between item format and test stakes. Students achieved lower scores

on constructed response items in both low- and high-stakes contexts, but these scores were significantly lower in the low-stakes condition. These findings supported the idea that performance suffers in low-stakes tests, and that test-taking motivation may be related to the differential in performance between low- and high-stakes testing contexts (DeMars, 2000; Wolf et al., 1995). These results also suggested that tasks requiring more investment in effort resulted in lower performance—in other words, a high cost for test-takers. However, no direct measure of examinee motivation was available to actually test this hypothesis.

Sundre and Kitsantas (2004) also investigated the impact of task type (constructed vs. selected response), motivation, and test stakes on student performance. Their study included both selected- and constructed-response tasks. Students completed two parallel tests, each consisting of 30 multiple-choice items and one essay question. One administration of the test counted for their grade and the other did not; thus, test stakes were counter-balanced and experimentally manipulated. In addition, student motivation and self-regulation for completing the tests was explicitly measured, unlike previous studies. Similar to previous research, student performance suffered more on constructed- than selected-response tasks in low stakes testing. More importantly, the results empirically supported what previous studies had alluded to: test-taking motivation was, in fact, related to performance in low-stakes testing contexts, although not in high-stakes contexts (Sundre, & Kitsantas, 2004). In other words, the construct-irrelevant variance introduced by the lack of student motivation in low-stakes testing attenuated test scores. These results were particularly startling in the low-stakes essay condition. Further, this study was conducted using a course-embedded testing context; thus, all students were

prepared to complete the assigned tasks. Despite student preparation and random assignment of the parallel test forms to high and low-stakes conditions, it was clear that students elected *not* to perform at their best level *and* reported this lack of motivation on the SOS scales.

More recently, studies have investigated the impact of motivation on value-added estimation. Liu et al. (2012) conducted a study that explored the impact of differential testing consequences on student motivation and performance in three conditions: Control (results of the test were only used for research purposes); Institutional (control instructions plus aggregate test scores were reported to the institution and may also be shared with potential employers) and Personal (control instructions plus individual test scores were reported to faculty and may also be shared with potential employers). Students completed the ETS Proficiency Profile (both the multiple-choice portion and the optional essay) as a measure of performance, and the SOS (Sundre & Thelk, 2007) as a measure of motivation. Even though the SOS has two established subscales (test-taking effort and perceived importance), only one total score was reported as the measure of motivation. Significant differences in motivation and performance were found between the Control and Institutional and Control and Personal conditions, but not between Institutional and Personal conditions. Although these findings may initially seem to be cause for alarm, it is likely that these significant differences between conditions were somewhat artificial. The Control condition told students that their scores would be used for research purposes and nothing more. There is a very small chance that an institution would engage in such a practice, making only the Institutional and Personal conditions

realistic. Recall that no significant differences in motivation or performance were observed between those two conditions.

In response to concerns about the effect of motivation in value-added estimation, a crude measure of learning over time was calculated by comparing scores on the ETS Proficiency Profile across academic levels. This comparison of scores, labeled value-added, was conducted via an ANOVA of group differences. In other words, this was not a true value-added estimate in that it was cross-sectional rather than longitudinal. Based on the ANOVA results, academic level was a significant predictor of test performance. This was a positive and much desired result. The fact that seniors performed better than sophomores is exactly what institutions would hope to observe, as it indicates seniors have likely learned more than sophomores due to spending more time learning at the institution. Difference scores between sophomores and seniors for each of the three motivational conditions along with an effect size were also calculated. The researchers then attempted to illustrate how motivation may differ between sophomores and seniors by comparing difference scores between the most motivated sophomores and least motivated seniors as well as comparing the least motivated sophomores and most motivated seniors. Results indicated that there were large gains (.72 SD for the multiple-choice portion and .65 SD for the essay portion) between the least motivated sophomores and most motivated seniors, but little to negative gain (-.23 SD) between the most motivated sophomores and least motivated seniors. These differences between students of different motivational levels may provide evidence for bias in value-added estimates due to test-taking motivation (Liu et al., 2012).

However, these statements were misleading. Value-added estimates comparing the least motivated sophomores to the most motivated seniors were calculated simply by extracting the sophomores from the Control group and comparing them to the seniors in the Personal group, rather than identifying low-motivated sophomores and high-motivated seniors based on their SOS scores. Similar procedures were followed to compare the most motivated sophomores to the least motivated seniors, by extracting the sophomores from the Personal group and comparing them to the seniors in the Control group (L. Liu, personal communication, January 20 2015). Motivation scores were calculated, and were higher for sophomores but no statistical testing was reported. Based on these procedures, there was no clear way to describe the impact of motivation on performance, nor empirically test the relationship.

This work was an important first step in exploring the impact of motivation level on value-added estimates—research that was sorely needed (Liu, 2011a, 2011b). Further, it provided the first evidence that the concerns regarding the effect of motivation on value-added estimates in higher education may have been well founded. However, while conclusions regarding bias due to differential motivation were certainly important, the way in which value-added was calculated was misleading: the sample was cross-sectional rather than longitudinal, and did not include freshmen. The rationale for not including freshmen was that freshmen may feel intimidated, and therefore try harder, than students who had been at the institution longer. In other words, pre-test motivation could have been artificially inflated if freshmen were included. However, if value-added estimates were desired, students *should* be tested prior to any college coursework to get an accurate baseline score—this was exactly what Astin (1982) advocated. As a result, not including

freshmen was a shortcoming and limits the inferences that drawn from the results of Liu et al.'s (2012) study. The attempt to include motivation in the value-added calculation was at least more than previous studies' efforts. However, as noted earlier there were problems with how the value-added calculations were carried out. Thus, we cannot know whether the effect sizes were a function of actual motivation, academic level, or other factors.

In response to Liu, et al.'s work, Finney, et al. (under review) conducted a longitudinal study to examine the impact of test consequence manipulation on test-taking motivation and performance. In contrast to previous research, longitudinal data (rather than cross-sectional) was gathered from the same cohort of students. Further, freshmen were included in the study with pre-test occurring prior to beginning the first semester and post-test after 45-70 credit hours of college coursework. Students were randomly assigned to one of three realistic testing conditions: Control (aggregate results would be used for institutional purposes); Feedback (aggregate results would be used for institutional purposes and students would receive feedback on their individual performance); and Personal (aggregate results would be used for institutional purposes, students would receive feedback on their individual performance, and personal scores would be released to faculty). After taking an arduous quantitative and scientific reasoning test students responded to the SOS, which asked students to think about the test they had just completed. Separate scores for perceived importance and test-taking effort were calculated, allowing the researchers to examine whether perceived importance and test-taking effort had different relationships with test performance across the three motivational conditions. Perhaps the most important finding of the study was that the

change in perceived importance and the *change* in test-taking effort significantly predicted the *change* in test scores over time. In other words, the longitudinal *change* in perceived importance of the test and *change* in test-taking effort were significant predictors of value-added scores. Depending on motivational condition, the change in test-taking motivation explained 27% - 41% additional variance in the change in test scores above and beyond gender and personality variables! This was the first empirical evidence that test-taking motivation directly biased value-added estimates, and validated the concerns previously raised regarding the impact of test-taking motivation on value-added estimates (Liu, 2011a, 2011b; Liu et al., 2012). Although this was compelling evidence for the impact of examinee motivation on value-added estimates, the study used the difference scores as the measure of value-added, not residual gain scores used in other research. Additionally, sophomores and juniors were used in the post-test sample. It is conceivable that had seniors been used for post-test, as in the Liu et al. (2012) study, the effects of change in motivation may have been even more pronounced, and test performance would also potentially be increased.

Interventions to combat low motivation. Based on the evidence just discussed, it is clear that low examinee motivation is a problem in low-stakes testing. Much research has been dedicated to addressing this problem, both through behavioral interventions and statistical methods. Both areas will be discussed next.

Behavioral interventions. Many methods of combating low motivation in testing have been investigated. Specifically, the effect of proctor training (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009), subtle variation in test stakes (Finney, et al., under review; Liu, et al 2012;), providing feedback on performance (Finney, et al., under

review; Swain, Williams, Hopkins, Sundre, & Finney, 2013; Williams, Swain, Hopkins, Sundre, & Finney, 2013; Wise & DeMars, 2005) and appealing to students' sense of academic citizenship (Huffman, Adamopolous, Murdock, Cole, & McDermid, 2011; Zilberberg, Brown, Harmes, & Anderson, 2009) have all been investigated as ways to increase student motivation in low-stakes testing. Results of these interventions have been mixed, however, with some interventions working better than others. Training proctors and appealing to academic citizenship seemed to increase student motivation, while providing feedback to students has not been an effective intervention. In contrast, subtle manipulation of test stakes has had mixed results.

When test proctors were intentionally trained to be more consistent in test instruction delivery and test monitoring behaviors, student motivation and performance both increased (Lau, et al., 2009). Prior to the beginning of testing sessions, proctors were instructed to engage in behaviors associated with higher student motivation, such as thanking students for their effort, conveying the importance of the tests, modeling a positive attitude about testing, and maintaining an environment favorable to testing. Not only did student motivation increase once these behaviors were encouraged among all proctors, but differences in motivation across testing rooms decreased. Another effective strategy for increasing student motivation and performance was appealing to students' academic citizenship (Huffman, et al., 2011). With this approach, telling students about the purpose of the tests and the use of scores on major-specific assessment tests increased student performance.

Feedback has been suggested as a potential solution for low motivation in low-stakes testing (Wise & DeMars, 2005). Providing feedback to students seemed to be an

intuitive and simple way to increase student motivation, but research indicated that it was ineffective. Several studies have compared groups of students in low-stakes testing who were offered the opportunity to receive feedback on their performance with groups of students taking the same tests who were not offered feedback on their performance, finding no significant difference in motivation between the groups (Baummert & Demmrich, 2001; Finney, et al., under review; Swain et al., 2013; Williams et al., 2013). Further, another study that investigated what type of feedback students preferred (norm-referenced vs. criterion referenced) found that less than 40% of students even took the time to look at their scores and feedback (Socha, Swain, & Sundre, 2013).

In some cases, subtly manipulating the stakes of a test have been effective (Liu, et al., 2012) while in other instances subtle manipulations in test stakes have been ineffective (Finney, et al., under review; Swain et al., 2013; Williams et al., 2013). In the Liu, et al. study, manipulating test stakes was effective. However as discussed earlier, this may have been due to the fact that the control condition was an unrealistic testing situation—no operational testing program would tell students that results would only be used for research purposes. In the two conditions where stakes were realistic, no differences in motivation or performance were observed. In contrast, other studies have not found manipulations of test stakes to impact student motivation and performance (Finney, et al., under review; Swain et al., 2013; Williams et al., 2013). In these studies, all conditions under which test stakes were manipulated were realistic, and part of an operational testing program. The results from those studies provided evidence that subtle manipulations in test stakes were not an effective method to increase student motivation while still keeping the test low-stakes.

Because behavioral interventions have had such mixed results on impacting motivation, other methods of dealing with low-motivated examinees have been developed. A statistical technique referred to as motivation filtering was one of the most promising methods and will be discussed next.

Statistical interventions. The most common statistical method for dealing with low motivation in testing was a technique called motivation filtering. Motivation filtering is a method whereby students exhibiting no or low motivation are removed from subsequent analyses. If it is implemented, the *Standards* recommend that "decision criteria regarding whether to include scores from individuals with questionable motivation" are reported (AERA, APA, & NCME, 2014, p. 213). Examinees exhibiting low motivation were identified through self-report effort measures (SRE) such as the SOS (Sundre & Wise, 2003; Wise, Wise, & Bhola, 2006) or response-time effort (RTE; Wise & Kong, 2005; Wise & DeMars, 2010). Response-time effort was a method for identifying unmotivated examinees based on the amount of time a student takes to respond to an item on a computer-based test. Unmotivated students often engaged in *rapid-guessing behavior*, where they respond to test questions very quickly, and their answers are essentially random guesses (Wise & Kong, 2005). Response-time effort measures used the amount of time a student spend when responding to a question on a test to determine a student's level of motivation, which can then be used for motivation filtering purposes. Assumptions of and specific procedures for motivation filtering are described next.

Motivation filtering required two assumptions: 1) a valid measure of motivation and 2) that student motivation was not related to actual ability (Sundre & Wise, 2003).

The first assumption is easily met by choosing a motivation instrument with good psychometric properties, such as the SOS (Sundre & Thelk, 2007) or through RTE. Response-time effort has been deemed a valid measure of motivation, as it was correlated with other measures of examinee effort, was unrelated to ability, provided item-level information, and demonstrated internal consistency (Wise & Kong, 2005). Although both methods (SRE and RTE) were legitimate, research suggested that RTE measures of motivation may be more effective in identifying low-motivated examinees (Rios, Liu, & Bridgeman, 2014).

The second assumption, that motivation and proficiency were unrelated, can easily be investigated through simple correlational studies. In higher education, SAT scores often serve as a measure of proficiency; calculating a correlation between SAT scores and test performance will quickly reveal whether the second assumption is upheld. Previous research has consistently reported that motivation and SAT scores were unrelated (Rios et al., 2014; Sundre & Wise, 2003; Wise & DeMars, 2010; Wise et al., 2006).

Once both assumptions are tested and upheld, motivation filtering can proceed by first removing non- or low-motivated students from the dataset and then proceeding with analyses (Sundre & Wise, 2003; Wise & DeMars, 2010; Wise et al., 2006). In order to classify non- or low-motivated examinees, first a threshold for low motivation must be identified. Determination of what constitutes a non- or low-motivated examinee is ultimately up to the researcher, but guidelines do exist. For example, if the SOS was used as a SRE measure, previous studies suggested that removing examinees with SOS effort scores below 13-17 on a 5-25 point scale was appropriate (Wise et al, 2006). When

using RTE measures of motivation, the thresholds for determining low-motivated examinees have been chosen via visual inspection of response time distributions (Wise & Kong, 2005; Wise & DeMars, 2010), or via the NT10 method, where the threshold was determined by calculating 10% of the average response time (Wise & Ma, 2012). Regardless of whether response time distributions or the NT10 method were used, RTE thresholds should not exceed 10 seconds, as it was difficult to characterize responses of more than 10 seconds as guessing behavior (Setzer, Wise, van den Heuvel, & Ling, 2013; Wise & Ma, 2012).

Once the threshold for non- or low-motivated examinees was identified, students whose scores fell below the motivation threshold were removed from the data set—this was the "filtering" portion of motivation filtering. Next, any desired analyses proceeded. Previous research has consistently shown that after motivation filtering, higher mean scores and higher convergent validity coefficients emerged, evidence that data was more trustworthy once non- or low-motivated examinees were removed from analyses (Rios et al., 2014; Sundre & Wise, 2003, Wise & DeMars, 2010; Wise et al., 2006). If motivation filtering was chosen as a method to deal with the effects of non- or low-motivated examinees, the *Standards* advised that "decision criteria regarding whether to include scores from individuals with questionable motivation should be clearly documented" (AERA, APA, & NCME, 2014, p. 213).

The Current Study

Based on the information presented in this literature review, it was important that the impact of examinee motivation on value-added estimates be explored more fully. In addition, such a study should be conducted in a context that recognizes both assessment

for improvement and accountability demands. It was clear from the literature previously discussed that there was a consistent relationship between examinee motivation and performance in low-stakes testing contexts, such that the lower the motivation, the lower the performance. Further, because low-stakes testing was the most common source of data collection for both assessment and accountability purposes, it was imperative that the impact of motivation on value-added estimates within that context be explored. To that end, this study built upon previous research by analyzing both cross-sectional and true longitudinal data measuring both test performance and examinee motivation. The longitudinal samples included students when they first entered the institution and those same students at the mid-point of their undergraduate career. Two value-added estimation procedures, raw difference scores and HLM, were invoked to provide additional comparative information regarding the impact of motivation when using different value-added methods. Ultimately, five research questions were posed and answered across two study phases. In the first phase, the impact of using cross-sectional vs longitudinal data and analyses was explored, followed by comparison of raw difference score and HLM value-added estimation procedures. In the second phase, the effect of motivation on value-added estimates was investigated in the context of institutional assessment. These phases and research questions were:

Phase 1: Comparing value-added estimates using longitudinal vs. cross-sectional data. As discussed earlier, many value-added methods used cross sectional data even though longitudinal data was preferable for measuring growth over time. This phase of the study was intended to compare the results of value-added estimates when using cross-sectional rather than longitudinal data. It also directly compared value-added

estimates for two value-added methods. There were two research questions for this phase of the study. They were:

1. Do value-added estimates generated from raw difference scores differ when using cross-sectional vs. longitudinal data?
2. When using only longitudinal data, do value-added estimates generated from raw difference scores and an HLM differ?

Phase 2: Investigating the effect of motivation on value-added estimates using longitudinal data. The second phase of this study explicitly investigated the effect of test-taking motivation on value-added estimates in an assessment context. Further, because longitudinal data was preferable when measuring growth, this phase only used longitudinal data. The research questions for this phase were:

3. Do value-added estimates for raw difference scores and an HLM differ when perceived importance of the test is included in the HLM?
4. Do value-added estimates for raw difference scores and an HLM differ from one another when test-taking effort is included in the HLM?
5. Do value-added estimates for raw difference scores and an HLM differ when both test-taking effort and perceived importance of the test, together, are included in the HLM?

