


Summer 2017

Student learning gains in higher education: A longitudinal analysis with faculty discussion

Catherine E. Mathers
James Madison University

Follow this and additional works at: <https://commons.lib.jmu.edu/master201019>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Quantitative Psychology Commons](#)

Recommended Citation

Mathers, Catherine E., "Student learning gains in higher education: A longitudinal analysis with faculty discussion" (2017). *Masters Theses*. 494.
<https://commons.lib.jmu.edu/master201019/494>

This Thesis is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Masters Theses by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.

Student learning gains in higher education: A longitudinal analysis with faculty
discussion

Catherine Elizabeth Mathers

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Masters of Arts

Graduate Psychology

August 2017

FACULTY COMMITTEE:

Committee Chair: Dr. Sara Finney

Committee Members/ Readers:

Dr. John Hathcoat

Dr. Keston Fulcher

Acknowledgments

My student learning gain exploration – on what at times seemed a veritable Everest of inter-connected assessment and measurement concepts – I would have never completed without the support of my committee, Drs. Sara Finney, John Hathcoat, and Keston Fulcher. Sara - thank you for your dedication to student learning. This thesis would not have come to fruition without your obvious enthusiasm for the project and your faith in me to do it justice (as well as your fast and keen edits). I know my own gains have exceeded both my expectations and desires under your advising these past two years. John – thank you for donning the hats of both Cluster 3 and Mixed Methods expert. There were a couple of times where a problem in either domain felt as if it eclipsed any plausible solution. I could rely on you to bring the problem into perspective and to talk through the next steps with me. As well, thank you for your sharp eye for numbers (even if it did add months of work to this thesis and subtract years from my life). Keston – thank you for your instruction on (and passion for) higher education accountability, assessment, and learning improvement. Thank you also for encouraging me to “see the bigger picture”. I additionally would like to thank Dr. Cara Meixner for her contributions to the mixed methods portion of this work. I have much to learn with regards to qualitative inquiry but am indubitably better educated than I was at the start.

My friends and family have also contributed greatly to this thesis, predominantly by keeping me sane as I researched and wrote. Aaron, Derek, Nick, and Madison – my cohort and the future members of the stat-rock band Get Data! – this adventure would not have been possible without your sass, commiseration, and friendship. CARS kids (there are too many of us to list) - CARS became a home away from home because of you (and

partly because we were here at all hours). Danielle, Mary, and Sister Teresa Joy - thank you for loving me unconditionally over the past four years and for your prayers. Danielle and Mary especially, thank you for Thursday night porch-sits and for constantly reminding me that I have a life outside my cubicle. YAMmers – thank you for friendship rooted in faith and laughter. James and Carla – I would not be treading this academic path without your love and support (and occasional strong-arming into math classes). Cara – thank you for listening to my ravings when things became tough and encouraging me to work harder and smarter. I hope I become half the academic you inspire me to be. Daniel, my (statistically and practically) significant other – thank you for hosting “work parties”, for playing the roles of chef, laundry-sorter, chauffeur, counselor, cheerleader...the list goes on. You have taught me what it means to love selflessly by example, which is, perhaps, the greatest lesson of all. To quote Billy Joel, “You’re wonderful so far and it’s more than I hoped for....” I love you.

Table of Contents

Acknowledgments.....	ii
Table of Contents.....	iv
List of Tables.....	vi
List of Figures.....	vii
Abstract.....	viii
I. Introduction.....	1
Conceptualizing and Measuring Student Learning	
Inferences about Learning Given Current Assessment Practice	
Exceptional Examples of Learning Gain Research	
Purpose of the Current Study and Hypotheses	
II. Literature Review.....	23
The Need to Assess Learning in Higher Education	
Accreditation and Financial Aid	
Two Models of Assessment	
The Importance of Using Results for Improvement	
Research Designs Used to Assess Learning	
Learning Gain Estimates	
Personal and Curriculum Characteristics Related to Learning Gains	
Test-Taking Motivation and Learning Gains	
How to Address Low Test-Taking Motivation: Motivation Filtering	
Determining an Adequate Amount of Learning Gain	
III. Methods.....	78
Participants and Procedures for Estimating Growth (Phase 1)	
Measures for Estimating Growth (Phase 1)	
Participants for Faculty Reactions (Phase 2)	
Procedures and Materials for Faculty Reactions (Phase 2)	
IV. Results.....	93
Hypothesis 1: Collapsing Across Courses, Students Should Have Moderate Gains	
Hypothesis 2: Gains Will Increase with Increased Coursework	
Hypothesis 3: Removing Unmotivated Students Will Increase Learning Gains	
Hypothesis 4: The Effort Measure Will Not Affect the Magnitude of Gain Scores	
Hypothesis 5: Coursework and Personal Characteristics Will Predict Gains	
Hypothesis 6: Faculty's Expectations Will Not Match Actual Gain Scores	
V. Discussion.....	109
Collapsing Across Courses, Students Appear to Have Moderate Gains	
Gains Did Not Increase with Increased Coursework	

After Removing Unmotivated Students, Learning Gains Did Not Increase
Test-specific and Test Session-Specific Gain Scores Are Similar
Coursework and Personal Characteristics Did Not Predict Learning Gains
Faculty’s Desired Gain Scores Did Not Match Actual Gain Scores
Limitations
Future Research
Conclusions

Appendices.....174
References.....182

List of Tables

Table 1.....	129
Table 2.....	132
Table 3.....	136
Table 4.....	138
Table 5.....	139
Table 6.....	140
Table 7.....	141
Table 8.....	142
Table 9.....	143
Table 10.....	144
Table 11.....	145
Table 12.....	148
Table 13.....	150
Table 14.....	152
Table 15.....	154
Table 16.....	155
Table 17.....	156
Table 18.....	157
Table 19.....	158

List of Figures

Figure 1.....	159
Figure 2.....	160
Figure 3.....	161
Figure 4.....	162
Figure 5.....	163
Figure 6.....	164
Figure 7.....	165
Figure 8.....	166
Figure 9.....	167
Figure 10.....	168
Figure 11.....	169
Figure 12.....	170
Figure 13.....	171
Figure 14.....	172
Figure 15.....	173

Abstract

Student learning is the primary desired outcome of a college education. To understand how educational programming and curricula affect students, colleges and universities must collect evidence of student learning gain. In this study, a longitudinal design was employed to investigate how a math and science general education curriculum impacted college students' quantitative and scientific reasoning. Quantitative and scientific reasoning gain scores were computed and predicted from personal (i.e., prior knowledge, gender) and curriculum (i.e., number of completed courses in the domain) characteristics to uncover what factors relate to learning gain. Collapsing across personal and curriculum variables, gain scores were moderate (average of 3.72 out of 66 points) with little variation and were not predicted by personal or curriculum characteristics. Disaggregating gain scores by coursework revealed that students had modest learning gains after completing one course but did not gain with additional coursework. Given performance on the quantitative reasoning test has no personal consequence for the students (i.e., low-stakes test), low examinee effort could attenuate student learning gain estimates. Therefore, gain scores and gain score predictions were estimated again after data from unmotivated students were removed (i.e., motivation filtering). Test-specific and test-session specific motivation measures were used to filter unmotivated students; results were compared to determine if they are measure-dependent. The learning gain estimates derived from using the two motivation measures were not different from each other or the unfiltered estimates. Faculty expectations of learning gain estimates were assessed. Faculty overestimated the learning gains of students with quantitative and

scientific reasoning coursework. Findings imply that students are not learning as much as expected or desired from their coursework and further investigation is necessary to explain why.

CHAPTER ONE

Introduction

A college degree is more than a piece of paper; it is a time capsule of the academic experiences intended to form students into professionals, thinkers, and leaders. Stakeholders expect these experiences to lead to positive educational outcomes for students. Specifically, students, faculty, and higher education administration typically believe university curricula should lead to gains in knowledge and skill. Scant data exist, however, to support these beliefs. Educational researchers (e.g., Ewell, 1983; 1985) and the U.S. Department of Education (U.S. Department of Education, 2006) have been calling for the collection of student learning data for decades. As Astin and colleagues noted in the mid-nineties, “As educators, we have a responsibility to the publics that support or depend on us to provide information about the ways in which our students meet goals and expectations.” (Astin et al., 1996, p. 3).

If faculty know how much or little students are learning, they may be energized to make improvements to curricula (Fulcher, Good, Coleman, & Smith, 2014). It is necessary that estimates of learning are of high psychometric quality to accurately inform curriculum modifications. Surprisingly, few institutions collect data that allow faculty to understand how much students are learning and what factors contribute to this academic growth. In this study, I estimated student learning gain across several cohorts of college students, and determined how an institution’s curriculum affected learning gain above and beyond personal characteristics (i.e., prior academic ability and gender). Additionally, faculty evaluated the learning gain estimates to determine if the estimates aligned with their expectations. Faculty also provided suggestions on how to improve

learning. The results from this study should facilitate greater understanding of learning in college and encourage a culture of learning improvement.

Conceptualizing and Measuring Student Learning

Before delving into the literature on how students' skills and knowledge are currently assessed, I clarify the distinctions between student performance, student learning gain, and learning improvement. I also discuss how learning gain should be estimated to best support inferences about student learning.

Student performance refers to knowledge and skills students have at the time of assessment. To measure student performance, practitioners collect data on proficiency at one point in time (e.g., students' math skills during the spring semester of their second year). Additional data regarding students' prior proficiency is not necessary to assess performance.

Student learning, on the other hand, refers to *change* in knowledge and skills within individuals. A positive change in proficiency is a learning *gain*. Thus, practitioners must collect data on students' prior proficiency as well as current proficiency (e.g., students' math skills during the spring semesters of the first and second year). Estimates of student performance and estimates of student learning are closely intertwined – the difference in a student's performance across multiple assessments is the student's estimated learning gain.

Student learning gains are also distinct from, yet related to, *learning improvement* (see Figure 1). Learning improvement is conceptualized as an increase in student learning gains between a cohort that experienced a modified program/curriculum and a cohort that experienced the original program/curriculum (Fulcher et al., 2014). These modifications

to improve the program are informed by previous student learning assessment results. After students have completed the modified program/curriculum, the program/curriculum is then reassessed to determine if the modifications increased student learning gains. Thus, the term ‘learning improvement’ applies to programs that have experienced effective program/curriculum modifications. The term ‘learning gains’, on the other hand, applies to students. However, these student-level learning gains may be aggregated across students participating in a particular program or who are enrolled at a specific institution. The comparison of aggregate student-level learning gains before and after program modifications inform inferences regarding learning improvement. Thus, student-level learning gains of different cohorts must be computed and assessed before and after interventions. The difference between these cohorts’ learning gains is used to determine the degree of improvement.

To assess learning gains, faculty must select the appropriate data collection and measurement (i.e., experimental) design. Longitudinal designs are most appropriate because they allow faculty to track students over time and thus obtain an estimate of learning (Castellano & Ho, 2013). In a longitudinal design, students complete the same test or psychometrically equivalent tests both before (*pretest*) and after (*posttest*) completing coursework. Faculty can then calculate the number of additional items/tasks students completed correctly to determine how much students are learning. This difference between pretest and posttest scores is known as a raw difference score, gain score, or unstandardized learning gain estimate. Faculty can use this unstandardized estimate to discuss gains in terms of the test’s metric (e.g., students, on average, gained

four points on an 80-item test)¹. The magnitude of this estimated gain can be evaluated by comparing the average gain score of students who have *not* completed the program/curriculum (i.e., comparison group) to the average gain score of students who *have* completed the program/curriculum (i.e., treatment group). Preferably, gain scores would also be compared to a predetermined faculty standard or expectation to determine if students' learning gains are sufficient.

To evaluate the magnitude of the gain scores, faculty need context regarding the tests' stakes. Large-scale, low-stakes tests are regularly used to assess students' abilities (Ewell, 2004). Students may not expend effort on low-stakes assessments because there are no personal consequences attached to poor test scores. Performance estimates (Wise & DeMars, 2005) and learning gain estimates (Finney, Sundre, Swain, & Williams, 2016; Wise & DeMars, 2010) have been shown to be attenuated by low test-taking motivation. Without correction for low test-taking motivation, faculty may come to the erroneous conclusion that students are not learning from coursework. Faculty or assessment practitioners should therefore control for low test-taking motivation to produce more valid estimates of student learning gain. These corrected estimates can then be regressed on personal and curriculum characteristics to better understand the effect of coursework on learning.