CHAPTER THREE

Methods

As discussed in the literature review, value-added models were used when engaging in assessment for improvement as well as assessment for accountability purposes. In higher education, two methods of value-added estimation prevail: raw difference scores and residual gain scores. Research to date has explored residual gain score methods by comparing OLS methods to HLM methods, concluding that HLM methods are superior (Liu, 2011b; Steedle, 2012). However, little research has compared the results of raw difference score methods to residual gain score methods, and even less research has investigated the impact of examinee motivation on value-added estimates from these methods (Liu, 2011a, 2011b; Liu, et al., 2012; Steedle, 2012). This study addressed these areas, in both assessment for improvement and assessment for accountability frameworks. Phase 1 investigated the impact of using cross-sectional vs. longitudinal data on value-added estimates generated using raw difference scores. Phase 2 examined the effects of motivation on value-added estimates generated with an HLM and compared those to raw difference score value-added estimates. The research questions for this study were:

Phase 1: Comparing Value-Added Estimates using Longitudinal vs. Cross-sectional Data

1. Do value-added estimates generated from raw difference scores differ when using cross-sectional vs. longitudinal data?

2. When using only longitudinal data, do value-added estimates generated from raw difference scores and an HLM differ?

Phase 2: Investigating the Effect of Motivation on Value-Added Estimates using Longitudinal Data

3. Do value-added estimates for raw difference scores and an HLM differ when perceived importance of the test is included in the HLM?
4. Do value-added estimates for raw difference scores and an HLM differ from one another when test-taking effort is included in the HLM?
5. Do value-added estimates for raw difference scores and an HLM differ when both test-taking effort and perceived importance of the test, together, are included in the HLM?

Participants and Procedures

Phases 1 and 2. In phases 1 and 2, this study analyzed archival data collected during a campus-wide institutional Assessment Day at a public, Mid-Atlantic, 4-year liberal arts institution. Data collected at these Assessment Days were primarily used for assessment of General Education and Student Affairs programing. At Assessment Day testing sessions, students were randomly assigned (based on the last two digits of their University-issued ID number) to attend either a morning or afternoon testing session; each session was scheduled for three hours. During these sessions, students completed a battery of cognitive and non-cognitive tests, including a measure of test-taking motivation. Not only were students randomly assigned to a morning or afternoon session, but they were also randomly assigned to testing rooms, with each room having a different configuration of tests. Such assignments resulted in large, random,

representative samples of students completing each test. All testing rooms were staffed by trained proctors, with one proctor serving as the lead and one or more assistant proctors providing support to the lead proctor. Although the data gathered during this institution-wide Assessment Day was high-stakes for administrators in that results were used to make decisions about curriculum as well as for accreditation and reporting purposes, the tests were low-stakes for the students themselves. That is, there were no consequences to students based on their performances. Hence, student motivation may not have been optimal given the lack of personal consequences to the students themselves.

Students included in this study participated in Assessment Day twice: the first time as an incoming freshman and the second after completing 45-70 credit hours of college coursework. The first testing occasion (pre-test) took place as an integral component of a required five-day Orientation program for entering new students; thus, students had not taken any college coursework at pre-test. At the second testing occasion (post-test), students had attended the university approximately three semesters and were classified as sophomores or juniors, depending on their credit hour completion. The post-testing session occurred on a Tuesday in mid-February, and all classes were cancelled until 4 p.m. to accommodate the morning and afternoon testing sessions. Students were typically assigned to complete the same tests as at pre-test, providing longitudinal data. Assessment Day participation was required for all students; those who did not attend had a hold placed on their registration until they completed assessment testing at a make-up session. Over the last several years, 93% of students participated on their assigned Assessment Day, and virtually 100% will complete the assessment tasks at some point.

The archival data analyzed in Phases 1 and 2 of this study were collected at Assessment Day in Fall *2010 (N* = 1401) for the pre-test measure and Spring 2011 (*N* = 1130) and Spring 2012 (*N* = 1072) for post-test data. Using data from two post-test occasions allowed both longitudinal and cross-sectional analyses to be conducted. The cross-sectional data set consisted of the Fall 2010 students (62.6% female; *N* = 1198 after examinees with missing data removed) for the pre-test measure, and Spring 2011 students (62.5% female; *N* = 932 after examinees with missing data removed) at post-test; students involved in these two testing sessions *were not* members of the same cohort; sample sizes after missing data were removed are reported in Table 2. The longitudinal data set (*N* = 621, 66% female) consisted of the Fall 2010 students at pre-test, and the Spring 2012 students at post-test; students involved in these two testing occasions *were* members of the same cohort. Thus, a matched longitudinal sample was obtained from the Fall 2010/Spring 2012 cohort. In other words, true repeated measures were calculated for the Fall 2010/Spring 2012 cohort.

While it was possible that students may have taken the post-test measure at a time other than 3 semesters after pre-test, these students were not included in the current study. Because this study is the first of its kind, it was important to limit "noise" as much as possible. Including only students who had taken the post-test at the midpoint of their undergraduate studies provided a control for maturity. If students who had taken the post-test earlier (due to entering the institution with accumulated AP, transfer, or dual enrollment credits) or later (due to slower credit hour accumulation or failing classes) may have added construct-irrelevant variance to the data set, thus potentially diluting the effects of interest in this study.

Instruments

Phases 1 and 2. ***Natural World Test, version 9 (NW-9).*** As a measure of quantitative (QR) and scientific reasoning (SR) skills, a random sample of A-Day participants took the 66-item Natural World test, version 9 (NW-9; Sundre, Thelk, & Wigtil, 2008). Sample sizes of students with complete data for each testing occasion are reported in Table 2 (that is, students who did not complete the test on both occasions and thus had missing responses were removed from the dataset). This particular cognitive test was chosen because it measured knowledge (as opposed to a test consisting of attitudinal items) and was arduous, resulting in scores that varied but were not susceptible to ceiling or floor effects.

The NW-9 purports to measure QR and SR skills in General Education courses. Rigorous test development and validity studies support the content alignment and construct validity of the test. For example, Sundre and Thelk (2010) conducted a multi-institutional study that included four additional institutions with a variety of missions and diverse student populations. Even when employed in such a variety of contexts, 92 - 100% of the NW-9 items mapped to the QR and SR student learning outcomes of each of the home institutions (Sundre & Thelk, 2010). As evidence of concurrent validity, "over 90% of correlations between relevant course grades and scores on [the NW-9] were positive" and generally ranged from 0.30−0.50 (Sundre & Thelk, 2010, p. 9). Further, Biology faculty at JMU mapped NW-9 items to program learning objectives and found that 25 of the 66 NW-9 items mapped to 7 out of the 14 skill objectives in the Biology major (Hurney, Brown, Griscom, Kancler, Wigtil, & Sundre, 2011). These skill objectives included distinguishing association from causation, formulating and evaluating

hypotheses, and using mathematics to understand biological phenomena, among others. Although mapping items to only 7 out of 14 skill objectives might seem like a low number, it was expected by faculty because the NW-9 is a general education assessment instrument. In addition, the Biology major objectives that did not have NW-9 items mapped to them were better suited to constructed-response and lab activities (e.g., obtaining data, evaluating sources of information, and communicating results effectively) than objectives mapped to the NW-9. The NW-9 is a multiple choice test; therefore, it was not expected that all or even most items would map to that type of learning objective. (Hurney, et al., 2011). Finally, historic reliability estimates range from .71 - .85 for the total test, indicating adequate internal consistency (Sundre & Thelk, 2010; Sundre, et al., 2008).

Student Opinion Scale. The Student Opinion Scale (SOS; Sundre & Moore, 2002; Sundre & Thelk, 2007; Thelk, et al., 2009) was administered for purposes of collecting information regarding student motivation while taking a test; recall from the literature review that this is the same measure of motivation used by Liu, et al. (2012) as well as Finney, et al. (under review). The SOS, built on the Expectancy-Value motivation framework (Wigfield & Eccles, 2000), is a 10-item instrument comprised of two subscales: Importance and Effort. Items on the Importance subscale ask students to respond to statements regarding the importance of the test to the examinee; an example item is: "Doing well on these tests was important to me." Items on the Effort subscale ask students to respond to statements regarding how hard they tried on the test; an example item is: "I gave my best effort on these tests." Each subscale consists of 5 items using a 5-point Likert scale (1= strongly disagree; 5= strongly agree); two items are

reverse coded. High scores on the subscales indicate high levels of effort and high levels of importance; possible scores range from 5-25. Historic reliability estimates in low-stakes testing contexts range from .80-.84 for importance items and from .83-.86 for effort items (Sundre & Thelk, 2007). Longitudinal measurement invariance is supported for the SOS (Sessoms, 2014). Student responses to the SOS referred to all of the tests students completed on Assessment Day, not just the NW-9. As described earlier, students complete a battery of both cognitive and non-cognitive instruments during the three-hour testing session; see Table 2 for a list of tests completed during each testing session (Fall 2010, Spring 2011, and Spring 2012). In each testing session, at least one test measured cognitive knowledge (noted in Table 2), while the remaining tests measured students' attitudes and other non-cognitive constructs. Although the NW-9 was not the only cognitive test administered during the testing session, it was the most arduous cognitive test completed. Thus, scores on the SOS were likely representative of students' attitudes toward completing the NW-9.

Table 2
Instruments and Testing Order

Testing Occasion	N^a	Test 1	Test 2	Test 3	Test 4	Test 5
Fall 2010	1198	NW-9*	ATL-10	USSP*	GAP-3	SOS
Spring 2011	932	NW-9*	ATL-10	CABS	GAP-4	SOS
Spring 2012	902	NW-9*	ATL-10	USSP*	SOS	

[a] sample size after students with incomplete data on the NW-9, SOS, or SAT math were removed
*denotes a cognitive test
Note: Test abbreviations are as follows:
NW-9 = Natural World, version 9
ATL-10 = Attitudes Toward Learning, version 10
USSP: United States Society and Politics
GAP: General Attitudes Packet, versions 3 (GAP-3) and 4 (GAP-4)
CABS: Civic Minded Student Sale
SOS: Student Opinion Scale

Treatment of Missing Data

Only students who had complete data on all measures were retained for analyses. Maximum likelihood (ML) estimation used in the analyses described later removed cases with missing data when calculating parameter estimates. Thus, retaining cases in the longitudinal data set with missing data would have resulted in the analyses described below being conducted on slightly different samples of students. Although listwise deletion is not a preferred method of dealing with missing data, cases with missing data were listwise deleted in this study to ensure that all analyses were conducted on *exactly* the same sample.

Analyses

Value-added models. To answer the research questions, Phase 1 of this study compared value-added estimates generated from longitudinal and cross-sectional data and Phase 2 compared value-added estimates generated from raw difference scores to value-added estimates generated from an HLM that included motivation variables in the model. Value-added estimates in both phases were calculated with the scores on the NW-9 test as the indicator of student learning. The following paragraphs outline the specific analyses used to address each research question in this study. Analyses were conducted using SAS 9.4 (SAS Institute, 2013) and plotting was done using R version 3.1.1 (R Core Team, 2014).

Phase 1: Comparing value-added estimates using longitudinal vs. cross-sectional data within an institution. Phase 1 of this study addressed the first two research questions. Specifically, these questions were:

1. Do value-added estimates generated from raw difference scores differ when using cross-sectional vs. longitudinal data?
2. When using only longitudinal data, do value-added estimates generated from raw difference scores and an HLM differ?

In Phase 1 of the study, the Fall 2010 data set served as pre-test data for both the longitudinal and cross-sectional analyses. The Spring 2011 data set served as post-test for the cross-sectional analyses, while the Spring 2012 data set served as the post-test for the longitudinal analyses.

To address the first research question (Do value-added estimates generated from raw difference scores differ when using cross-sectional vs. longitudinal data?), a difference score was calculated for the cross-sectional data set by subtracting the mean NW-9 score on the pre-test (Fall 2010 sample) from the mean NW-9 score on the post-test (Spring 2011 sample). The calculation of a raw difference score was slightly different for the longitudinal data set, as this dataset consisted of a matched sample of students. A difference score was first calculated for each student by subtracting the NW-9 score on the pre-test (Fall 2010) from the NW-9 score on the post-test (Spring 2012); an overall mean difference score was then calculated from the mean of individual difference scores.

To test whether analyses from the cross-sectional and longitudinal data sets would result in different conclusions about student learning, statistical significance tests were conducted. An independent-samples *t*-test was conducted on the Fall 2010/Spring 2011 NW-9 scores (the cross-sectional sample) to determine whether the difference in means was statistically significant, and a dependent-samples *t*-test was conducted on the NW-9

scores in the Fall 2010/Spring 2012 matched sample (longitudinal data). Effect sizes of Cohen's *d* were computed to indicate the practical significance of the differences. As change over time was the question of interest in this study, both effect sizes were calculated in the repeated-measures metric and could be directly compared with one another (see Appendix B for a full explanation of and formula used to transform effect sizes to a common metric).

If the two *t*-tests resulted in the same conclusion (that is, both tests and effect sizes were non-significant or both tests and effect sizes were significant), then it was evidence that using cross-sectional data instead of longitudinal data is reasonable given that cross-sectional data is less resource-intensive to collect than longitudinal. If, however, the *t*-tests resulted in different conclusions (for example, the *t*-test for cross-sectional data was non-significant and the *t*-test on the longitudinal data was significant), it would suggest that cross-sectional and longitudinal data and analyses are not interchangeable when calculating value-added estimates. This finding would support the assertion that longitudinal data and analyses are preferable to cross-sectional data and analyses due to potential issues in cross-sectional data regarding cohort and historical effects (Biesanz, West, & Kwok, 2003; Nesselroade & Ghisletta, 2003).

To address the second research question (When using only longitudinal data, do value-added estimates generated from raw difference scores and an HLM differ?) the raw difference score from the longitudinal data set was compared to a value-added estimate generated from an HLM.

As discussed in the literature review, the current VSA method of value-added is a residual gain score obtained from a hierarchical linear model (HLM) that controls for

entering academic ability and freshman-level CLA scores at the institutional level and entering academic ability at the student and institutional level. Controlling for ability at the institutional level by using freshman CLA scores allows the CLA method to use cross-sectional data instead of longitudinal. The resulting value-added estimates are used to determine whether institutions are performing at, above, or below expectations, compared to their peers (see Table 1 in Chapter 2). The CLA model is represented by the following equation (Steedle, 2012, p. 652):

$$
\begin{aligned}
CLA_{ij} &= \beta_{0j} + \beta_{1j}(EAA_{ij} - \overline{EAA}_j) + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\overline{EAA}_j) + \gamma_{02}(\overline{CLA}_{fr,j}) + u_{0j} \\
\beta_{1j} &= \gamma_{10}
\end{aligned}
\qquad (1)
$$

$$
CLA_{ij} = \gamma_{00} + \gamma_{01}(\overline{EAA}_j) + \gamma_{02}(CLA_{fr,j}) + \gamma_{10}(EAA_{ij} - \overline{EAA}_j) + u_{0j} + r_{ij}
$$

Where:

EAA_{ij}: entering academic ability of student i at school j, measured with SAT scores

$\overline{EAA}_j$: average entering academic ability of school j

$CLA_{fr,j}$: average CLA score for freshmen in school j

γ_{00}: school-level value-added equation intercept; mean CLA score when entering academic ability and freshman CLA scores are 0.

γ_{01}: school-level value-added equation slope coefficient for senior mean EAA; it indicates the increase in school-level senior CLA scores for every one-point increase in mean entering academic ability, controlling for other variables in the HLM.

γ_{02}: school-level value-added equation slope coefficient for freshman mean CLA; it indicates the increase in school-level senior CLA scores for every one point increase in mean freshman-level CLA scores, controlling for other variables in the HLM.

γ_{10}: student-level value-added equation slope coefficient for EAA (assumed to be same across schools); it indicates the overall increase in CLA scores for every one-point increase in the student-level EAA, controlling for other variables in the HLM.

r_{ij}: residual for student *i* in school j. σ^2 is the variance of the residuals, which is the pooled within-school variance of CLA scores after controlling for entering academic ability

u_{0j}: value-added intercept residual for school *j*. This is the residualized difference score, or value-added estimate, which indicates how much the senior-level CLA performance for school *j* differs from the predicted performance for school *j*, given its average entering academic ability and average freshmen CLA score. Large, positive residuals indicated above-expected performance whereas large, negative residuals indicated below-expected performance. Residuals near zero indicated expected performance.