Taking into consideration these practices, I compared estimated learning gains of students with quantitative and scientific reasoning coursework to students without such

¹ The average raw difference/gain score can be divided by the estimated standard deviation of scores to produce Cohen's *d*, the standardized difference between pretest and posttest scores (Cohen, 1992). These standardized effect sizes are useful for comparing learning gain estimates computed from different tests.

coursework, after controlling for low test-taking motivation. Moreover, faculty compared these empirical learning gain estimates to their expected and desired learning gains and provided reactions.

Inferences about Learning Given Current Assessment Practice

Faculty want to infer from assessment data that students are learning from coursework. Unfortunately, the data institutions currently gather do not allow for such inferences. Institutions often simply assess student performance (U.S. Department of Education, 2006) and attempt to infer student learning from data collected using cross-sectional designs (Liu, 2011b). In these designs, a group of first-year students is typically compared to an independent group of upper-class students who have completed particular coursework. To make valid inferences about learning *gains* from this type of design, the prior academic ability (and other personal characteristics) of the upper-class group must be equivalent to the academic ability (and other personal characteristics) of the first-year group. However, this assumption, and therefore the decision to employ a cross-sectional design, may be untenable. That is, the difference between the two groups is most interpretable when this assumption is met (and the assumption is more often met by longitudinal designs). Moreover, the data to test this assumption (pretest scores for both groups) are likely not gathered. If one had the initial academic ability of the students to check this assumption, there would be no need for the cross-sectional design. Instead, learning gains could be computed for the upper-class group who experienced the coursework (i.e., a longitudinal design could be employed).

That is not to say all higher education institutions use cross-sectional designs to gauge student learning. In 2006, the U.S. Department of Education encouraged states to

collect student learning data via the Spellings Report (U.S. Department of Education, 2006). To accommodate this request, the state of Virginia briefly required its institutions to report how much they contributed to student learning and development (State Council of Higher Education for Virginia, 2007). Most Virginia institutions did so with a longitudinal design (Erwin & DeFilippo, 2010). However, little information exists on whether these institutions continue to use longitudinal designs (i.e., assess learning gain), or have reverted to cross-sectional designs (i.e., assess performance).

Although the institutions themselves may not employ longitudinal designs, researchers have investigated student learning gains using this methodology. For example, Blaich and Wise, lead researchers on the Wabash National Study, collected student learning data over a span of four years from 19 American colleges and universities (Blaich & Wise, 2011). Their results indicated that, after four years, students' estimated critical thinking gain was 0.44 standard deviations. Though the researchers measured students' critical thinking skills at the end of each academic year, they did not link these skills to critical thinking coursework. Thus, they estimated the overall effect of college on students' critical thinking.

Because students may be learning from particular coursework, or their learning gains might be influenced by other variables (e.g., maturation, out of class activities), it is imperative that faculty who claim their students are learning from particular courses connect student learning gains to this coursework. Moreover, by connecting learning gains to coursework, faculty may be better able to direct resources to courses that need improvement.

Nonetheless, connecting learning gains to coursework, albeit necessary, is not sufficient for making valid statements about how courses affect student learning. Faculty can only make limited claims given student differences on personal characteristics (e.g., prior academic ability, motivation), which often affect how or when students complete the coursework². Consequently, it is difficult to separate the effects of personal characteristics from the effects of coursework when examining learning gains³. In their book *Academically Adrift*, Arum and Roksa (2009) stated that educational researchers need to measure learning longitudinally *and* investigate the effects of both curriculum and personal characteristics on learning gains. Informing the need for the current study, the authors also remarked how few researchers were conducting such studies. A review of the literature seems to support this statement. The Wabash National Study investigated how personal and curriculum characteristics related to student learning gain, finding that prior academic ability, gender, and type of coursework (though type of coursework was not specified) moderated student learning gains (Pascarella & Blaich, 2013).

Most studies investigating the impact of curriculum and personal characteristics examine performance rather than student learning gains (e.g., Bray, Pascarella, & Pierson, 2004). Some researchers predict upper-class performance from these

² Random assignment is one experimental solution that effectively minimizes differences in personal characteristics among student groups (Shadish, et al., 2002). However, randomly assigning students to courses is hardly feasible in higher education because students enroll in courses relevant to their majors and career goals.

³ Though true experimental designs that employ randomization to control for confounds are the best methods available for making causal statements about the effects of coursework, other, albeit inferior, solutions are available. For instance, statistical modeling (i.e., regression) can be used to partition the effects of coursework on student learning gains from those of personal characteristics. This partitioning of variance does not support causal inferences unless students are randomly assigned to classes.

characteristics and then compare the predicted performance to students' actual performance; they interpret this residual as "a measure of interpretable change" (e.g., Herzog, 2011, p. 28). However, this residual score (i.e., difference between predicted and actual performance) is not a learning gain estimate. The residual score only represents how well the model with those specific predictors was able to predict actual performance. A better estimate of learning gain is the difference between posttest performance and pretest performance, which will be computed in this study. Without actual estimates of learning gains, faculty and practitioners likely cannot make valid claims about how curriculum affects student learning gains.

Exceptional Examples of Learning Gain Research

Given contemporary assessment practices, most faculty, assessment practitioners, and policy makers cannot make valid claims about how college courses influence student learning. In the section below, I describe three studies that employ designs closest to the ideal methods discussed earlier. (i.e., assess learning gains longitudinally and investigate what characteristics affect learning gains). Each study can only support limited claims about how coursework affects learning due to inadequate or absent modeling of personal or curriculum characteristics, inadequate or absent correction for low test-taking motivation, or other methodological flaws. Thus, these studies and their limitations informed the need for the current study.

Pastor, Kaliski, and Weiss (2007). Pastor, Kaliski, and Weiss (2007) estimated history and political science learning gains across five cohorts of college students. As part of the university's general education curriculum, students were required to complete two history and political science courses before graduation. However, credit for these two

courses could be obtained through Advanced Placement (AP)/International Baccalaureate (IB) or transfer credit. Given no significant demographic differences among the cohorts, the authors conducted a meta-analysis to estimate the average history/political science learning gain. Students' history/political science knowledge was assessed using an 81-item test during a university-wide assessment of general education outcomes. This test was administered once before the students began their first year of college and again halfway through their second year. The authors computed both raw and standardized difference scores⁴ and related these learning gain estimates to coursework. Specifically, Pastor and colleagues (2007) examined how much students learned after completing 0, 1, or 2 courses in the domain. Additionally, the authors investigated how completing coursework outside the university (i.e., AP/IB credit, transfer credit) affected learning gains.

After a year and a half, students who completed either the history or political science course had moderate standardized gains ($d = 0.41$ or 0.54). This standardized effect translates to an average increase of 4 points on the 81-item test. Students who completed both courses at the university had larger gains: $d = 0.90$, or an average increase of 7 points on the test. In contrast, students who received outside credit (i.e., AP/IB, transfer) had smaller learning gains ($d = 0.04$ and 0.18 , respectively). The authors postulated that these students, who scored higher on the pretest than their peers, likely had smaller gains because they already completed coursework in that domain. Thus, these

⁴ Cohen's d was computed as the raw pretest/posttest difference divided by the standard deviation of the pretest scores.

students had more history or political science knowledge and therefore less to master by posttest.

Though the authors employed adequate methodology for investigating learning gains, the study is subject to several limitations. Pastor and colleagues (2007) did not examine how personal characteristics or interactions between personal characteristics and curriculum exposure affect learning gains. The authors also did not assess the influence of test-taking motivation on gain scores. Thus, it is possible that the reported gains are actually underestimates of students' history/political science gains. The domain of interest, though not a limitation, is also a consideration. That is, history/political science learning gains may not need to be as heavily investigated as other domains. In fact, the Spellings Commission explicitly suggested more research on math and science learning gains (U.S. Department of Education, 2006). In the current study, I addressed the aforementioned limitations of Pastor et al. (2007) by assessing how students' characteristics and test-taking motivation affects these estimates in the content domain of quantitative and scientific reasoning. Similar to Pastor and colleagues (2007), though, I examined how coursework influenced learning gains across several cohorts.

Roohr, Liu, and Liu (2016). A decade after the Spellings Report and the Pastor et al. (2007) study, Roohr, Liu, and Liu (2016) investigated student learning gains across three cohorts of college students. Longitudinal data were gathered from students who completed the short-form of the ETS Proficiency Profile (EPP) in their first year of college and again after one/two years (cohort one; $N = 44$), three years (cohort two; $N = 39$), or four/five years (cohort three; $N = 85$). In other words, Roohr and colleagues conducted three longitudinal analyses, one for each cohort. As the researchers explained,

the three cohorts were convenience samples. For each of the three cohorts, the researchers estimated unstandardized and standardized learning gains in the domains of critical thinking, reading, writing, and math⁵. Unlike Pastor and colleagues (2007), Roohr and colleagues (2016) did not examine how coursework impacted learning gains. Instead, they investigated how personal characteristics affected gain scores in each cohort across the four domains. Specifically, Roohr and colleagues (2016) predicted gain scores from gender, race, STEM major status, prior academic ability, and time in college.

On the overall test (i.e., collapsing across the four domains), students' average test scores ranged from about 451 points to about 459 points. Within each domain-specific test, students' average scores ranged from about 113 points to about 123 points.

Collapsing across the domains, the researchers found that students had a gain of $d = 0.13$ after one/two years of college and an overall gain of $d = 0.61$ after four or five years of college. These standardized gains translate to raw score gains of 1.80 points and 10.88 points, respectively. With respect to domain, students made similar gains reading ($d = 0.46$ or 2.63 points after three years; $d = 0.41$ or 2.85 points after four/five years) and math ($d = 0.42$ or 2.72 points after three years; $d = 0.41$ or 2.70 points after four/five years). Roohr and colleagues (2016) found that prior academic ability (i.e., first-year GPA) statistically significantly but not practically predicted writing and reading gains (3-

⁵ Cohen's d estimates were computed by dividing the gain score by the standard deviation of the difference scores. Although desirable to compare the gains that Roohr and colleagues estimated to those from the Pastor and colleagues study, the two research teams used different standard deviations when computing d . Roohr and colleagues used the standard deviation of the difference scores, which put the effects on the gain score metric. Pastor and colleagues used the standard deviation of the pretest scores, which put the effects on the raw score metric. The two effect sizes are on different metrics; they cannot be compared. See Chapter 2 for a detailed explanation.

4% of variance explained in gains), and time spent in college statistically significantly but not practically predicted reading gains (4% of variance explained in gains). No personal characteristics statistically significantly or practically predicted math or critical thinking gains (e.g., gender explained 1% of variance in gains).

Although unclear why students differed in learning gains across years in school, one can hazard a few guesses. The difference in learning gain between students with one/two years of exposure and the other cohorts could be due to sample composition due to attrition. Students who completed the posttest two years after the pretest were not the same students who completed the posttest five years after the pretest. Thus, students in the four/five year cohort did not contain those students who left the university due to poor grades, which the one/two year cohort is likely to contain. Consequently, students in the one/two year cohort may vary more in their academic ability

It is equally likely that the difference in learning gains between cohorts is a function of maturation, coursework, or other unmeasured variables. The researchers speculated coursework may affect student learning gains. However, they examined how length of time in college, rather than curriculum, affects learning gain. Furthermore, students in this study were not randomly assigned to complete the test at different time points, which may have led to unbalanced attributes among the groups (e.g., motivation). In their discussion, they speculated that motivation may affect learning gain and recommended that motivation be examined in future research. In the current study, I examined how coursework related to student learning gains while holding length in time in college constant. Moreover, per Roohr and colleagues' (2016) recommendation, I investigated the effect of test-taking motivation on learning gain estimates.