This model generates a value-added estimate that indicates whether a school is performing below, at, or above expectations (see Table 1 in Chapter 2 for more information), essentially comparing institutions to one another. However, the purpose of this study is to investigate the effects of motivation on value-added estimates for a single institution. Thus, although the CLA model is simply an HLM and other relevant variables can be added in if they are of interest, it must be modified for use at a single institution. An HLM was chosen for value-added estimation to accommodate the nesting

of students within institutions, but nesting of data can be thought of in alternative ways, one of which is time points nested within students for repeated-measures data. If the CLA model is reconceptualized as time points nested within students at a single institution, a modified version of the CLA could be:

$$Y_{ti} = \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti}$$
$$\pi_{0i} = \beta_{00} + \beta_{01}(EAA_i - \overline{EAA}) + r_{0i}$$
$$\pi_{1i} = \beta_{10} + r_{1i} \quad (2)$$

$$Y_{ti} = \beta_{00} + \beta_{01}(EAA_i - \overline{EAA}) + \beta_{10}(Time_{ti}) + r_{1i}(Time_{ti}) + r_{0i} + e_{ti}$$

Where:

Y_{ti}: dependent variable (test score at time *t* for student *i*)

π_{0i}: intercept for student *i*

π_{1i}: slope for person i: the change in the dependent variable between time points

$Time_{ti}$: time indicator, time *t* for student *i*

e_{ti}: residual for every observation (student *i* at time *t*); indicator of the within student variation once controlling for time

β_{00}: overall intercept across all students

β_{01}: slope for centered entering academic ability; this indicates the effect of entering academic ability on each person's intercept.

EAA_i: entering academic ability for student *i*

$\overline{EAA}$: average entering academic ability for all students

r_{0i}: deviation of student *i*'s intercept from the overall intercept; this would be analogous to the value-added score in the institutional-level CLA model.

β_{10}: overall slope across all students (average change in the DV across time points). If only two time points were used, this would be equivalent to the change in scores from pre- to post-test, and is analogous to an average difference score estimate of value-added, after controlling for the other variables in the model.

r_{1i}: deviation of student *i* from overall slope; the parameter can only be estimated in a dataset with 3 or more time points. This is the deviation from the average change over time for student *i*. With three or more time points, this parameter represents a residual gain value-added estimate; it indicates the deviation of individual slopes (e.g., change over time) from the average slope.

This HLM can be applied to a data set that contains pre- and post-test data and also include any theoretically relevant variables. In the case of the NW-9, prior coursework is known to predict student performance (Williams, Socha, & Sundre, 2013) and was included in the model. Although the effect of SAT math scores to predict performance on the NW-9 has not been empirically investigated, SAT math scores were included here for two reasons. First, as the NW-9 measures quantitative and scientific reasoning skills, it is reasonable to think that SAT math scores may predict NW-9 performance. Second, SAT scores have been used in other value-added analyses as a

proxy for academic ability when estimating value-added (Liu, 2011b; Steedle, 2012). Thus, the HLM used here becomes:

$$
\begin{aligned}
NW9_{ti} &= \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti} \\
\pi_{0i} &= \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + r_{0i} \\
\pi_{1i} &= \beta_{10}
\end{aligned}
\tag{3}
$$

$$
NW9_{ti} = \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + \beta_{10}(Time_{ti}) + r_{0i} + e_{ti}
$$

where $course_i$ is a variable that indicates the number of General Education science and math credits student *i* has taken. All of the coefficients maintain the same interpretation as before, with the addition of β_{02}, which is the slope coefficient for courses. Notice also that there is no longer a random effect for the slope at Level 2. When only two time points are available for longitudinal data, the data are essentially recreated during parameter estimation and limited information is available to describe growth trajectories. Thus, it is not appropriate to let the Level-1 slope (π_{1i}) vary (Raudenbush & Bryk, 2002; Rogosa et al, 1982; Singer and Willet, 2003). As a result, a residual gain score was not possible as a value-added estimate for a single institution when using a pre/post dataset with only two time points.

This model served as the base for answering research questions 2-5, with β_{10} the coefficient of interest, as it is the estimate of average change between pre- and post-test.

Since r_{1i} could not be estimated in this model[1], β_{10} was used as the indicator of value-added. Thus, β_{10} indicated the average change in NW-9 scores between pre-and post-test after controlling for SAT math scores and number of science and math credits earned. The β_{10} parameter estimate was on the same metric as raw change scores, making the two estimates comparable. Both $course_i$ and SAT math scores were grand-mean centered, making intercepts interpretable as the average NW-9 scores for students with average amount of math and science coursework and average SAT math scores. All HLM analyses in research question 2-5 used full maximum likelihood estimation (ML), as ML estimation is preferred when the parameters of interest are fixed effects (Enders, 2005).

To answer the research question, the HLM described above was used to estimate parameters and obtain β_{10}. This parameter was then compared to the mean raw difference score of the longitudinal sample and its 95% confidence interval. A β_{10} that was outside the 95% CI was considered to be different enough as to indicate that value-added

---

[1] Note that variability in slopes could be estimated if the error terms were fixed (Voelkle, 2007). Specifically, Finney, et al. (in press, Footnote 3) fixed the error terms to a value of ((1-reliability estimate)*variability of the variable). Doing so did not affect the value-added means, yet permitted estimation of the interindividual differences in change (Voelkle, 2007). The model estimated in the current study was an attempt to emulate current value-added models, and thus freely estimated error terms.

estimates from an HLM and raw difference scores were not comparable. In other words, if the HLM estimate of value-added fell outside the 95% confidence interval of the raw difference score estimate of value-added, the two value-added estimates were judged to *not* indicate the same conclusions about student growth over time.

Phase 2. Investigating the effect of motivation on value-added estimates using longitudinal data within an institution. The second phase of this study analyzed only longitudinal data, and addressed three research questions. Specifically, these questions were:

3. Do value-added estimates for raw difference scores and an HLM differ when perceived importance of the test is included in the HLM?
4. Do value-added estimates for raw difference scores and an HLM differ from one another when test-taking effort is included in the HLM?
5. Do value-added estimates for raw difference scores and an HLM differ when both test-taking effort and perceived importance of the test, together, are included in the HLM?

For research questions 3, 4, and 5, the raw difference score calculated for the longitudinal data in Phase 1 served as the raw difference score to which the HLM value-added estimate was compared. The HLM used to answer Research Question #2 served as the baseline HLM for the models that included motivation, which will be described further in the explanation of analyses for each research question.

Perceived importance. To address Research Question 3 (Do value-added estimates for raw difference scores and an HLM differ when perceived importance of the

test is included in the HLM?), the raw difference score for the longitudinal data set (that is, the matched sample of Fall 2010 and Spring 2012 students) was compared to the value-added estimate generated by the following HLM:

$$
\begin{aligned}
&NW9_{ti} = \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti} \\
&\pi_{0i} = \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + r_{0i} \\
&\pi_{1i} = \beta_{10} + \beta_{11}(\Delta imp_i - \overline{\Delta imp})
\end{aligned}
\tag{4}
$$

$$
NW9_{ti} = \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + \beta_{10}(Time_{ti}) + \beta_{11}(\Delta imp_i - \overline{\Delta imp})(Time_{ti}) + r_{0i} + e_{ti}
$$

Notice that Equation 4 is the same as Equation 3, with the addition of Δimp_i, a variable to indicate the change in student *i*'s perceived importance of the test between pre- and post-test. Previous research indicated that the *change* in perceived importance of the test, rather than importance at either pre- or post-test, may be a significant predictor of the *change* in student performance over time (Finney et al., under review), hence its inclusion here. For ease of interpretation Δimp_i was centered, as the scale for importance does not have a meaningful zero point.

To generate value-added estimates that illustrate the effect of varying levels of perceived importance of the test on value-added estimates (i.e., an interaction between perceived importance and time), Equation 4 was estimated three times: once with Δimp_i centered at the mean change in importance (shown in Equation 4), once with Δimp_i centered one standard deviation above the mean, and once with Δimp_i centered one standard deviation below the mean. However, the interaction term (β_{11}) in Equation 4 was significant, which meant it was inappropriate to interpret β_{10} as the value-added

estimate in the presence of the significant interaction term. In order to illustrate the effect of the change in effort on value-added estimates (β_{10}), values of 0 for Δimp_i were substituted into each of the three versions of Equation 4 generated from the estimation of Δimp_i centered at the different levels of change in importance. Doing so allowed two things: first the interaction term dropped out of the equation and second, β_{10} could now be interpreted as predicting the change over time for students at each of the three levels of change, controlling for number of credits and SAT math scores. Three separate equations resulted, each of which had its own β_{10}; see Chapter 4 for the final equations.

To determine whether the two methods of estimating value-added differed, each β_{10} in the three equations (see Chapter 4) was compared to the 95% confidence interval of the raw difference score estimate of value-added. If any of the three β_{10}'s fell outside the 95% CI of the raw difference score estimate of value-added, it was evidence that test-taking importance biased value-added estimates of student learning over time.

Test-taking effort. To address Research Question 4 (Do value-added estimates for raw difference scores and an HLM differ from one another when test-taking effort is included in the HLM?), the same procedure was followed as when addressing Research Question 3, but using the following equation:

$$
\begin{aligned}
&NW9_{ti} = \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti} \\
&\pi_{0i} = \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + r_{0i} \\
&\pi_{1i} = \beta_{10} + \beta_{11}(\Delta effort_i - \overline{\Delta effort})
\end{aligned}
\qquad (5)
$$

$$
\begin{aligned}
&NW9_{ti} = \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + \beta_{10}(Time_{ti}) \\
&+\beta_{11}(\Delta effort_i - \overline{\Delta effort})(Time_{ti}) + r_{0i} + e_{ti}
\end{aligned}
$$

where $\Delta effort_i$ indicated the change in test-taking effort between pre- and post-test. Again, previous research indicated that *change* in test-taking effort was a significant predictor of the *change* in performance, not test-taking effort at pre- or post-test. To illustrate the effect of change in test-taking effort on value-added estimates, Equation 5 was estimated three times: once with $\Delta effort_i$ centered at the mean change in effort (shown in Equation 4), once with $\Delta effort_i$ centered one standard deviation above the mean change, and once with $\Delta effort_i$ centered one standard deviation below the mean change. However, similar to the previous HLM analyses including Δimp_i as a predictor, the interaction between time and $\Delta effort_i$ (β_{11}) in Equation 5 was significant, which meant it was inappropriate to interpret β_{10} as the value-added estimate in the presence of the significant interaction term. The same process was followed as in the analyses including Δimp_i to illustrate the effect of the change in effort on value-added estimates (β_{10}). Values of 0 for $\Delta effort_i$ were substituted into each of the three versions of Equation 5 generated. Doing so allowed two things: first, the interaction term dropped out of the equation and second, the equation was now interpreted as predicting the change in NW-9 scores over time for students with change in effort at each of the three levels, after controlling for number of credits and SAT math scores.

Three separate equations were obtained (see Chapter 4) and each resulting β_{10} was compared to the 95% confidence interval of the raw difference score estimate of value-added to determine whether the two value-added estimates differed. If any of the three β_{10}'s fell outside the 95% CI of the raw difference score estimate of value-added, it was evidence that test-taking effort biased value-added estimates of student learning over time.

Perceived importance and test-taking effort, combined. Finally, to address research question 5, (Do value-added estimates for raw difference scores and an HLM differ when both test-taking effort and perceived importance of the test, together, are included in the HLM?), a third version of the HLM was constructed, shown in Equation 6:

$$
\begin{aligned}
NW9_{ti} &= \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti} \\
\pi_{0i} &= \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + r_{0i} \\
\pi_{1i} &= \beta_{10} + \beta_{11}(\Delta effort_i - \overline{\Delta effort}) + \beta_{12}(\Delta imp_i - \overline{\Delta imp})
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
NW9_{ti} = {} & \beta_{00} + \beta_{01}(SAT_i - \overline{SAT}) + \beta_{02}(course_i) + \beta_{10}(Time_{ti}) + \beta_{11}(\Delta effort_i - \overline{\Delta effort})(Time_{ti}) \\
& + \beta_{12}(\Delta imp_i - \overline{\Delta imp})(Time_{ti}) + r_{0i} + e_{ti}
\end{aligned}
$$

The version of the HLM in Equation 6 included both $\Delta effort_i$ and Δimp_i as a way to model the combined influence of the change in test-taking motivation and perceived importance on value-added estimates. Similar to the methods to address research questions 3 and 4, both $\Delta effort_i$ and Δimp_i were centered one standard deviation below the mean changes, at the mean changes, and one standard deviation above mean changes in importance and effort to illustrate the effect of different combinations of levels of importance and effort on value-added estimates. Equation 6 was initially estimated nine times, once for each combination of mean, +1SD, and -1SD levels of the $\Delta effort_i$ and Δimp_i variables. As with the previous analyses, the interaction between time and $\Delta effort_i$ (β_{11}) as well as the interaction between time and Δimp_i (β_{12}) in Equation 6 were significant, which meant it was inappropriate to interpret β_{10} as the value-added estimate in the presence of the significant interaction terms. To aid interpretation, the same process was followed as in the previous analyses to illustrate the combined effect of the

change in importance and effort on value-added estimates (β_{10}). Values of 0 for Δimp_i and $\Delta effort_i$ were substituted into each of the nine versions of Equation 6 generated. Once again, the interaction terms dropped out of the equations, and the resulting β_{10} parameters were interpreted as predicting the change in NW-9 scores over time after controlling for SAT math scores and number of science and math credits.

Each of the nine resulting β_{10}'s was compared to the 95% confidence interval of the raw difference score estimate of value-added to determine whether the two value-added estimates differed. If any of the nine β_{10}'s fell outside the 95% CI of the raw difference score estimate of value-added, it was evidence that the particular *combination* of the change in test-taking effort and change in perceived importance of the test biased value-added estimates of student learning. If all nine β_{10} estimates fell outside the 95% CI of the raw difference score estimate of value-added, then it was considered evidence that *any* change in effort or importance biased value-added estimates.

Finally, AIC and BIC fit statistics were examined to determine which of the four models (no motivation variables, importance only, effort only, or importance and effort combined) had the best fit. Additionally, the nested models (importance and effort combined vs. importance only and importance and effort combined vs. effort only) were compared with a chi-square goodness-of-fit test to determine whether using the both the importance and effort scores was necessary, or if one subscale would suffice to measure motivation when calculating value-added estimates.

CHAPTER FOUR

Results

This chapter presents results for all analyses conducted to answer the research questions posed in Chapter 3. I will first provide a description of the sample and descriptive statistics for variables used in the subsequent analyses. Following the sample description, results for each of the research questions will be presented in order and organized by the two phases of the study. Brief interpretation of results will be presented here, but full discussion of results and their implications for future research, policy, and assessment practice will be presented in Chapter 5.

Sample Description.

As detailed in Chapter 3, three samples were used in this study: one pre-test and two post-test samples. Students in all three samples completed the Natural World 9 (NW-9), a cognitive test of quantitative and scientific reasoning, and the Student Opinion Scale (SOS), a measure of test-taking motivation. In addition, SAT math scores and number of science and math credits completed at the time of post-testing were included in the analyses. Table 3 reflects the descriptive statistics for the samples used in analyses reported in this chapter.

Sample sizes and descriptive statistics for variables all three data sets are reported in Table 3. These sample sizes are for the full samples, with those students missing data on any of the variables removed via listwise deletion; a full explanation of the decisions regarding missing data is found is Chapter 3. As seen in Table 3, the post-test samples were similar to one another in SAT math scores, number of science and math credits

earned, self-reported importance and effort, and total NW-9 scores. SAT math scores and the number of science and math credits were similar across all samples. Importance as well as effort decreased between the pre- and post-test, although importance decreased more than effort. Even though both importance and effort show a decrease over time, it is important to note that overall, both motivation variables are fairly high at both time points (recall that possible scores range from 5-25), given that this is a low-stakes test for students. Both post-test samples showed an increase in NW-9 scores from the pre-test group despite decreased importance and effort. These trends will be explored through the analyses presented later in the chapter.

Table 3
Descriptive Statistics of Samples

		SAT Math	Science and Math Credits	Importance	Effort	NW-9
Fall 2010*	Mean	580.03	n/a	15.91	18.68	46.41
	SD	66.21	n/a	3.76	3.60	6.69
	Min	340	n/a	5	5	27
	Max	800	n/a	25	25	66
Spring 2011**	Mean	583.95	7.96	12.68	18.5	49.83
	SD	70.58	4.39	4.44	4.27	7.82
	Min	360	0	5	5	16
	Max	800	26	25	25	65
Spring 2012***	Mean	578.31	8.03	13.57	17.79	49.42
	SD	69.3	4.65	4.41	3.95	7.50
	Min	200	0	5	5	16
	Max	800	24	25	25	64

* N = 1198
** N = 932
*** N = 902

The students in the Fall 2010 and Spring 2012 samples were members of the same cohort, and therefore a matched longitudinal sample was possible. Table 4 reports descriptive statistics for the matched longitudinal sample including reliabilities of all

scores used in these analyses as well as difference scores. Not all students were retained in the longitudinal sample, as 577 students in the pre-test sample did not have post-test data, and 281 students in the post-test sample did not have pre-test data. Recall from Chapter 3 that difference scores for the longitudinal sample were simply the average of the difference scores for each student; these were calculated simply by subtracting pre-test scores from post-test scores. All scores, including difference scores, were considered reliable as the Cronbach alphas all fell above the .7 cutoff mark of acceptable reliability; all but one difference score fell above the more rigorous .80 cutoff recommended for applied research contexts (Nunnally, 1978; Lance, Butts, & Michaels, 2006).