Hathcoat, Sundre, and Johnston (2015). While Roohr and colleagues (2016) were conducting their study, Hathcoat, Sundre, and Johnston (2015) were investigating learning gains in quantitative and scientific reasoning. As part of the university's general education curriculum, students at the institution were required to complete 10 credit hours of quantitative and scientific reasoning courses. Two relatively large cohorts of students ($N = 761$, $N = 867$) were randomly assigned at the beginning of their first year to complete a 66-item quantitative and scientific reasoning test. They completed this test again halfway through their sophomore year of college. Similar to Pastor and colleagues (2007), Hathcoat and colleagues (2015) examined how fulfillment of quantitative and scientific curriculum coursework related to learning gains. They also examined estimated learning gains of students who received credit from other institutions. Although not reported in the study, the authors used motivation filtering to remove students from the sample (Hathcoat, personal communication, September 2016). This study design (sampling, assignment, and length of time) is almost identical to Pastor et al. (2007) except for the difference in content domain and use of motivation filtering.

After a year and a half of exposure to college coursework, which may have included quantitative and scientific courses, students had moderate estimated standardized gains ($d = 0.42$ or 0.67 , depending on the cohort)⁶, which corresponded to point increases of 3.13 to 3.23 points. Students who completed the 10 credit hour requirement also had moderate estimated standardized gains ($d = 0.46$ or 0.52 , depending

⁶ Unfortunately, the researchers did not specify the denominator used to compute the standardized gain estimates.

on the cohort), which corresponded to point increases of 3.49 and 2.97 points, respectively.

Estimated learning gains did not increase with additional quantitative and scientific reasoning coursework. In one cohort, students who completed the curriculum requirements (i.e., 10 credit hours) gained on average only 0.44 more points compared to those who had partially fulfilled the requirements. In the other cohort, students who partially fulfilled requirements gained on average 0.35 points more than those who had completed the curriculum.

A few methodology concerns must also be addressed. First, the authors grouped students based on credit hour completion rather than number of courses. If results from learning gain studies are used to improve curriculum, it would be simpler for faculty to know how many courses, rather than credit hours, should be required to maximize learning. Second, akin to Pastor and colleagues (2007), the authors did not examine how personal characteristics affect learning gains (e.g., prior ability, gender). The researchers examined pretest scores to detect if differences in pretest performance were due to students' prior academic abilities. Results indicated that students who received AP/IB credit came to college with higher academic ability than students with transfer credit or no credit at all. However, the researchers did not model the interactions between credit hour completion status and personal characteristics. Specifically, prior academic ability may moderate the impact of credit hour completion on learning gains (e.g., academically adept students may learn more than their non-adept peers as each group completes more courses). In the current study, I tested interactions among coursework, prior academic

ability, and gender when predicting learning gains to assess if learning gain is bivariately related to coursework or if the relationship is moderated by personal characteristics.

The last limitation in Hathcoat et al. (2015) concerns students' test-taking motivation. The level of student motivation was not reported in the published article. In a personal conversation, the first author explained that test-taking motivation data was collected and used for motivation filtering (J. Hathcoat, personal communication, September 2016). This technique entails measuring students' motivation and removing data from students with motivation scores below a set threshold (Sundre & Wise, 2003). Hathcoat explained that the filtering methods were inconsistent across cohorts. Students were filtered using test-specific motivation scores, using test session-specific motivation scores, or if they completed less than 50% of the test. In the published study, however, the authors did not report the level of test-taking motivation (e.g., was motivation low for the majority of students) or explain the filtering process. In the current study, I report the level of test-taking motivation. Test-taking motivation was measured using two motivation measures: test specific motivation and test session motivation. Scores from both measures were used to filter unmotivated students from the sample and results were compared.

Purpose of the Current Study and Hypotheses

Faculty can make more valid inferences about student learning gain and, in turn, more informed modifications to curriculum if learning gain data are appropriately collected and measured, potential moderators are assessed, and learning gain estimates are corrected for low test-taking motivation. However, documentation of appropriate measurement and informed curriculum modifications is sparse. In this study, I addressed

these issues. I estimated learning gains in quantitative and scientific reasoning for several cohorts of students. These students were randomly assigned to complete a quantitative and scientific reasoning test at the beginning of their first year of college and again after completing three semesters of college coursework. Thus, the samples represent the university population. I computed two learning gain estimates: Cohen's d estimates and raw gain scores. Cohen's d estimates from this study were compared to those from other studies (Pastor et al., 2007; Roohr et al., 2016). The unstandardized gain estimates were communicated to faculty to determine if desired or expected gains were observed.

As low test-taking motivation may bias learning gain estimates, I employed motivation filtering using scores from test-specific and test session-specific self-report motivation measures. I compared results from the unfiltered and filtered samples to determine if filtering produced different estimates of learning gain, and if these estimates were affected by choice of motivation measure. The unstandardized gain estimates from the unfiltered and filtered samples were predicted from personal and curriculum characteristics to uncover what characteristics relate to learning gain. Specifically, I predicted learning gains from gender, prior academic ability, number of quantitative and scientific reasoning courses, and the interactions of these variables.

Lastly, I discussed the learning gain estimates with faculty. I conducted interviews to assess faculty reaction to how the empirically estimated gains compared to faculty expectations of learning gains and faculty desired learning gains.

Hypothesis 1: Collapsing Across Courses, Students Should Have Moderate Gains

I predicted that, collapsing across the number of courses completed, students experiencing three semesters of college coursework on average would have moderate

learning gain in quantitative and scientific reasoning. In math, gains of $d = 0.22$ have been reported after one/two years of college, which may or may not have included math coursework (Roohr et al., 2016). In research predating 1991, gains in math and science after four years of college have been reported between 0.22 SDs to 0.41 SDs; more recent work suggests this gain is about .55 SDs (Pascarella & Terezini, 2005). However, these gains were not tied to coursework. Most recently, gains of up to $d = 0.32$ and 0.48 have been reported after three semesters of college, which may or may not have included quantitative courses (Hathcoat et al., 2015).

Given the students in this study completed a 66-item quantitative and scientific reasoning test, a moderate gain of 0.5 SD should be associated with an increase of only three items correct from pretest to posttest (Hathcoat et al., 2015)⁷. Support for this hypothesis would imply that students are learning in college, although the gain is not tied to how many courses students complete in quantitative and scientific reasoning. Therefore, testing this hypothesis had little value with respect to learning improvement. How much learning gain occurs due to specific coursework, arguably the answer most faculty and administrators want to know, requires separating learning gain estimates by coursework. This analysis is detailed below.

Hypothesis 2: Gains Will Increase with Increased Coursework

I predicted that gains in quantitative and scientific reasoning would increase as number of quantitative and scientific courses increased. Research in the domain of

⁷ In Cohen (1992), the author discusses the magnitude of effects between two independent groups. Gains of 0.2 SDs computed the within-groups standard deviation are considered small effects, gains of 0.5 SDs are considered moderate and gains of 0.8 SDs are considered large.

history/political science found that students who completed one course had moderate learning gains ($d = 0.41$ or 0.54) whereas students who completed two courses had large learning gains ($d = 0.90$). However, research in the domain of quantitative and scientific reasoning did not find this effect (Hathcoat et al., 2015). Given the incongruity between these findings, research is needed to determine how much students are learning from their quantitative and scientific reasoning courses. Thus, it is expected that learning gains will increase a small to moderate amount with each course that students complete. Support for this hypothesis would imply that quantitative and scientific reasoning coursework positively affects student learning gains.

Hypothesis 3: Removing Unmotivated Students Will Increase Learning Gains

I predicted that, after removing unmotivated students via motivation filtering, estimates of learning gains in quantitative and scientific reasoning would increase. Performance estimates have been shown to double in size when unmotivated students are removed from the sample (Wise & DeMars, 2005). However, the research on the attenuating effects of low motivation on learning gains is mixed. Learning gain estimates have been shown to increase by 0.34 SDs when data from unmotivated students are removed (Wise & DeMars, 2010). In contrast, low motivation at pretest and posttest has been shown to attenuate estimated learning gain by less than 0.25 points on a measure where students scored about 222 points on average, even though 11% of the sample was removed due to low motivation (Wise, 2015). Researchers who employed motivation filtering have reported quantitative and scientific learning gains of 0.46 SDs, corresponding to a 3-point increase on a 66-item test, after three semesters of college (Hathcoat et al., 2015). Thus, I expected smaller estimates of learning gains before

filtering and larger estimates approximating 0.5 SD after filtering. Support for this hypothesis would imply that faculty must measure and control for low test-taking motivation when estimating student learning gains.

Hypothesis 4: The Effort Measure Will Not Affect the Magnitude of Gain Scores

I predicted that learning gain estimates of students with adequate test-specific effort would be similar to the learning gain estimates of students with adequate test session-specific effort. Test-specific and session-specific motivation measures assess similar but distinct types of motivation ($r = 0.75$), with test-specific effort being slightly more correlated with test performance than session-specific effort ($r = 0.47$ and $r = 0.40$, respectively; Hathcoat et al., 2015). Test-specific motivation measures tend to identify more students as unmotivated than test session-specific measures (Hathcoat et al., 2015; Swerdzewski et al., 2011). Nonetheless, the two measures produce similar filtered performance estimates (Hathcoat et al., 2015; Swerdzewski et al., 2011). The two measures also tend to similarly classify students as being motivated or unmotivated (78.7% agreement; Hathcoat et al., 2015). Given that students appear to be equally motivated on the test and the test battery, it is likely that filtering via test-specific measure will not produce larger learning gain estimates. Support for this hypothesis would indicate that either measure may be used to make more valid inferences regarding learning gains.

Hypothesis 5: Coursework and Personal Characteristics Will Predict Gains

I predicted that coursework significantly predicts learning gains after controlling for personal characteristics. Higher education researchers investigated the effects of personal characteristics on student performance, finding that gender (Pacarella & Blaich,

2013) and prior academic ability (Wholuba, 2014) affects student performance. Prior academic ability (Grigorenko, Jarvin, Diffley, Goodyear, Shanahan, & Sternberg, 2009) and gender (Finney et al., 2016) have also been shown to affect student learning gain estimates. Fortunately, some researchers have shown that students' coursework affects student performance after controlling for prior academic ability (Bray et al., 2004). This latter result supports the premise of postsecondary education that college coursework affects student learning gains above and beyond the effects of personal characteristics. Thus, support for this hypothesis would suggest that college coursework does indeed foster student learning. On the other hand, lack of support for this hypothesis – that is, if coursework is not associated with larger learning gains – would indicate a need for learning improvement.

Hypothesis 6: Faculty's Expected Gain Scores Will Not Match Actual Gain Scores

I predicted that when discussing learning gains with faculty, faculty's expected and desired magnitude of learning gain would not align with the magnitude of empirically estimated learning gains. More specifically, I believed faculty would expect larger gains than those estimated. No research has been conducted regarding how much faculty expect students to learn from college coursework. However, research in K-12 settings have found that teachers tend to either overestimate (e.g., Rubie-Davies, Hattie, & Hamilton, 2006) or accurately estimate (e.g., Hinnant, O'Brien, & Ghazarian, 2009) student performance. One can also make predictions about the overestimation from the literature on faculty perceptions of student attitudes and behaviors. Faculty commentary on students' behaviors and performance in classrooms suggest that students are performing below expectations (Frame & Pearse, 2001). As these authors state, "Many

students don't recognize that their personal standards and perceptions of quality are well below what is expected." (p. 42).

Faculty at this university have high expectations for student competency in general education (DeMars, Sundre, & Wise, 2002). Considering quantitative and scientific reasoning competency, most students do not meet these desired competency levels (Hathcoat et al., 2015). Specifically, faculty expect that students who completed the quantitative and scientific reasoning curriculum requirements should answer 50 out of 66 items correctly at posttest, but less than 60% of students with domain-specific course exposure meet this standard. With respect to learning gains, students at the university have demonstrated 3.49 point gains on a 66-item quantitative and scientific reasoning test (Hathcoat et al., 2015) and 7 point gains on an 81-point history/political science test (Pastor et al., 2007) after completing all required coursework in the domain. Although these gains are considered moderate by *my* values, faculty with more informed opinions may not find these gains to be moderate. Thus, I expected when discussing learning gains with the faculty that they would overestimate how much their students learn--that students' actual learning gains would be less than desired by faculty.

If faculty expected learning gain were less than their desired learning gains, I believed that explanations would center on lack of student interest or motivation. In an investigation into student characteristics, researchers found that college students spend less than 12 hours per week studying and 5 hours per week preparing for their courses (Arum & Roksa, 2009). One may easily assume that college students would spend more time engaging with academic material if they were interested in it. As well, middle and high school teachers have ascribed low student learning to lack of student motivation

(Harris, 2012; Falconer-Medlin, 2014). Although these teachers work with younger student populations, it is likely that college faculty perceive these same attributes in their undergraduate students.