Table 4
Descriptive Statistics of Scores from the Longitudinal Sample (N = 621)

Variable	Mean	Std Dev	Min	Max	α
SAT Math	583.93	64.92	410	800	n/a
Science and Math Credits	7.79	4.46	0	22	n/a
Effort (pre-test)	18.81	3.49	5	25	.83
Effort (post-test)	17.81	3.87	5	25	.84
Effort change	-1.00	3.93	-19	14	.84
Importance (pre-test)	16.21	3.52	5	25	.80
Importance (post-test)	13.59	4.25	5	25	.85
Importance change	-2.62	4.37	-16	13	.83
NW-9 (pre-test)	46.79	6.47	27	66	.73
NW-9 (post-test)	49.95	7.20	16	64	.81
NW-9 change	3.16	6.04	-33	20	.77

Since change in performance on the NW-9 over time is of primary interest in this study, a plot of the change in NW-9 scores over time for individuals and the overall longitudinal sample is found in Figure 1. It is clear from Figure 1 that there is considerable variability in performance at each time point as well as over time for these students. To better illustrate the range of change over time, Figure 2 contains a plot of the distribution of change scores. Notice that in the distribution of change scores, the

distribution is fairly normal, but with a slight negative skew. Although the overall mean change scores is positive, it is clear from the distribution that some students' scores actually decrease over time, and in some cases by a substantial amount. The following analyses will focus on these change scores and shed some light on why this phenomenon may be occurring.

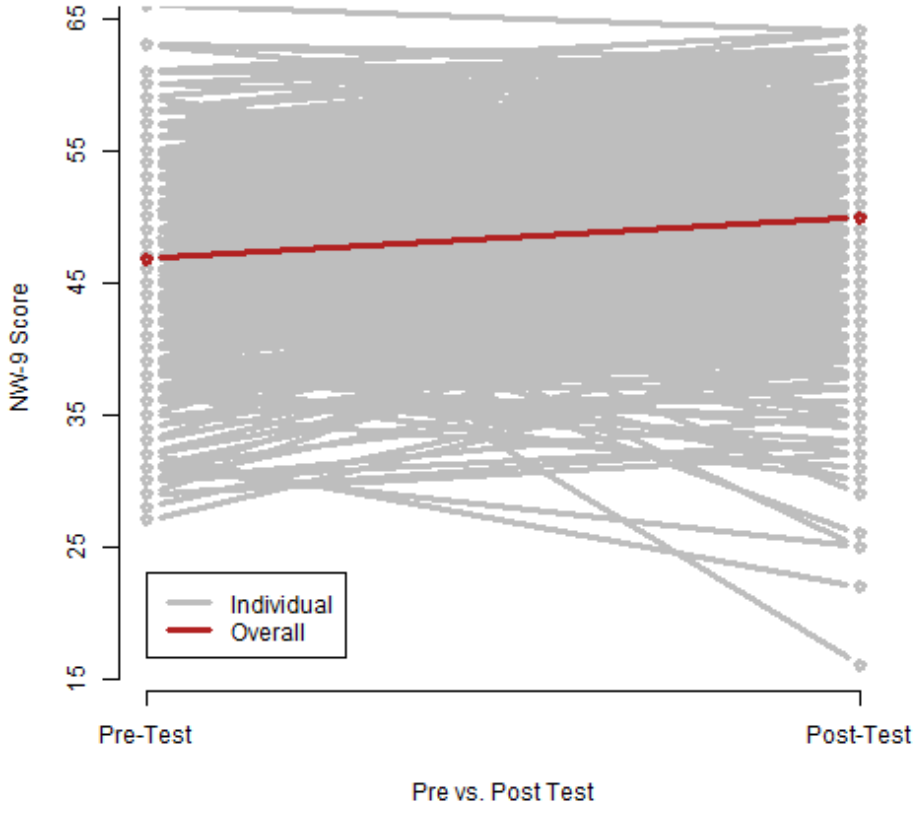

Figure 1. Change in NW-9 performance over time.

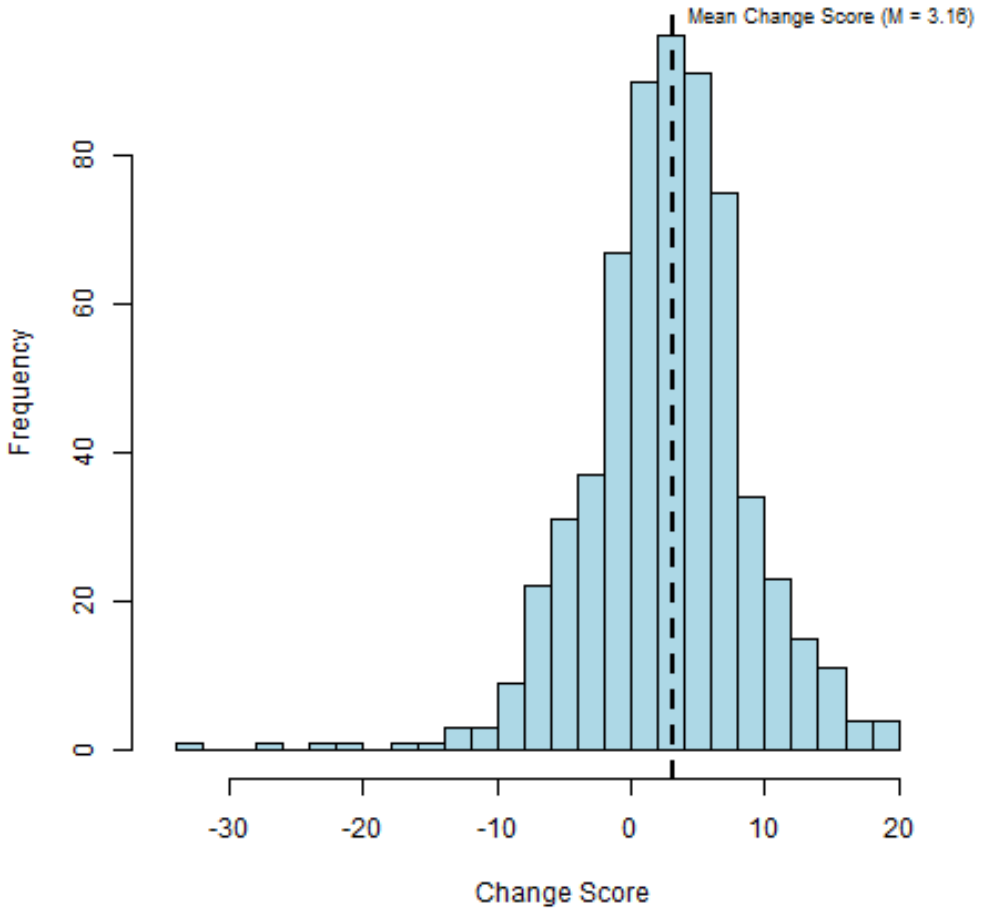

Figure 2. Distribution of NW-9 change scores.

Phase 1: Comparing Value-Added Estimates using Longitudinal vs. Cross-sectional Data

Research question #1: Do value-added estimates generated from raw difference scores differ when using cross-sectional vs. longitudinal data? On average, students' scores on the NW-9 increased by just over 3 points between pre- and post-test for both the cross-sectional and longitudinal samples; see Table 5. The

difference in performance for the cross-sectional data was tested with an independent t-test of differences between means, and a dependent t-test of mean differences was used to analyze the longitudinal data. Both tests showed significant differences in mean performance between pre- and post-test, as shown in Table 5. As described in Chapter 3, effect sizes were calculated so that both the cross-sectional and longitudinal effect sizes would be on a change score metric, as change in performance is the focus of the current study. Effect sizes for both the cross-sectional and longitudinal sample indicated that performance increased about a half of a standard deviation between pre- and post-test. The results of the *t*-tests and effect sizes indicated that in these samples, cross-sectional and longitudinal analyses result in similar conclusions about student performance over time. However, the similarity in results may be an artifact of the exceptionally high retention rate at the institution where the study was conducted (92.3% for the 2012 cohort; James Madison University, 2014); this finding may not generalize to other institutions.

Table 5
Comparison of Cross-Sectional and Longitudinal Estimates

	Mean Difference	95% CI	*t*	*p*	Effect size (*d*)
Cross-Sectional	3.18	2.58-3.78	10.21	< .001	0.53
Longitudinal	3.16	2.68-3.64	13.03	< .001	0.52

Research question #2: When using only longitudinal data, do value-added estimates generated from raw difference scores and an HLM differ? The raw difference score estimate of value-added for the longitudinal sample, shown in Table 5, was 3.16 with a 95% confidence interval of 2.68-3.64. This indicated that on average, student performance on the NW-9 increased by 3.16 points from pre- to post-test

assessments, and the plausible range of values for average change in NW-9 scores between pre- and post-test ranged from 2.68-3.64.

Prior to fitting the HLM model of value-added, an intercept-only HLM was conducted to obtain the intraclass correlation coefficient (ICC), which is an indicator of the amount of variability in the NW-9 scores due to nesting of data. The ICC for the intercept-only model was .53, indicating that over 50% of the variability in NW-9 scores was due to differences among students. After obtaining the ICC via the intercept-only model, parameter estimates were calculated for the HLM model used to generate value-added estimates that control for SAT math scores and number of science and math credits earned (Equation 3 from Chapter 3); results are presented in Table 6. The full model fit significantly better than the intercept-only model ($\chi^2 = 343.5$, $p < .001$), indicating that SAT scores and number of science and math credits should be included in the model, as they explain a significant amount of variability in NW-9 performance above and beyond that accounted for by the intercept alone.

The 95% confidence interval of the raw difference scores was compared to the HLM value-added estimate of 3.16 (β_{10}; see Table 6). Recall from Chapter 3 that all variables in the HLM model were grand-mean centered, making β_{10} interpretable as the average change in NW-9 scores from pre- to post-test for students of average SAT math scores and average number of math and science credits earned. The raw difference score and HLM estimates were identical, and the confidence interval included the parameter estimates indicating that the two value added estimates did not meaningfully differ. Parallel to the results from cross-sectional and longitudinal analyses in Research Question #1, this finding may be a distinction of the particular institution where the data

was collected and may not be generalizable to other locations. This finding and its implications will be discussed in greater detail in the final chapter of this dissertation.

Table 6
Summary Table of Parameter Estimates for Longitudinal Model of Quantitative and Scientific Reasoning Achievement

	Intercept-only Model			No Motivation Variables Included		
Parameter	Estimate	SE	*p*	Estimate	SE	*p*
Fixed Effects						
Intercept (β_{00})	48.34	.24	<.001	46.79	.24	< .001
SATmath (β_{01})				.05	.003	< .001
Credits (β_{02})				.10	.05	.04
Time (β_{10})				3.16	.24	< .001
Random Effects						
σ^2		23.23			18.24	
τ_{00}		26.08			18.51	

Note: SAT Math scores and credits were grand mean centered, to make the intercepts interpretable as the NW-9 score for students with average SAT scores and average number of science and math credits.
* ICC = 0.53

Phase 2: Investigating the Effect of Motivation on Value-Added Estimates using Longitudinal Data

Research question #3: Do value-added estimates for raw difference scores and an HLM differ when perceived importance of the test is included in the HLM?

Parameter estimates for an HLM that includes perceived importance (Equation 4 from Chapter 3) are found in the column labeled "Importance Only Included" in Table 7. Only the estimate of β_{10} centered at the mean change in importance is shown Table 7; Table 8 contains parameter estimates for each of the three estimations of β_{10} along with the raw difference score and its 95% CI for comparison.

Table 7

Summary Table of Parameter Estimates for Longitudinal Model of Quantitative and Scientific Reasoning Achievement, Motivation Variables Included

	Importance Only Included			Effort Only Included			Importance and Effort Included		
Parameter	Estimate	SE	*p*	Estimate	SE	*p*	Estimate	SE	*p*
Fixed Effects									
Intercept (β_{00})	46.79	.24	< .001	46.79	.24	< .001	46.79	.24	< .001
SATmath (β_{01})	.05	.003	< .001	.05	.003	< .001	.05	.003	< .001
Credits (β_{02})	.10	.05	.03	.10	.05	.03	.10	.05	.03
Time (β_{10}) (centered at mean Δimport)	3.16	.24	< .001						
Time (β_{10}) (centered at mean Δeffort)				3.16	.24	< .001			
Time (β_{10}) (centered at mean Δimport and mean Δeffort)							3.16	.24	< .001
Time*Δimp(β_{11})	.15	.05	.002				.15	.06	.009
Time*Δeffort(β_{12})				.19	.05	< .001	.10	.05	.04
Random Effects									
σ^2		17.89			18.04			17.84	
τ_{00}		18.77			18.35			18.56	

Note: SAT Math scores and credits were grand mean centered, to make the intercepts interpretable as the NW-9 score for students with average SAT math scores and average number of science and math credits.

If model parameters from the HLM estimation with change in importance centered at its mean are substituted into Equation 4, the following equation results:

$$NW9_{ti} = \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti}$$
$$\pi_{0i} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + r_{0i} \quad (7)$$
$$\pi_{1i} = 3.16 + .15(\Delta imp_i - \overline{\Delta imp})$$

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.16(Time_{ti}) + .15(\Delta imp_i - \overline{\Delta imp})(Time_{ti}) + r_{0i} + e_{ti}$$

In Equation 7, note that β_{10} = 3.16, which is the intercept of the Level 2 equation predicting the slope of time. In other words, for students of average change in importance NW-9 scores increased by 3.16 points between pre- and post-test, after controlling for SAT math scores and number of science and math credits earned.

Note that the HLM includes a significant interaction between testing occasion (*Time*) and change in importance (parameter β_{11} in Table 7). Because it is inappropriate to interpret a main effect (β_{10}, the parameter for time) in the presence of a significant interaction, in order to truly get a sense of the effect of perceived importance on value-added estimates it is necessary to rearrange Equation 7. Since change in importance was centered, if zero is substituted for change in importance in Equation 7 for each of the three estimations of the HLM, three separate equations result after the interaction drops out of the model, shown in Equations 8-10 below. These equations now represent the prediction of NW-9 scores for students one standard deviation below, at, and one standard deviation above the mean change in importance.

1SD below mean *Δimp*:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 2.50(Time_{ti}) + r_{0i} + e_{ti} \quad (8)$$

At mean *Δimp*:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.16(Time_{ti}) + r_{0i} + e_{ti} \quad (9)$$

1SD above mean Δimp:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.82(Time_{ti}) + r_{0i} + e_{ti} \quad (10)$$

The resulting equations now illustrate the slope for time (HLM value-added estimate) at three levels of the change in test-taking importance over time. Figure 3 shows the three equations graphically for students who have average SAT math scores and average number of science and math credits. From this plot, it is clear that while students score the same at pre-test regardless of their self-reported perceived importance, post-test scores differ depending on how much perceived importance of the test changes between pre- and post-test. Performance at post-test is lowest for those students whose perceived importance decreased over time.

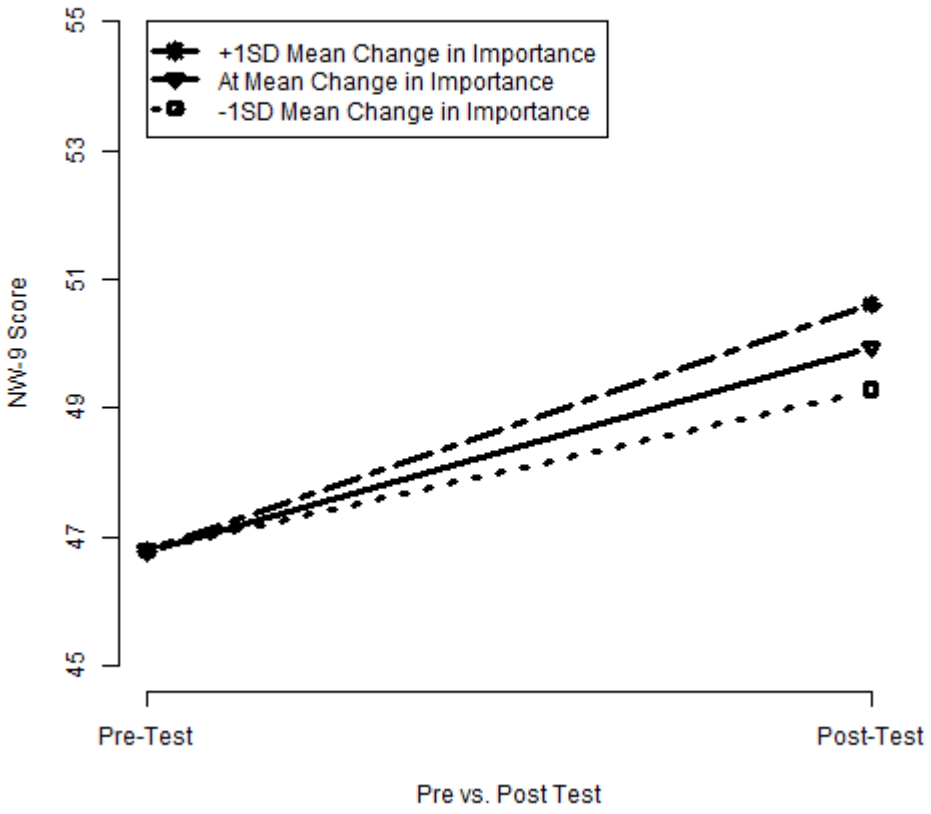

Figure 3. The effect of change in importance on HLM value-added estimates below, at, and above the mean change in perceived importance.