Addressing this hypothesis has several implications. Misalignment between faculty expectations and empirically estimated gains suggests that either more realistic expectations should be set for student learning in higher education or a need for learning improvement. Perhaps most importantly, if student learning gains are negligible, it would suggest that students are not learning from their college coursework. This finding is problematic for higher education, as it undermines the academic value of postsecondary education. If faculty observe what they consider minimal learning gains, they may be motivated to take part in the learning improvement process.

CHAPTER TWO

Literature Review

The Need to Assess Learning in Higher Education

Student learning assessment has long been discussed in higher education circles, although most higher education administration and faculty were not particularly concerned with demonstrating student learning gains to external audiences. Peter Ewell, a champion of student learning assessment, drew attention to this need for most of the 1980's (e.g., Ewell 1983; Ewell, 1985; Ewell, 1987). In fact, he had written that "Only in rare cases, however, are students typically re-tested using the same (or any) instruments to ascertain the competency achieved, or to assess the effectiveness of remediation." (Ewell, 1987, p. 15). Other notable figures in higher education assessment, such as Alexander Astin and Trudy Banta, had also attempted to impress upon their colleagues the need for both student learning assessment and data on student learning outcomes (Astin et al., 1996). Largely due to federal mandates enacted in the 2000s, greater attention from higher education administration and other stakeholders has focused on student learning outcomes assessment. In 2006, the U.S. Department of Education formed the Spellings Commission, named after U.S. Secretary of Education Margaret Spellings. The U.S. Department of Education assigned the Commission the task of investigating the status of higher education in the four areas of accessibility, affordability, quality, and accountability. Additionally, the Commission was tasked with using this information to recommend areas for improvement in higher education to the federal government. The impetus for this Commission stemmed from a number of reports generated earlier in the 21st century on the downward turn of American educational

outcomes and an absence of evidence that could explain why. The Commission's final report noted the necessity for the restructure of higher education accountability systems: the U.S. ranked 12th in degree attainment among industrialized nations, employers complained that college graduates were entering the workforce without the skills supposedly taught at universities, and evidence from the National Assessment of Adult Literacy suggested a *decline* in students' literacy abilities over time (U.S. Department of Education, 2006). Unfortunately, the systems used by universities to collect and disseminate student learning gain data were woefully inadequate to hold institutions accountable for providing quality instruction. Institutions regularly collected and reported on student competencies (i.e., performance) and other student outcomes (e.g., graduation rates), but not on students' performance *throughout* their college careers. A few researchers external to these institutions had collected student learning gain data to obtain a national perspective on student academic learning gain (e.g., Pascarella & Terezini, 2005). This aggregate data, however, could not fully capture the contributions of each institution to students' academic development. Moreover, this lack of student learning gain data resulted in little to no information to explain *why* American students were performing poorly (U.S. Department of Education, 2006). The Commission lamented the absence of reported student learning gain data, as this information was key to both holding institutions accountable for the performance of their students and initiating conversations about learning improvement. As the Commission stated, "Compounding all of these difficulties is a lack of clear, reliable information about the cost and quality of postsecondary institutions, along with a remarkable absence of accountability mechanisms to ensure that colleges succeed in educating students." (U.S. Department of

Education, 2006, p.vii). Consequently, stakeholders were left without intuition as to which institutions were most successful in teaching students.

The Commission was not the only educational body to recognize the lack of sufficient student learning gain data. The National Center for Public Policy and Higher Education (2006), a nonpartisan, higher education organization, published a “national report card” on student financial and educational outcomes. This report card, *Measuring Up*, indicated weak student learning evidence in almost all states (The National Center for Public Policy and Higher Education, 2006). In the *Measuring Up* report card series, the National Center for Public Policy and Higher Education had hoped to address weaknesses in the U.S. education system and stimulate policy changes for learning improvement (Miller & Ewell, 2005). The *Measuring Up* authors were frustrated to find current state university assessments of student learning outcomes did not enable normative comparisons of student academic abilities across states. Interstate comparisons of college student academic ability were hindered by lack of a nation-wide measure of learning on which scores could be compared, much to the consternation of the report authors. A specific model of learning assessment had been recommended in past *Measuring Up* reports that included the National Assessment of Adult Literacy (NAAL), a measure of prose, document, and quantitative literacy. Although nine states did follow the recommended model and employed either the NAAL or its state-counterpart, the State Assessment of Adult Literacy (SAAL), the other 42 states did not apply these measures. The authors of *Measuring Up* dismissed the results from these 42 states as “incomplete” assessments of college student achievement because the assessments did not follow the recommended model of learning assessment. The evidence of achievement

presented was not sufficient to address student learning gains. America, to the chagrin of both higher education practitioners and the federal government, was lax in its assessment of student learning gains (Atwell, et. al., 2006; U.S. Department of Education, 2006).

To put the U.S. educational system on track, the Spellings Commission advised the U.S. Department of Education to require institutions to empirically demonstrate student learning gains and development. American universities and colleges needed to be held accountable for how they prepared their students. Such assent from institutions was necessary to begin to reestablish the U.S. as a leader in education and to improve job and financial prospects for citizens. The Spellings Commission noted that it would be important for American universities to “embrace a culture of continuous innovation and quality improvement.” (U.S. Department of Education, 2006, p.5). As well, the Commission called for better measurement of educational outcomes and amended accountability systems to improve student learning gains, and recommended the U.S. Department of Education provide incentives for institutions that developed “outcomes-focused accountability systems” to improve programming.

American institutions had purportedly been held accountable for providing quality education, but the poor outcomes (i.e., low graduation rates, employer concerns, decrease in literacy) uncovered by the Spellings Commission called into question what occurred behind the closed doors of the academy. As stated by Ewell (2009), “Accountability requires the entity held accountable to demonstrate, with evidence, conformity with an established standard of process or outcome.” (p.7). Accreditation had long been the apparatus for accountability, and was meant to ensure the institutional quality of colleges

and universities. Something, however, was not adding up: why were accredited universities not able to empirically demonstrate their value to stakeholders?