Table 8 contains the value-added estimates generated by the raw difference scores and the HLM estimations. The HLM estimate of value-added for students with below- and above-average change in importance fall outside the plausible range of raw difference scores. Thus, there is evidence that value-added estimates are inaccurate when *change* in importance is either below or above average. In other words, the value-added estimates are not just a reflection of quantitative and scientific reasoning skills, SAT scores, and relevant course taking experience, but are also significantly influenced by the change in how important students perceive the test to be. When value-added estimates

are downwardly biased due to negative change in perceived importance of tests, student ability is underestimated and inferences regarding quantitative and scientific reasoning skills or curriculum effectiveness based on these scores are questionable, a clear threat to validity.

Table 8
Comparison of Value-Added Estimates, Importance included in HLM (Equation 8-10)

	Value-Added Estimation Method			
	Raw Difference Score	HLM, centered 1SD below mean effort	HLM, centered at mean effort	HLM, centered 1SD above mean effort
Estimate	3.16 (95% CI: 2.68-3.64)	2.50	3.16	3.82

Research question #4: Do value-added estimates for raw difference scores and an HLM differ from one another when test-taking effort is included in the HLM? Parallel to methods used to answer Research Question #3, parameter estimates for an HLM that includes change in test-taking effort alone (Equation 5 from Chapter 3) were generated; results are shown in the column labeled "Effort Only Included" in Table 7. Only the estimation at the mean level of change in effort is included in the summary table; refer to Table 9 for estimates of β_{10} one standard deviation below, at, and one standard deviation above the mean change in effort. Again, these estimates indicate the change in NW-9 scores between pre-and post-test at each level of change in effort, after controlling for SAT math scores and number of science and math credits completed.

Similar to the model including change in importance, the model including change in effort includes a significant interaction between testing occasion and change in effort (parameter β_{12}). Because it is inappropriate to interpret a main effect (β_{10}) in the presence of a significant interaction, the equation for the model including effort was rearranged in

the same manner as the equations containing only the importance motivation variable. By substituting zero into the equations resulting from the three estimations of the model, the following final equations result:

1SD below mean *Δeffort*:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 2.42(Time_{ti}) + r_{0i} + e_{ti} \quad (11)$$

At mean *Δeffort*:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.16(Time_{ti}) + r_{0i} + e_{ti} \quad (12)$$

1SD above mean *Δeffort*:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.91(Time_{ti}) + r_{0i} + e_{ti} \quad (13)$$

Equations 11-13 now contain parameter β_{10} without the presence of an interaction and illustrate the slope for time (HLM value-added estimate) at the different levels of the change in test-taking effort over time. Figure 4 shows Equations 11-13 graphically for students who had average SAT math scores and average number of science and math credits. Similar to the results including only importance, the graph shows that although students all scored similarly at pre-test, the change in effort significantly impacted their post-test scores. Students whose change in effort decreased less (dashed line) scored higher at post-test than those students whose change in effort was average (solid line) or below average (dotted line).

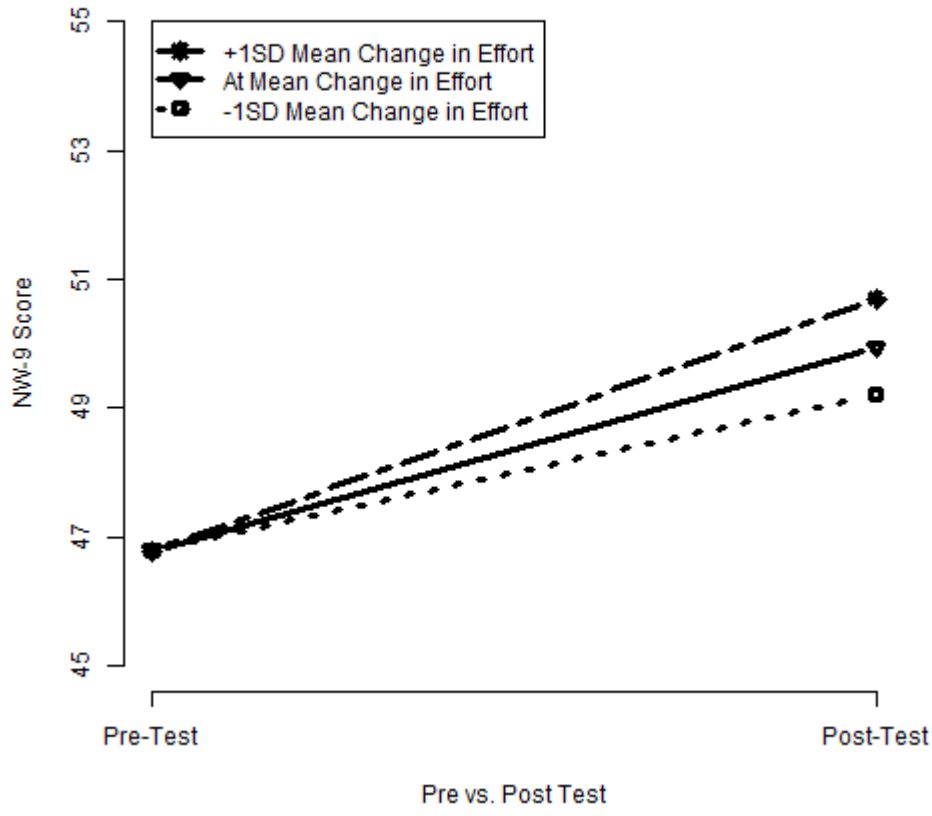

Figure 4. The effect of test-taking effort on HLM value-added estimates below, at, and above the mean change in test-taking effort.

For ease of comparison, Table 9 contains the value-added estimates generated by the raw difference scores and the HLM estimations. The HLM value-added estimates at the mean change in test-taking effort fall within the plausible range of raw difference score estimates, but the HLM estimates both 1SD below and 1 SD above mean change in test-taking effort falls outside the range of plausible raw change scores. In other words, the raw difference score and HLM value-added estimates differed from one another below and above average change in effort. Parallel to the results from the model

including only importance, these results indicated *change in* student effort over time on a test impacted NW-9 scores; therefore, inferences and decisions based on these scores reflect effort *and* ability, rather than just ability, SAT math scores and relevant course experiences. The influence of change in test-taking effort poses a significant threat to the validity of many potential inferences.

Table 9
Comparison of Value-Added Estimates, Test-Taking Effort Included in HLM (Equations 11-13)

	Value-Added Estimation Method			
	Raw Difference Score	HLM, centered 1SD below mean effort	HLM, centered at mean effort	HLM, centered 1SD above mean effort
Estimate	3.16 (95% CI: 2.68-3.64)	2.42	3.16	3.91

Research question #5: Do value-added estimates for raw difference scores and an HLM differ when both test-taking effort and perceived importance of the test, together, are included in the HLM? Parameter estimates for a model that includes both change in effort and change in importance (Equation 6 from Chapter 3) are found in the column labeled "Importance and Effort Included" in Table 7. Because nine estimations of Equation 6 were necessary to estimate β_{10} for all possible combinations of levels of change in effort and importance, only one parameter estimate (both change in importance and change in effort centered at their mean) is reported in Table 7; all nine estimates of β_{10} are found in Table 10.

Parallel to results from the models including only importance and only effort, there was a significant interaction between time and change in importance (β_{11}) and time and change in effort (β_{12}). As with the previous models, values of 0 for change in importance and change in effort were substituted into the full equations resulting from the

nine estimations. Doing so generated equations that contained values of β_{10} representing the change in NW-9 scores for different levels of change in importance and effort, after controlling for SAT math scores and number of science and math credits completed. The resulting equations were:

Both change in effort and importance -1SD:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 2.13(Time_{ti}) + r_{0i} + e_{ti} \qquad (14)$$

Change in effort -1SD, change in importance at mean:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 2.58(Time_{ti}) + r_{0i} + e_{ti} \qquad (15)$$

Change in effort -1SD, change in importance +1SD:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.03(Time_{ti}) + r_{0i} + e_{ti} \qquad (16)$$

Change in effort at mean, change in importance -1SD:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 2.71(Time_{ti}) + r_{0i} + e_{ti} \qquad (17)$$

Change in effort at mean, change in importance at mean:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.16(Time_{ti}) + r_{0i} + e_{ti} \qquad (18)$$

Change in effort at mean, change in importance +1SD:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.61(Time_{ti}) + r_{0i} + e_{ti} \qquad (19)$$

Change in effort +1SD, change in importance -1SD:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.29(Time_{ti}) + r_{0i} + e_{ti} \qquad (20)$$

Change in effort +1SD, change in importance at mean:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 3.74(Time_{ti}) + r_{0i} + e_{ti} \qquad (21)$$

Change in effort +1SD, change in importance +1SD:

$$NW9_{ti} = 46.79 + .05(SAT_i - \overline{SAT}) + .10(course_i) + 4.19(Time_{ti}) + r_{0i} + e_{ti} \qquad (22)$$

A plot of Equations 14, 15, 18, 21, and 22 is found in Figure 5; these equations reflect those instances where β_{10} estimates significantly differed from the raw difference score of value-added (Equations 14, 15, 21, and 22) along with the equation where both importance and effort are centered at the mean change in both motivation variables (Equation 18). Parallel to results from the other models that included only importance or only effort, students scored similarly at pre-test, but the amount of change in importance and effort influenced scores at post-test. In other words, change in both perceived importance and test-taking effort moderated the change in NW-9 scores over time. It is clear from the plot in Figure 5 that the more a student's change in importance and/or effort deviated from average change in importance and/or effort, the more their post-test scores differed from the raw difference score value-added estimate. A clear, intuitive, and theoretically supported pattern was observed.

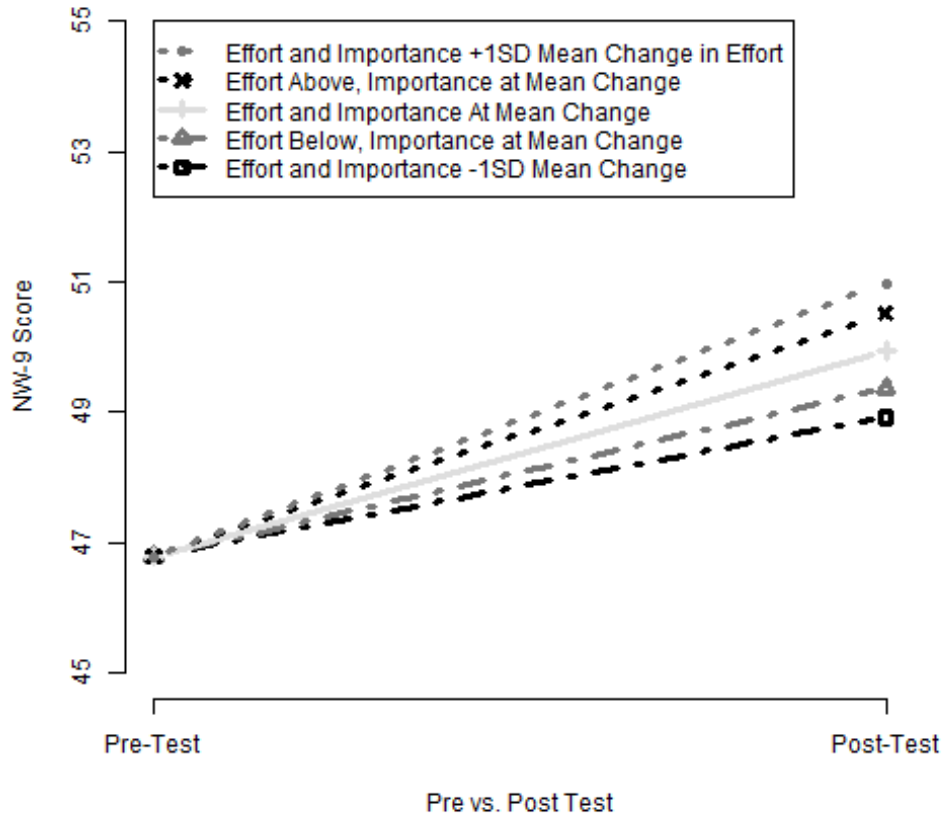

Figure 5. The effect of change in importance and change in effort on HLM value-added estimates.

Table 10 contains only the β_{10} parameters for ease of comparison with the raw difference score and its 95% CI (3.16 and 2.68-3.64, respectively). Notice that most values of β_{10} in the model including both effort and importance fell within the range of plausible values for raw difference scores, indicating that for those levels of change in importance and effort, raw difference score value-added estimates and an HLM value-added estimate including motivation were similar after controlling for SAT math scores

and science and math credits earned. Only β_{10}'s from models with extreme change in effort and/or importance fell outside the range of plausible values for raw difference scores. This provides evidence that value-added estimates are biased when above- or below-average changes in effort and importance occur, even after controlling for SAT math scores and number of science and math credits earned.

Table 10
Comparison of Value-Added Estimates, Importance and Test-Taking Effort Included in HLM (Equations 14-22)

		Importance		
		-1SD	At Mean	+1SD
Effort	-1SD	2.13*	2.58*	3.03
	At mean	2.71	3.16	3.61
	+1SD	3.29	3.74*	4.19*

* Values of β_{10} outside the plausible range of raw difference scores

Fit indices for all models are found in the summary Table 11. Using the AIC and BIC fit indices, the model including both importance and effort fit best. A likelihood ratio test of the nested models revealed that the model including only importance fit significantly worse than the full model ($\chi^2 = 6.8$, $p = .033$), while the model including only effort fit just as well as the full model ($\chi^2 = 4.1$, $p = .129$). In other words, for the current sample, it was best to include both importance and effort in value-added estimation, although including only effort would produce similar results. Because the model including only importance fit significantly worse than the model including both importance and effort, it is not recommended to only include importance in an HLM value-added model.

Table 11
Fit Indices

	Intercept-Only	No Motivation Variables (Equation 3)	Importance Only (Equation 4)	Effort Only (Equation 5)	Importance and Effort (Equation 6)
Deviance	8162.5	7819.0	7809.1	7806.4	7802.3
AIC	8168.5	7831.0	7823.1	7820.4	7818.3
BIC	8181.8	7857.6	7854.1	7851.4	7853.8

In summary, the results of this study indicated that both importance and effort moderated value-added estimates of student learning. This means that student dispositions toward test-taking negatively influenced estimates of student learning outcomes. These were important findings, and have implications for assessment practice, research, and policy. A thorough exploration of these issues as well as a discussion of the limitations of the current study and suggestions for future research follow in Chapter 5.

CHAPTER FIVE

Discussion

The results presented in Chapter 4 clearly indicate that value-added estimates are biased by test-taking motivation in low-stakes settings. For the sake of clarity, in this discussion, use of the word "bias" indicates that value-added estimates obtained through an HLM that includes motivation differ from raw difference score value-added estimates. In this chapter, I will explore the implications of these findings as well as offer suggestions for future research. First, I will briefly recap the findings from Chapter Four. Next, I will discuss implications of these findings in terms of assessment practice. Finally, this chapter will close with limitations of the current study as well as directions for future research.

Summary of Findings: Phase 1

Research Question #1 Do value-added estimates generated from raw difference scores differ when using cross-sectional vs. longitudinal data? In this phase, the equivalence of cross-sectional vs. longitudinal sampling and data analysis was explored, as well as the comparability of raw difference score and HLM value-added estimates. When comparing cross-sectional data and analyses with longitudinal data and analyses, parameter estimates and conclusions based on statistical test results and effect sizes were similar. In other words, for the institution in this study, the answer to Research Question #1 is "no"; raw difference score estimates of value-added did not differ when using cross-sectional vs. longitudinal data and analyses.

The similarity in findings between the cross-sectional and longitudinal samples could simply be an artifact of the institutional profile. The retention rate at this institution is exceptionally high, at 92.3% (James Madison University, 2014). In contrast, the average national retention rate for public four-year institutions in the United States is 79.5% while two-year institutions average 59% (National Center for Educational Statistics, 2014); the Virginia state average for four-year institutions is 78.6% and two-year institutions average 56.3% (National Center for Higher Education Management Systems, 2015). When collecting data for institutional assessment and accountability purposes, many institutions use a cross-sectional design because data are much easier to obtain. At institutions where the retention rate is not nearly as high as at the institution under study, demographics and students' abilities in a cross-sectional post-test sample would in all likelihood be vastly different from a cross-sectional pre-test sample. For example, in a cross-sectional design, students who drop out or are dismissed would be included in the first time point, but not the second. In contrast, a longitudinal sample would only include students who started and persisted at the institution. Thus, when performances of the two groups in the cross-sectional analysis are compared to one another, differences in student performance may be inaccurate due to the post-test sample only including stronger students who would persist at the institution and therefore inflate post-test scores. Cross-sectional and longitudinal results may differ more at another institution, where the longitudinal and cross-sectional samples are not as similar as they were in the current study. In addition, longitudinal sampling methods allow for inferences of *change* in learning, whereas a cross-sectional sample can only infer about *differences* in the two groups of students. This is a subtle but important distinction

between inferences that can be made with a cross-sectional vs. longitudinal design. Institutions that use cross-sectional assessment designs often make statements about *change in student learning* when the design can only legitimately support inferences about *differences in groups of students*.

Research Question #2: When using only longitudinal data, do value-added estimates generated from raw difference scores and an HLM differ? When comparing a raw difference score value-added estimate with a value-added estimate generated via an HLM that controlled for student ability (SAT math scores and number of science and math credits completed), there was no difference in the two estimates. In other words, the answer to Research Question #2 is "no"; the value-added estimates generated from raw difference scores and an HLM did not differ. Since the two value-added methods resulted in similar estimates when motivation was not included in the HLM, raw difference scores may be the preferred estimation method as they are computationally less complex. However, given the results of Phase 2 of this study, the preference for raw difference score value-added estimates should only occur if it is impossible or not feasible to measure examinee motivation.

Previous research has indicated that different value-added methods can result in different value-added estimates (Liu, 2011b; Steedle, 2012). This research compared OLS and HLM value-added methods, finding that while the estimates were correlated highly, they could still differ quite a bit from one another. However, no research had compared raw difference scores to more complex methods of value-added. This was an important area to investigate, because raw difference scores are a common metric for value added, particularly when engaging in assessment for improvement purposes. Given

that raw difference scores do not control for any other variables that might influence test performance, it was important to investigate the similarity of results. The results of the current study were reassuring in that raw difference scores were comparable to a longitudinal growth model, meaning that the less computationally complex method was acceptable. However the more concerning relationship between examinee motivation and performance was not accounted for in this phase of the study. Apprehension regarding this relationship has come up repeatedly in literature (Liu, 2011a; Liu et al, 2012) and is well worth investigation. This effect is studied and discussed in Phase 2.

Summary of Findings: Phase 2

Phase 2 was designed to investigate the effects of change in perceived importance and change in test-taking effort on value-added estimates. First, models including only change in importance and only change in effort were fit, to evaluate the individual effects of these variables. Then, a model was fit to investigate the combined impact of change in importance and change in effort. These research questions and subsequent analyses were sequenced intentionally, as previous literature has indicated importance and effort both impact performance in low-stakes testing, but in different ways. Specifically, importance theoretically impacts effort, and effort is related to performance (Knekta & Eklöf, 2014). The Phase 2 research questions investigated the individual relationships that importance and effort have with value-added estimates, as well as their combined effect. The following paragraphs will first summarize results of each of the research questions and then discuss the significance of these findings.

Research Question #3: Do value-added estimates for raw difference scores and an HLM differ when perceived importance of the test is included in the HLM?

and Research Question #4: Do value-added estimates for raw difference scores and an HLM differ from one another when test-taking effort is included in the HLM? When perceived importance (Equation 3, Chapter 3) or test-taking effort (Equation 4, Chapter 3) were included in a longitudinal HLM value-added model, the resulting estimates from the HLM differed from raw difference score estimates if change in importance or test-taking effort was below- or above-average. Thus, the answer to both Research Questions # 3 and #4 is "yes", when the change in importance and test-taking effort is either below or above the average change. Specifically, when only change in importance was included in the model, NW-9 scores for students with below average change in importance increased by 2.50 points after controlling for SAT math scores and prior science and math coursework, whereas scores for students reporting an average change in importance increased by 3.16 points. Further, scores for students reporting above-average changes in importance increased by 3.82 points. This same pattern reoccurred when change in effort alone was modeled. Students with below-average change in effort increased their NW-9 scores by only 2.42 points, compared to the 3.16-point increase for students with average change in effort, and a 3.91-point increase was obtained for students with above-average change in effort, after controlling for SAT scores and prior science and math coursework. These results indicated that both change in importance and effort singly moderated value-added estimates. Specifically, the effect of change in effort was more dramatic than the effect of change in importance. This was an interesting and important finding given that average change in importance over time, a decrease of 2.62 points, was more than 2 ½ times greater than the average change in effort, a decrease of only 1 point (both on a scale of 5-25). In other words, even though

the change in effort over time was smaller in an absolute sense, its impact on value-added estimates was greater than the change in importance. This finding aligns with developing test-taking motivation theory and previous research that indicates perceived importance has a weaker relationship with performance than effort does (Knekta & Eklöf, 2014). The larger impact of change in effort on change in performance, compared to the impact of change in importance, likely signals that there is a stronger relationship between change in effort and change in performance.

To reiterate, value-added estimates were biased for students whose reported importance or effort changed substantially more or less than the mean change over time. Although these biases were not large in an absolute sense—for example, the value-added estimate for students 1SD below the mean change in effort was less than 1 point lower than the raw change score estimate—the difference in estimates for students of differing motivation levels could have very meaningful implications. Given that the raw change score estimate was an increase of just over 3 points, even a small bias in estimates due to low examinee motivation could translate into a meaningful difference in value-added estimates and greatly impact effect sizes. To illustrate this point, Table 12 reports repeated-measures Cohen's *d* effect sizes for change in NW-9 scores at each of the levels of change in importance and change in effort discussed here. Notice that effect sizes at above- and below-average change in importance and effort do differ from the effect size at average change. These repeated-measures effect sizes were calculated using the standard deviations at pre- and post-test of the NW-9 scores. It is conceivable that NW-9 scores obtained from students with below-average changes in motivation would have even larger standard deviations that those used here, due to additional variation from

construct-irrelevant variance introduced by low motivation. Consequently, these value-added estimates and effect size differences would be even more pronounced. This could have serious implications. For example, since value-added estimates calculated based on CLA test scores are used to assign performance categories to individual institutions (see Table 1 in Chapter 2), even a small bias in scores due to examinee motivation could mean that a school's performance designation changes from, for example, "Above Expectations" to "At Expectations."

Table 12
Repeated-Measures Cohen's d of Change in NW-9 Scores for Different Levels of Change in Importance and Effort

	1 SD Below Average Change	Average Change	1SD Above Average Change
Change in Importance	.41	.52	.63
Change in Effort	.40	.52	.65

Research Question #5: Do value-added estimates for raw difference scores and an HLM differ when both test-taking effort and perceived importance of the test, together, are included in the HLM? When including both change in importance and change in effort in the HLM (see Equation 6 in Chapter 3), value-added estimates systematically differed from raw difference score estimates in four of the nine possible conditions: 1) change in both importance and effort were 1SD below mean change; 2) importance was at mean change and effort was 1SD below mean change; 3) importance was at mean change and effort was 1SD above mean change and; and 4) both importance and effort were 1SD above mean change. In other words, the answer to Research Question #5 was "yes", but only for certain combinations of change in effort and importance. These combinations make both intuitive and theoretical sense. Low motivation introduces systematic error in the form of construct-irrelevant variance into

the data, resulting in biased estimates of learning. When both change in importance and effort were 1SD above or 1SD below mean change (i.e. extreme values of change), the combined effects of change in importance and change in effort biased value-added estimates even more than the individual effects of change in importance and effort. For example, when changes in importance *and* effort were 1SD below average, student scores on the NW-9 increased by 2.13 points, after controlling for SAT and prior coursework. In contrast, when only change in importance was included in the model, NW-9 scores increased by 2.50 points, and when only change in effort was included in the model NW-9 scores increased by 2.42. As another example, when changes in importance *and* effort were 1SD above average, student scores on the NW-9 increased by 4.19 points, after controlling for SAT and prior coursework. In contrast, when only change in importance was included in the model, NW-9 scores increased by 3.82 points, and when only change in effort was included in the model NW-9 scores increased by 3.91. In other words, although both importance and effort, individually, biased value-added estimates their combined influence on value-added estimates resulted in more severe bias.

Differences between HLM and raw difference score value-added estimates were also observed when change in importance was at its mean and change in effort was either above or below average. This is clear evidence that when both change in importance and effort were modeled, the effect of change in effort moderated change in performance more strongly than change in importance. The effect is analogous to a strong magnet: change in effort essentially pulls an estimate upwards (when change in effort was above average) or downwards (when change in importance was below average), overpowering the effects of a less extreme change in motivation. If importance and effort had equal

impact on biasing value-added estimates, more symmetric results would be expected. That is, in addition to the biases observed when change in effort was above and below average and importance was at its mean, biases would also exist when change in importance was at extreme values and change in effort was at its mean. The fact that the bias was not symmetric aligns with previous research indicating effort has a stronger relationship with performance than does importance (Knekta & Eklöf, 2014). This is an important finding and should be explored more carefully in future research.

These results indicate that the change in motivation over time is a serious issue for value-added estimation, because test-taking motivation has been shown to change over time and negatively impact test scores. Given that the purpose of higher education is to influence student learning, and that institutions are struggling to show evidence of student learning (Fulcher, Good, Coleman, & Smith, 2014), the fact that value-added estimates are biased in the presence of low motivation has enormous implications for assessment and educational policy. Indeed, the claims made by Arum and Roska (2011) regarding the minimum impact institutions appear to have on student learning may simply be at least partially explained by low motivation for examinees, not the ineffectiveness of higher education. When value-added estimates are biased, as they are in this study, the validity of score interpretations is not readily clear without additional information. Decisions about curriculum or programming are tenuous at best. It is clear that assessment practitioners must not only be concerned about test-taking motivation at individual low-stakes testing occasions (e.g., pre- or post-test), but also consider how test-taking motivation may *change* over testing occasions. Modeling this change is rarely included in assessment practice, largely for logistical and design inadequacies, even

though the *Standards* recommend gathering motivation information to aid in interpretation of test results (AERA, APA, & NCME, 2014). Based on the recommendations from the *Standards* and on the results of this study, it is imperative for practitioners to model motivation over time, and attempt to mitigate the effects of low motivation when collecting and analyzing data. Motivation theory indicates that importance is a form of value (Wigfield & Eccles, 2000), and that it also predicts effort which in turn predicts performance (Knekta & Eklöf, 2014). In other words, even though effort has the larger absolute impact on value-added estimates, it is vital to influence student's perceived importance of a test as well to increase effort. Strategies for addressing these issues will be discussed in the next section.

Implications for Assessment Practitioners

The results of this study have practical implications for assessment practitioners. Practitioners should investigate whether cross-sectional or longitudinal data and analyses are interchangeable at their institutions given that the two sampling methods potentially resulting in nonequivalent results and interpretations. Practitioners should also consider and empirically explore how examinee motivation may be biasing estimates of students learning at their institutions, and contemplate ways to combat the effects of examinee motivation on value-added estimates. To address these issues, I will discuss four considerations: 1) sampling procedures; 2) measuring motivation; 3) behavioral interventions; and 4) statistical interventions.

Sampling procedures. When questions of student growth and learning are of interest, as they are in value-added modeling, longitudinal data is best (Castellano & Ho, 2013; Liu, 2011a, 2011b; Singer & Willett, 2003). At the core of value-added modeling

is a desire to measure growth in student learning and development; longitudinal data is the *only* way that such inferences can be made. Sadly, longitudinal designs appear elusive in higher education assessment designs. Although research continually emphasizes that longitudinal data is necessary for making inferences regarding growth, there is little documentation of the types of designs institutions employ to gather data for accountability purposes and even less evidence that longitudinal data is gathered at all.

At the institution under study, rigorous sampling methods are employed, resulting in high-quality true longitudinal data for estimating student learning over time. Attendance at Assessment Day is mandatory, and all students participate in pre-testing during their freshman Orientation experience. Students are invited back for post-testing after they have completed 45-70 credit hours of study, thus providing pre- and post-test data for a longitudinal sample of students. This second post-test assessment controls for student collegiate maturity in terms of credit hours earned, regardless of where and how those credits were earned. In contrast, many other institutions use convenience sampling, volunteer sampling, or pay students to participate, and rarely obtain longitudinal samples for measuring student learning. When sampling in this manner, institutions cannot be sure that their sample is truly representative of their student body, and claims about student growth may not be warranted. Further, because pre- and post-test samples may not represent the same population due to nonrandom attrition, different forms of bias may be introduced. At best, statements regarding the *difference* between knowledge at freshman and senior levels could be made, but none regarding student *growth*. In other words, if institutions wish to discuss student growth, they *must* use longitudinal data.

It is true that longitudinal data collection strategies can be costly and time-consuming due to the additional complexity of tracking students over time rather than choosing a random sample of students at the beginning and end of their academic career. However, these additional costs are well worth the added information gained from longitudinal sampling, since longitudinal data allows inferences regarding learning and growth over time. After all, what is more central to the mission of higher education than impacting student learning and *growth*? The results of this study indicate that longitudinal and cross-sectional data and analyses may be interchangeable. However, this will most likely not be the case at all institutions due to varying retention rates and key data collection design issues. If an institution wishes to substitute cross-sectional for longitudinal data and only make comparisons between students at different academic levels rather than making statements about student growth and learning, the institution should first examine the interchangeability of the two data types and analysis methods. A study similar to that conducted to answer Research Question #1 would be advisable. Even if longitudinal and cross-sectional data and analyses yield similar significance test results and effect sizes, viable inferences about student growth cannot be made, since the performance of two *different* samples of students are being compared, not the difference in performance of the *same* students at two time points. Thus, if inferences regarding student learning or growth are desirable, cross-sectional designs are inappropriate; only longitudinal data will suffice.

Measuring motivation. Another consideration for practitioners is that measuring test-taking motivation is important to assessment practice. Previous research has clearly established a link between low-stakes testing and test-taking motivation (DeMars, 2000;

Finney et al., under review; Liu et al., 2012; Sundre & Kitsantas, 2004; Wolf, Smith & Birnbaum, 1995). The results from the current study indicate that after controlling for ability and amount of prior coursework, value-added estimates from raw difference scores and an HLM were not different for students with average motivation. However, value-added estimates obtained from the two methods were different for students with below- or above-average motivation. Clearly, it is important to investigate student motivation and its relationship to performance in order to appropriately interpret results emanating from low-stakes testing conditions. The *Standards* advise practitioners to gather supplemental data in evaluation or accountability settings to better inform interpretation of test results (AERA, APA, & NCME, 2014). By simply administering a short motivation measure, such as the Student Opinion Scale (SOS; Sundre & Thelk, 2007), assessment practitioners have the information necessary to contextualize scores obtained from low-stakes testing. While the SOS is only ten items, some test administrators may be very pressed for time and may not have the space to administer all ten questions. However, operational estimates suggest that only 3-4 minutes are required for completion of the entire 10-item SOS when the items are directly added to the end of cognitive measures. This time should be considered an investment in the validity of inferences made from the data gathered and can surely be found in any testing administration.

Information regarding student motivation during testing allows practitioners to inform their interpretations regarding student learning in light of motivation levels. For example, at the institution under study, faculty in the Science and Math domain of General Education are often disheartened to see an average increase of only 3 points on

the NW-9. However, once faculty learn of the way in which student motivation has been biasing these estimates, they may be more interested in and willing to take the time to obtain and interpret measures of student motivation. This information would then allow them to make better-decisions about curriculum and resources. Further, information regarding student motivation could also strengthen the assessment culture: faculty might be more likely to endorse assessment activities and encourage students to give good effort on the tests. Administrators and policymakers would also have better data from which to make decisions regarding educational quality if they had access to measures of examinee motivation. Estimations of true learning gains could be much better approximated and reported externally. There does not appear to be any downside to collecting motivation data in low-stakes testing; the practice can only improve data quality and enhance validation of decisions based on test scores.

Behavioral interventions. While there has been research attempting to influence examinee motivation in low-stakes testing contexts, results have been mixed (Baumert & Demmrich, 2001; Finney, et al., under review; Huffman, et al., 2011; Lau, et al., 2009; Liu, et al., 2012). However, the results of this study indicate that perhaps we do not need to affect motivation in an absolute sense, but rather attempt to simply maintain motivation over time. In the current study, recall that student motivation did not decrease that much in an absolute sense. Even so, bias in value-added estimates was present. Data collection and testing procedures at the institution where the study was conducted are highly standardized and include many of the motivational strategies discussed in the literature review. At other institutions, where such practices are not regularly used, the decrease in motivation would likely be more pronounced, and value-added estimates

substantially more influenced by motivation. For these reasons, it is important to consider strategies to impact motivation, particularly in an effort to maintain motivation levels over time.

One strategy found to positively impact examinee behavior is training proctors (Lau, et al., 2009); this strategy is regularly implemented at the institution under study. As discussed in Chapter 2, prior to implementing proctor training procedures, there was significant variability in motivation across testing rooms. Assessment practitioners at the institution then attempted to identify behaviors associated with higher levels of motivation such as greeting students upon arrival, thanking students for their effort, making eye contact, actively monitoring the testing environment, and maintaining a positive testing environment. Proctors were then specifically instructed to implement these behaviors when monitoring testing rooms; as a result, student motivation increased, and variability in motivation across rooms decreased. This strategy requires very little additional time or energy for assessment practitioners to implement, as some sort of proctor training should take place before any assessment testing. Proctors also emerge with a greater sense of their impact and professional importance. The rewards in terms of data quality and student motivation far outweigh any additional time or resources it may take to train proctors on motivational strategies.

Another strategy that has some evidence of affecting motivation is educating students about the purpose of the tests (Huffman, et al., 2011; Zilberberg, et al., 2009). These studies suggested that students do not try on low-stakes tests because they are not educated about the purpose of the test or the ways in which test results are used. To this point, Huffman, et al. (2011) told students how their test results were going to be used,

for what purposes, and why putting forth their best effort on the tests was important. Student scores increased when this intervention was implemented, indicating that perhaps students took the tests more seriously and gave better effort than previous cohorts. It may be that instructions such as these need to be emphasized at post-testing in hopes of counteracting the decrease in motivation that is so ubiquitous in the literature. At the institution under study, the instructional video that students view prior to testing was recently modified to include a message from the university President. Although data regarding student motivation are not yet available to see whether this change had any mediating effect on the decrease in student motivation over time, observational evidence from the most recent Assessment Day suggests that students paid close attention to the video and were pleased to see the President appear in the video. Even though training proctors and educating students on the purpose of tests have shown promise, there is currently no research regarding how to influence *change* in student motivation over time. As such, other methods of mitigating the effect of low motivation must be considered. Statistical interventions for dealing with low-motivated students will be discussed next.

Statistical interventions. Behavioral interventions are not the only way to address the effect of low motivation on test scores. There have been suggestions in the literature regarding statistical interventions to deal with low motivation. In particular, motivation filtering has shown promise. Recall from Chapter 2 that motivation filtering is a technique that identifies and removes examinees exhibiting low motivation when taking a test. For purposes of motivation filtering, motivation can be measured either through a self-report measure (such as the SOS) or through response-time effort (Rios et al., 2014; Sundre & Wise, 2003; Wise et al., 2006; Wise & DeMars, 2010; Wise & Kong,

2005). After motivation filtering, mean performance and validity coefficients increase, resulting in more trustworthy scores (Rios et al., 2014; Sundre & Wise, 2003; Wise et al., 2006; Wise & DeMars, 2010; Wise & Kong, 2005). The assumption that motivation and ability are unrelated has been supported empirically in a number of studies (Rios et al., 2014; Sundre & Wise, 2003; Wise et al., 2006; Wise & DeMars, 2010; Wise & Kong, 2005), thus making motivation filtering a very appealing option for assessment practitioners.

There are two common methods for identifying low-motivated examinees when employing motivation filtering: self-report effort measures and response-time effort. Self-report effort (SRE) data are obtained by asking students to respond to a self-report motivation scale, such as the SOS; SRE data can be gathered when either paper-and-pencil or computer-based tests are used. Response-time effort (RTE) is a measure of how long a student spends on an item and can only be gathered when computer-based testing is used. Students identified as low-motivated have "item response times so short that examinees could not have read and fully considered the item" (Wise & Kong, 2005, p. 167). Both SRE and RTE methods of identifying low-motivated examinees are effective, but research suggests that RTE may better classify low-motivated examinees (Rios et al, 2014). Additionally, RTE provides item-level response time data that could be very useful for item development and revision purposes (Rios et al., 2014; Wise & Kong, 2005).

Motivation filtering methods are well-researched and provide compelling evidence for use when analyzing and interpreting scores in low-stakes testing contexts. Further, because increased validity coefficients are observed, motivation filtering is a

viable option for assessment practitioners to deal with construct-irrelevant variance due to low-motivated students. For these reasons, motivation filtering is recommended as a viable solution to this problem. Either SRE or RTE measures of motivation can be used to identify and remove low-motivated students from the data set prior to analyses, but given that RTE better classifies examinees, it is the preferred method. However, RTE can only be used in computer-based testing sessions, and there must be a way to capture the amount of time a student spends on the test. Luckily, many survey software platforms offer this option as part of their standard package. In the absence of resources to collect data via computer-based testing, however, SRE measures are certainly acceptable. Regardless of which method of identifying non- or low-motivated examinees is chosen, the *Standards* advise assessment practitioners to carefully document the process by which these students are removed from the data set (AERA, APA & NCME, 2014)

Limitations and Future Research

As with most research designs, there are limitations to this study. In particular, three main issues are present: 1) this study only used two time points; 2) the longitudinal design included students at the mid-point of their academic career; and 3) results may not be generalizable. All three issues will be discussed next as well as suggestions for future research.

Number of time points. First, as discussed earlier, when only two time points are used in a HLM longitudinal growth model, the slope for time cannot vary randomly. However, Figures 1 and 2 clearly indicate that students displayed different patterns of change in NW-9 scores between pre- and post-test. Further, if only two time points are

used in the model, the increase in student learning is assumed to be linear, yet there is reason to believe that learning (like other types of growth) is not a linear process (Singer & Willett, 2003). Including more than two time points is necessary to model nonlinear growth; thus, nonlinear growth could not be captured with this model. The plot of change scores for individual students shown in Chapter 4 clearly indicates that students change in different ways: some students perform similarly on pre- and post-test, some drastically increase, and some decrease in performance. If at least 3 time points were used, both the potential nonlinear patterns as well as variation in student growth could be captured by the HLM. Further, allowing the slope of time to be a random effect would allow for the value-added residual gain scores to be estimated for the actual *change* in scores, rather than just intercepts. Astin (1982) argued that the change in performance is an ideal indicator of value-added. Therefore, it makes sense to model change as accurately as possible by both correctly specifying the nature of the change (linear, curvilinear, exponential, etc.) and modeling the variability in change (random effects modeling). Implementing this design change represents a major challenge for most higher education institutions.

If collecting three or more time points and allowing the slope for time to vary randomly, the HLM used in this study could easily be expanded to a three-level model. This three-level HLM could include time points as Level 1, students as Level 2, and institutions as Level 3, and appropriate predictor variables at each level (e.g., student ability, coursework, and motivation at Level 2 and institutional characteristics such as overall SAT scores and overall motivation at Level 3). The HLM could then estimate value-added residual gain scores on a longitudinal sample for multiple institutions in

terms of change in performance over time. This modeling would represent an improvement over current methods that typically only calculate residuals on cross-sectional samples to predict performance at the exit point. Clearly, if operationalizing value-added as the change in student performance over time, a residual gain score that also models change over time is superior to a residual score that only models predicted performance at an exit point. Although there are issues with using residual gain scores to estimate value-added at the institutional level (see Chapter 2), a three-level longitudinal growth model that includes motivation as a predictor would be a substantial improvement over the current two-level cross-sectional models that do not include motivation.

Sampling at post-test. This study included students in the longitudinal sample who had taken their pre-test prior to their first semester of classes, and took the post-test three semesters later; in other words, post-testing occurred at the midway point of the college career. However, data from many students were not captured. Students become eligible to attend Assessment Day for post-testing when they complete 45-70 college credit hours. A central assessment purpose is the evaluation of the institution's general education program. Thus, 45-70 credit hours represents the true mid-point of the undergraduate career and is considered ideal by faculty and administrators. This design allows study of many students who have completed their general education requirements in one or more of the five areas and includes credit hours transferred in via AP, IB, dual enrollment, and traditional transfer credits, as well as controls for maturity. In contrast, many other institutions use convenience sampling at best, which would not allow for any of these controls to be in place.

The design employed in this study did not include a number of students that may be important to the analyses. Some students will not be included in this credit hour window for one of two reasons: 1) they have not completed enough credit hours to be eligible, or 2) because they came in with transfer, AP, or dual enrollment credits, they may skip over the credit hour window for the post-testing session. Our research indicates that most students will fall into the first category. In other words, this sample only includes students who are on the typical path to graduation and are taking the post-test at the midpoint of their undergraduate career. Some students end up taking the post-test five semesters after the pre-test while others may take it only one semester after post-test. It could be that the change in motivation for these students is different than the students who were included in this study. If that is the case, value-added estimates could also be biased because of *when* the data was collected, and differentially biased for different groups of students. For example, the change in motivation may be greater for those students who take the post-test five semesters after pre-test, and consequently their value-added estimates more biased than the sample of students included in the current study. Sample sizes are insufficient to actually answer this question. These effects need to be investigated.

In addition, because this design conducted post-testing midway through the college experience, no data was collected for seniors. This is a limitation for two reasons: 1) student performance was not measured at exit, as Astin (1982) indicated was most appropriate for institutional value-added purposes; and 2) student motivation for seniors is expected to be lower than that observed for sophomores. First, as student performance was not measured at the exit point, post-test scores may be lower than what

we would observe for the same students tested as seniors. The current design boasts considerable variability in relevant course work experience, and is therefore useful. However, if the purpose of a value-added estimate is to examine the contribution of the entire educational experience to student knowledge, then post-testing when students are seniors would be a better design than at the collegiate midpoint.

Second, student motivation as sophomores was lower than as freshmen; recall importance decreased by nearly 3 points, and effort decreased by 1 point, both on a scale of 5-25. If this trend continues, and it is realistic to think that it would, student motivation would be even lower at the senior exit time point. Based on the results of this study, a larger decrease in motivation would result in value-added estimates being even more negatively biased. This is an area in need of further investigation, particularly since this sampling design is advocated by the VSA and others.

Generalizability of results. These results may not generalize to other institutions for several reasons: 1) the culture of assessment; 2) rigorous methodology; and 3) high student motivation. The institution at which this study was conducted is a leader in assessment practice, and the conditions are likely not found anywhere else in the country. The strengths of the institutional assessment process reported in this study paradoxically represent a set of limitations that should be considered unsettling for other institutions. These three characteristics of our institution and assessment practices are perhaps the most notable and most likely causes for limited generalizability. All three will be discussed in detail next.

Culture of assessment. First, the institution under study has a long and illustrious history of rigorous assessment practice. We have been conducting Assessment Days for

nearly 30 years, with full support from the upper administration. The Orientation Office actively collaborates to ensure that students attend the pre-test sessions, and classes are cancelled for the day of the post-test data collection. The institution provides ample staff and resources to support the planning and implementation of Assessment Days, and there is a standing agreement with the Institutional Review Board to support the research conducted regarding student learning and assessment practice. Students at the institution are accustomed to assessment, and although they may complain about Assessment Day attendance, our empirical evidence indicates that they do try on the tests. Because of these factors, the data we gather are high-quality. These are conditions not easily found at other institutions, and consequently these results may not transfer easily to other contexts. Other institutions can and should aspire to high quality assessment practice to obtain data that warrants confidence.

Rigorous methodology. Unlike other institutions, 100% of students at the institution participate in Assessment Day, either at the original testing date or a mandatory make-up session. Chapter 3 provides a full description of Assessment Day procedures. Although the tests themselves remain low-stakes, students are required to attend—if a student does not attend their assigned testing session, they must attend a makeup session or a registration hold is placed on their student account. Because of the high attendance rate and random assignment of students to testing rooms, the sample of students who take any given test are representative of their student cohort. Instruments are developed by faculty who teach in the content area, are aligned to faculty-developed student learning outcomes, and mapped to curriculum. Test proctors are trained specifically in strategies to increase student motivation, are trained prior to every testing

day, and many have proctored for several years. In contrast, other institutions rely on convenience sampling, volunteers, or offer incentives for students to participate in assessment testing, and do not employ rigorous proctoring practices. These methods of obtaining participants and administering tests often result in non-representative samples and low student motivation, thus the results from the tests may not actually be representative of cohort in question.

Student motivation. Students in this sample gave fairly high effort at both time points. As discussed earlier, the culture of assessment at the institution under study is quite unique, and students understand that participation in Assessment Day is part of their responsibility as a member of the university community. Though students may grumble about the tests, they still report putting forth a great deal of effort, even at post-test: on a scale of 5-25, average effort for the longitudinal sample was 18.8 and 17.8 on pre- and post-test, respectively. For this reason, the results of this study may not generalize to other institutions. Indeed, the bias observed here would likely be more pronounced at other institutions where motivation may not be nearly as high—estimates could be biased at all levels of change in importance, effort and their combinations! Recall that importance decreased a little over 2 points and effort decreased 1 point over time in this study. In other words, relative change in both importance and effort was modest at this institution. Over a decade of applied research has informed a multitude of changes to the assessment design and practice, and has been demonstrated to mitigate the impact of low test-taking motivation, At another institution where assessment was not such an integral part of the culture and where assessment practice was not informed by such a rigorous

research agenda, student motivation may decrease more markedly, thus biasing value-added estimates much more than observed in the current study.

In the current study, *change* in motivation was the variable of interest, not motivation at pre- or post-test. As mentioned earlier, the post-test sample consisted of sophomores and juniors, not seniors. In contrast, most institutions test seniors when gathering data to measure student knowledge for accountability and reporting purposes. If seniors had been included in the current at post-test instead of sophomores and juniors, it is possible that the change in motivation would have been even larger than observed here, and consequently the bias of value-added estimates more distinct. This is an area ripe for investigation.

Future research

This study explored uncharted territory in the realm of motivation and value-added estimation, and as with most studies, it illuminates several opportunities for future research. Specifically, it is recommended that future research include: 1) replication at other institutions; 2) inclusion of multiple time points; and 3) inclusion of data from seniors.

Replication. This study only investigated the effects of test-taking motivation on value-added estimation at one institution. However, typical value-added models rely on data from multiple institutions to inform average performance across schools and consequently indicate which institutions are performing, at, above, or below expectations. These classifications are based on a residual gain score for each institution. As mentioned in Chapter 2, residual gain scores are not an ideal metric for comparing performance, as they are comprised of unexplained variance in the dependent variable.

Further, residual gain scores do not provide any diagnostic information regarding why an institution might perform above or below average (Steedle, 2012). However, residual gain scores are the current metric used in higher education value-added estimates. As such, the impact of motivation on their estimation should be further investigated.

In addition, levels of motivation as well as the relationship between motivation and performance could differ across institutions. In order to truly understand how examinee motivation impacts value-added estimates, this study should be replicated with data from multiple institutions. Given current assessment practice, this may not be feasible. In the absence of data from multiple institutions, a simulation study may be appropriate. Replication of the current study will shed further light on the impact of motivation on value-added estimates. Specifically, a replication study could illustrate how value-added estimates may fluctuate once motivation is included in an HLM residual gain score model. Of particular interest would be how estimates might fluctuate given that the relationship between motivation and performance may vary across institutions. Previous research investigated the similarity of value-added estimates when using OLS vs. HLM models, finding that results of the two methods were correlated fairly highly at .70-.84 (Liu, 2011b; Steedle, 2012). However, performance classification of institutions sometimes varied quite a bit (Liu, 2011b) and value-added estimates were more precise in the HLM framework (Steedle, 2012). Similar studies are needed to investigate the comparability of value-added estimates when motivation is and is not included in a multi-institutional HLM.

This study applied a longitudinal growth HLM to model change in student performance over time. In other words, the nesting was of time points within students.

In traditional value-added HLM's, however, students are nested within institutions, and cross-sectional samples used. Because of the differing nesting structures and use of longitudinal rather than cross-sectional data, results from this study may not be strictly comparable to those obtained from multiple institutions and cross-sectional data. As discussed earlier, using cross-sectional data for analyses such as these may result in dissimilar results when compared to results obtained from longitudinal data. In other words, the results from cross-sectional analyses may be different from those obtained in longitudinal analyses due to retained students inflating post-test scores. Further, the interpretations based on cross-sectional value-added HLM's are limited to those of differences between two groups, not growth in students over time. In order for the latter interpretation to be made, a three-level HLM should be used, with Level 1 as time points, Level 2 as students, and Level 3 as institutions. In order to fully explore the impact of change in motivation over time on student growth and learning across multiple institutions, this study must be replicated with a multi-institutional longitudinal growth model.

Multiple time points. As discussed above, using only two time points in an HLM forces a linear relationship on the data. However, since learning may not be linear (Anderman, et al., 2014; Singer & Willett, 2003; Rogosa, Brandt, & Zimowski, 1982), it is important to replicate the current study with three or more time points, (Singer & Willett, 2003). Doing so will allow the slope for individual students to randomly vary, and a non-linear model to be fit if appropriate and at least 4 time points are present (Singer & Willett, 2003). Proper model specification is one of the most important concepts in any statistical analysis in order to obtain the best, most precise parameter

estimates. Additionally, information about the actual pattern of student learning over time could help administrators better target interventions regarding student learning. This information could also help administrators temper interpretations of student growth, and not be disheartened by an increase of only 3 points from pre- to post-test, as observed in this study.

Include seniors. As mentioned in the previous section, seniors were not included in the current study. Yet, the definition of value-added indicates that students should be tested when entering and again when exiting an institution (Astin, 1982; Astin & antonio, 2012). Thus, when estimating value-added, samples of students at the beginning and end of their college careers are desirable. Further, change in student motivation was observed in the current study, which included sophomores and juniors at post-test. This change in motivation might be even larger if seniors were included at post-test rather than sophomores and juniors, and value-added estimates be more biased than observed in the current study. It is imperative that this phenomenon be investigated, particularly given that most other institutions conduct post-tests at the senior level. If motivation decreases even further than observed here, and value-added estimates are biased further, this is important information for assessment practitioners, administrators and policymakers. Replications of this study that include seniors at post-test are warranted, and imperative for understanding how value-added estimates may be affected by student motivation changes over the entirety of the college experience. This research is necessary to more fully understand the impact that motivation has on the validity of decisions made based on value-added estimates.

General Conclusions

In conclusion, there are three main considerations for practitioners and policymakers. First, value-added estimates are biased by student motivation in low-stakes testing. Second, motivation must be modeled when calculating value-added estimates—not doing so will result in model misspecification. Finally, and perhaps most importantly, value-added estimates based on cross-sectional data should *not* be used to make causal inferences about growth.

First, the results of this study consistently show that value-added estimates are moderated by student motivation in low-stakes testing. The difference observed in this study between raw difference score and HLM value-added estimates represents systematic error in estimation, not random noise. When this systematic error is the result of low examinee motivation, studies of student learning will result in an underestimation. With the increased pressure to show the value of a college education, practitioners and policymakers should consider ways to model motivation when reporting student learning. Modeling motivation should occur whether reporting learning for external purposes or for making decisions about curriculum and resources at individual institutions. It may be that the cause of the clamor regarding lack of evidence to show substantive student learning in higher education (Arum & Roska, 2011) is simply due to the non-inclusion of motivation when analyzing and reporting data regarding student learning. It is crucial for practitioners and administrators to consider the bias motivation has on value-added estimates and contemplate ways to mitigate its effects, whether through behavioral interventions or the statistical techniques described earlier.

Second, given that the results of this study indicate value-added estimates are moderated by motivation, not including motivation results in model misspecification. When model misspecification occurs, the variance associated with the non-modeled variable becomes part of the error term, or residual. In residual gain score value-added estimates, currently used for VSA, the residual is the value-added estimate. Thus, the value-added estimate contains systematic error when motivation is not modeled, and performance category classifications are influenced by that systematic error. It is imperative that when calculating value-added estimates, motivation is included in the modeling method used to produce such estimates. Raw difference scores will not suffice. The HLM used in this study is ideal for modeling motivation, as it can be used at a single institution or expanded to a three-level model for use with multiple institutions.

Finally, and perhaps most importantly, value-added estimates should not be used to make causal inferences about student learning. Factors other than student ability, teachers, or schools can influence estimates of student learning. Further, interpretations regarding institutional impact based solely on value-added estimates are causal in nature; in reality, the conditions under which value-added data are collected and analyzed do not and cannot support such inferences (Braun, 2005). In other words, the effects attributed to institutions may not actually be due to the institution itself—the results could be due to factors such as student motivation. This is a prime example of construct irrelevant variance. As such, value-added estimates need to be supplemented by other measures of quality prior to interpretation and use (ASA, 2014). Additionally, value-added estimates should be accompanied by measures of precision and interpreted in consideration of

assumptions of and limitations to the model used to generate the estimates (AERA, APA, & NCME, 2014; ASA, 2014; Steedle, 2012).

In conclusion, the present study has made some unique contributions to value-added modeling in higher education. Value-added estimates from raw difference scores and a longitudinal growth HLM are comparable when change in motivation over time is not included in the HLM. However, once the effects of change in motivation are modeled, value-added estimates from the two methods are no longer similar. Specifically, value-added estimates are biased when below- and above- average changes in importance and effort are observed. Additionally, change in effort results in a larger difference in estimates, particularly when coupled with extreme changes in importance. These findings clearly indicate that it is imperative to measure student learning by collecting longitudinal data. Doing so is the only way to model change in motivation and learning, resulting in more accurate representations of student learning. Better estimates of student learning will then result in more valid value-added estimates and interpretations made on those estimates. Institutions must continue to explore ways of dealing with low motivation in low-stakes testing, and policymakers must be educated on the ways in which examinee motivation biases test results in low-stakes settings. Future research that replicates the current study by gathering longitudinal data from multiple institutions is necessary in order to more thoroughly understand the impact of change in examinee motivation on value-added estimates.

Appendix A

The Student Opinion Scale

Please think about the test that you just completed. Mark the answer that best represents how you feel about each of the statements below.

A = Strongly Disagree
B = Disagree
C = Neutral
D = Agree
E = Strongly Agree

1. Doing well on this test was important to me.
2. I engaged in good effort throughout this test.
3. I am not curious about how I did on this test relative to others.
4. I am not concerned about the score I receive on this test.
5. This was an important test to me.
6. I gave my best effort on this test.
7. While taking this test, I could have worked harder on it.
8. I would like to know how well I did on this test.
9. I did not give this test my full attention while completing it.
10. While taking this test, I was able to persist to completion of the task.

Appendix B

Effect Size Calculations

When comparing effect sizes across different types of analyses, the effect sizes must be transformed to a common metric in order to make meaningful comparisons (Morris & DeShon, 2002). In the case of the current study, methods developed for comparing effect sizes when conducting a meta-analysis were used.

When comparing effect sizes between studies that used repeated-measures (longitudinal data) and independent groups (cross-sectional data) designs, the effect size should be converted to a common metric. The choice of metric rests in what comparisons are of interest: individual change over time (change-score metric), or differences between independent groups (raw-score metric). Additionally, "effect sizes from alternate designs will be comparable only if…standard deviations are the same or can be transformed into a common parameter" (Morris & DeShon, 2002, p. 108-109). The current study was most interested in examining individual change over time, and as a result the effect sizes were transformed into a change-score metric. Examination of descriptive statistics for each of the three data sets revealed very similar standard deviations. Morris & DeShon (2002) recommend the following transformation to convert independent-groups effect sizes (raw-score metric) into a change-score metric:

$$d_{RM} = \frac{d_{IG}}{\sqrt{2(1-\rho)}}$$

where:

d_{RM} = repeated-measures effect size

d_{IG} = independent groups effect size

ρ = correlation between pre- and post-test measures

This transformation was used to convert the Cohen's *d* calculated using the cross-sectional data set into a change-score metric that was comparable with the Cohen's *d* calculated using the longitudinal data set. The transformed effect size for the cross-sectional data is reported and interpreted in Chapter 4.

References

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

American Statistical Association (2014). ASA statement on using value-added models for education assessment. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf.

Amrein-Beardsley, A. (2008). Methodological Concerns About the Education Value-Added Assessment System. *Educational Researcher*, *37*(2), 65–75.

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education.* New York, NY: Routledge.

Anderman, E. M., Gimbert, B., O'Connell, A. A., & Riegel, L. (2014). Approaches to academic growth assessment. *British Journal of Educational Psychology.*

Arum, R., & Roska, J. (2011). *Academically adrift: limited learning on college campuses*. Chicago, IL: University of Chiago Press.

Astin, A. W. (1982). *Excellence and equity in American education*. Paper presented at the Meeting of the National Commission on Excellence in Education, Washington, DC.

Astin, A. & antonio, a. l. (2012). *Assessment for excellence (2nd ed.).* New York, NY: Rowman & Littlefield Publishers, Inc.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, XVI*(3), 441-462.

Biesanz, J. C., West, S. G., & Kwok, O. M. (2003). Personality over time: Methodological approaches to the study of short-term and long-term development and change. *Journal of Personality*, *71*(6), 905–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14633053

Braun, H. I. (2004). *Value-added modeling: what does due diligence require?* Princeton, NJ: Educational Testing Service.

Braun, H. I. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Princeton, N.J.: Educational Testing Service

Briggs, D. C. & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *American Education Finance Association, 4*(4), 384-414.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models.* Washington D.C.: Council of Chief State School Officers. [Available at http://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf]

Cohen, J., Cohen, P., West, S. G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. New York, NY: Routledge.

Cronbach, L. J. & Furby, L. (1970). How should be measure "change"--or should we? *Psychological Bulletin, 74*(1), 68-80.

Cunha, J. M., & Miller, T. (2014). Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review, 42*, 64-77.

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

Eccles, J. S., Barber, B. L., Updegraff, K., & O'Brien, K. M. (1998). An expectancy-value model of achievement choices: The role of ability self-concepts, perceived task utility and interest in predicting activity choice and course enrollment. *Interest and learning*, 267-279.

Educational Testing Service. (2008). *ETS Proficiency Profile VSA learning gains sample report.* Retrieved from http://www.ets.org/s/proficiencyprofile/pdf/learning_gains_report.pdf.

Enders, C. K. (2005). Maximum likelihood estimation. In B.S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral scienc*e (1164-1170). Chichester: John Wiley & Sons, Ltd.

Finney, S.J., Sundre, D. L., Swain, M. S., & Williams, L. M. (under review). *The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences.*

Fulcher, K, H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014). *A simple model for learning improvement: Weigh pig, feed pig, weigh pig* (Occasional Paper #23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment

Fulcher, K. H., & Willse, J. T. (2007). Value-added: Back to basics in measuring change. *Assessment Update, 19*(5), 10-12.

Hopkins-Whetstone, D. N., Swain, M. S., Williams, L.M., Finney, S. J., & Sundre, D. L. (2013). *Investigating the dimensionality of test-taking motivation across consequential testing conditions*. Poster presented at the Association for Psychological Science conference, Washington, DC.

Huffman, L., Adamopolous, A., Murdock, G., Cole, A., & McDermid, R. (2011). Strategies to motivate students for program assessment. *Educational Assessment, 16*(2), 90-103.

Hurney, C. A., Brown, J., Griscom, H. P., Kancler, E., Wigtil, C. J., & Sundre, D. L. (2011). Closing the loop: Involving faculty in the assessment of scientific and quantitative reasoning skills of Biology majors. *Journal of College Science Teaching, 40,* (6), 18-23.

James Madison University (2014). *Common data set 2013-14.* Retrieved from http://www.jmu.edu/instresrch/cds/2013/CDS2013_B.pdf

Kim, H., & Lalancette, D. (2013). *Literature review on the value-added measurement in higher education*. Retrieved from http://www.oecd.org/edu/skills-beyond-school/Litterature%20Review%20VAM.pdf

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review, 31*(5) p. 415-439.

Knekta, E. & Eklof, H. (2014). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value based questionnaire. *Journal of Psychoeducational Assessme*nt, 1-12.

Kreft, I. & DeLeeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. *Organizational Research Methods*, 9(2), 202-220.

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

Liu, O.L. (2009). Measuring learning outcomes in higher education. *R&D Connections, 10*. Educational Testing Service, Princeton, NJ

Liu, O.L. (2011a). Measuring value-added in higher education: Conditions and caveats--results from using the Measure of Academic Proficiency and Progress (MAPPTM). *Assessment & Evaluation in Higher Education, 36*(1), 81-94.

Liu, O. L. (2011b). Value-added assessment in higher education: a comparison of two methods. *Higher Education*, *61*(4), 445–461.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring Learning Outcomes in Higher Education: Motivation Matters. *Educational Researcher*, *41*(9), 352–362.

McPherson, P. & Shulenburger, D. (2006). *Elements of accountability for public universities and colleges: Discussion draft.* National Association of State Universities and Land-Grant Colleges: Washington, DC. Retrieved from http://assessment.uconn.edu/docs/resources/ARTICLES_and_REPORTS/NASULGC_Accountability_Discussion_Paper_Revised_Draft_6July2006.pdf

Miller, M. A. (2012). Demonstrating and improving student learning: The role of standardized tests. In *Seven Red Herrings about standardized assessments in higher education.* National Institute for Learning Outcomes Assessment.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent groups designs. *Psychological Methods, 7(1),* 105-125.

Musca, S. C., Kamiejski, R., Nugier, A., Mèot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology, 2*, 1-6.

National Center for Higher Education Management Systems (2015). *Retention rates--First-time college freshmen returning their second year*. Retrieved from http://www.higheredinfo.org/dbrowser/index.php?measure=92.

National Center for Educational Statistics (2014). *The condition of education (2014*). Retrieved from http://nces.ed.gov/pubs2014/2014083.pdf.

Nesselroade, J. R., & Ghisletta, P. (2003). Structuring and measuring change over the life span. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding Human Development* (pp. 317–337). New York, NY: Springer Science & Business Media.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*(3), 622-637.

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

Payne, E. (2012). *Q&A: What's behind the Chicago teacher's strike?* Retreived fromhttp://www.cnn.com/2012/09/17/us/chicago-strike-explainer/

Pike, G. R. (1992). Lies, damn lies, and statistics revisited: a comparison of three methods of representing change. *Research in Higher Education, 33*(1), p. 71-84.

Pike, G. R. (2006). Value-added models and the Collegiate Learning Assssment. *Assessment Update, 18*(4), p. 5-7.

Raudenbush, S. W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods. (2nd ed.*). Thousand Oaks, CA: Sage Publications.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: a comparison of two approaches. *New Diretions for Institutional Research, 161*, p. 69-82

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90,* 726-748.

R Core Team (2014) [Software]. R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. Available from http://www.R-project.org/.

SAS Institute (2013) [Software]. SAS 9.4. SAS Institute: Cary, NC.

Sessoms, J. (2014). Predicting change in examinee effort on low-stakes tests (Master's thesis). James Madison University, Harrisonburg, VA

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investivation of examinee test-taking effort on a large-scale assessement. *Applied Measurement in Education, 26,* 34-49.

Shulenburger, D. (2007). *NASULCG and AASCU's Voluntary System of Accountability. National Association of State Universities and Land-Grant College and the American Association of State Colleges and Universities.* Retrieved from http://assessment.uconn.edu/docs/resources/ARTICLES_and_REPORTS/NASULGC%20and%20AASCU%20Voluntary%20System%20of%20Accountability.pdf

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Socha, A., Swain, M. S., & Sundre, D. L. (2013, October). *Do examinees want their scores? Investigating the relationship between feedback, motivation, and performance in low-stakes testing contexts.* Paper presented at the annual conference of the Northeastern Educational Research Association annual conference. Rocky Hill, CT

Southern Association of Colleges and Schools Comission on Colleges (2012). *The Principles of accdreditation: Foundations for quality enhancement.* Retrieved from http://www.sacscoc.org/pdf/2012PrinciplesOfAcreditation.pdf.

State Council of Higher Education in Virginia (2007). *Guidelines for assessment of student learning.* Retrieved from http://www.schev.edu/Reportstats/2007AssessmentGuidelines.pdf?from=

Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education, 37*(6), 637–652.

Sundre, D. L. (1997, April). *Differential examinee motivation and validity: A dangerous combination.* Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*(1), 8-9.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26.

Sundre, D. L., & Thelk, A. D. (2007). *The student opinion scale (SOS), a measure of examinee motivation: Test manual*. Center for Assessment and Research Studies: James Madison University, Harrisonburg, VA.

Sundre, D. L. & Thelk, A. D. (2010). Advancing assessment of quantitative and scientific reasoning. *Numeracy, 3* (2),Article 2. http://services.bepress.com/numeracy/vol3/iss2/art2

Sundre, D. L., Thelk, A., & Wigtil, C. (2008). *The Natural World Test, Version 9: A measure of quantitative and scientific reasoning, Test Manual*. Harrisonburg, VA. James Madison University, Center for Assessment and Research Studies.

Sundre, D. L., & Wise, S. L. (2003). *'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests.* Paper presented at the National Council of Measurement in Education Annual Conference. Chicago, IL.

Swain M. S., Williams, L. M., Hopkins, D. N., Sundre, D. L. & Finney, S. J. (2013). *Investigating the (neglected) role of personality in testing.* Poster presented at the Association for Psychological Science conference, Washington, DC.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education, 58*(3), 129–151.

U.S. Department of Education (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC. Retrieved from http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

Voelkle, M. C. (2007). Latent growth curve modeling as an integrative approach to the analysis of change. *Psychology Science, 49*(4), 375-414.

Voluntary System of Accountability (2008). *Voluntary system of accountability: information on learning outcomes measures.* Retrieved from https://cp-files.s3.amazonaws.com/21/LearningOutcomesInfo.pdf

Voluntary System of Accountability (2011). *About VSA.* Retrieved from http://www.voluntarysystem.org/about

Voluntary System of Accountability (2014). *Participation Agreement.* Retrieved from https://cp-files.s3.amazonaws.com/45/VSAParticipationAgreement_final012813.pdf

Wheelan, B. S. (2014, October 24). SACSCOC information/updates [Electronic mailing list message].

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 2*5, 68-81.

Williams, L. M., Socha, A. B., & Sundre, D. L. (2013). *Cluster 3 General Education Report.* James Madison University.

Williams, L. M., Swain M. S., Hopkins, D. N., Sundre, D. L. & Finney, S. J. (2013). *Do the stakes matter? The interplay of conscientiousness, effort, and performance.* Poster presented at the American Educational Research Association conference, San Francisco, CA.

Williams, L. M. & Swanson, M. R. (2014). *But it doesn't count: A mixed-methods investigation of student test-taking motivation in low-stakes contexts.* Paper presented at the Northeastern Educational Research Association Conference, Trumbull, CT.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*(1), 59-69.

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

Wise, S. L. & DeMars, C. M. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*, 27-41

Wise, S. L. & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Wise, S., L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the National Council on Measurement in Education Annual Meeting, Vancouver, Canada

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*(1), 65-83.

Wolf, L. F., Smith, J.K. & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*(4), 341-351

Zilberberg, A., Brown, A. R., Harmes, J. C., & Anderson, R. D. (2009). How can we increase student motivation during low-stakes testing? Understanding the student perspective. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 255-277). Greenwich, CT: Information Age Publishing.