The Spellings Commission called for a revamp of the current accreditation framework to improve the U.S. education system, stating, “Accreditation agencies should make performance outcomes, including completion rates and student learning, the core of their assessment as a priority over inputs or processes. A framework that aligns and expands existing accreditation standards should be established to...require institutions and programs to move toward world-class quality relative to specific missions and report measurable progress in relationship to their national and international peers.” (U.S. Department of Education, 2006, p.34). Accrediting agencies are the watchdogs of accountability, but as the Spellings Commission pointed out, their scrutiny of institutional quality did not necessarily include student learning gains or student progress.

Accreditation and Financial Aid

Accreditation is the multi-year, federally delegated process that requires institutions to empirically demonstrate their value to stakeholders by meeting federal, regional, and state standards of institutional effectiveness and student performance (Eaton, 2011; Council for Higher Education Accreditation, 2002). Presently, these standards require measurement of student achievement as defined at the federal, regional, and state levels. Table 1 outlines what is currently *required* for accreditation and what is *recommended* by the federal government, the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC; an accrediting body which oversees colleges and universities in the southeastern part of the U.S), and the state of Virginia. Evidence of student achievement in the form of student performance data (e.g.,

competency) is required, whereas evidence of student learning gains is largely recommended.

Currently, accredited status only ensures that students are performing at an acceptable level and implies that graduates of accredited institutions have achieved a standard level of skill (i.e., “evidence of student achievement”). For example, SACSCOC mandates that an institution provides evidence of improvement. However, this improvement could take the form of an increased percentage of students meeting the desired competency rather than a student increasing in skill from his first year to his last year. Many accreditors couch their standards in terms of improvement but are vague about what improvement means (Smith, Good, Sanchez, & Fulcher, 2015). As well, the State Council of Higher Education in Virginia (SCHEV) requires institutions to assess the value the institutions add, but do not explicitly state that the evidence should be in the form of student learning gains. As Erwin and DeFilippo describe SCHEV’s mandate, “As long as they [institutions] could demonstrate value added in accordance with the operating conception, a range of instruments and designs would be acceptable. So questions such as whether a longitudinal, cross-sectional, or residual-analysis approach would be taken were left to the institutions to settle (most institutions elected a longitudinal design).” (Erwin & DeFilippo, 2010, p. 42). Moreover, recent requirements from SCHEV emphasize an institution’s outputs (e.g., number of degree recipients, number of students enrolled) rather than the value it adds to students (e.g., SCHEV, 2013). Thus, institutions may be able to measure student learning gains but are incentivized to assess other student outcomes. Furthermore, accreditation requires institutions to document and report changes made to programs based on past assessments,

but does not explicitly require institutions to document and report program improvements. Accreditation does not ensure that students are gaining in what they know, think or can do as a function of college curricula (i.e., “evidence of student learning gains”). If an institution does not submit itself to the accreditation process, or does undergo the accreditation review but fails to meet the accrediting standards, federal financial aid is withdrawn and the institution is denied accredited status. Lack of accredited status casts the institution’s academic curriculum and value into doubt. Further repercussions include preventing credits to transfer from the unaccredited institution to any other university.

Though federal money is involved, the federal government does not accredit publically-funded higher education institutions; this job is left to a third party of national or regional accreditors. National accreditors work to ensure the academic quality of for-profit, non-degree granting higher education institutions (e.g., Advanced Technology Institute); regional accreditors assess the academic quality of non-profit, degree granting institutions (CHEA, 2002). Regional accreditors require each institution to collect and document evidence on how well it meets those standards and disseminate the results to the accrediting body. Accreditors review the report and conduct an on-site visit to determine if accreditation standards have been met. If standards are met, the institution is put on a public list and can then qualify for federal financial aid. The institution is monitored until a set date of reevaluation of status, which can range from five to ten years (Eaton, 2009; CHEA, 2002).

There are six regional accrediting bodies; each works with the institutions in its area to specify institution-level standards particular to the region and to assess both these

standards and federal requirements. Federal requirements for institutions entail the collection of data related to degree completion and student retention. Conversely, as assessment expert Michael Middaugh (2010) describes, the standards specified by the collaboration of accreditors and institutions generally fall into three categories: student learning outcomes, institutional effectiveness, and current strategic planning.

These categories are not always distinct. For example, student learning outcomes and institutional effectiveness blend in SACSCOC's standards. SACSCOC standards for accreditation require publically-funded higher education institutions within its region to meet SACSCOC "core requirements" and "comprehensive standards" as well as federal requirements (SACSCOC, 2012). Examples include facilitation of a review process for continual improvement (core requirement), the identification and assessment of student learning outcomes, identification and assessment of student competencies (comprehensive standards) and assessment of student achievement (federal requirements). The "comprehensive standards" align with both of Middaugh's (2010) "student learning outcomes" and "institutional effectiveness" categories. Delineating further, "Institutional Effectiveness", Standard 3.3 from the SACSCOC *Principles of Accreditation: Foundation for Quality Enhancement* (2012), calls for the identification, assessment, and evidence of student learning outcomes from an institution's educational programs (SACSCOC, 2012).

States also have input in how student learning outcomes are assessed. In Virginia, SCHEV works to improve the quality of the state's institutions in order to assure regional accreditation standards are met. SCHEV's initial guidelines for the assessment of student learning gains called for the documentation of student learning outcomes, as well as use

of results to improve student learning (SCHEV, 2007; see Table 1). These state standards differed from those put forth by federal and regional bodies in that they explicitly required institutions to gather value-added (i.e., learning) data. However, current state standards call for information on outputs (e.g., number of degrees granted) rather than student learning gain evidence (SCHEV, 2013).

In sum, accreditation requires evidence of both student achievement and the documentation of data used to improve student achievement. It is not explicit whether evidence of student performance or actual student learning gains should be collected and reported to accrediting bodies. Given the ambiguity, institutions must make the call on what “student achievement” evidence to report. It has been suggested that the climate of accountability plays a large role in whether institutions report competencies (i.e., performance) or evidence of actual learning gains.

Two Models of Assessment

The Spellings Commission placed student learning gains in the national spotlight by *requesting* that “Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a ‘value-added’ basis that takes into account students’ academic baseline when assessing their results.” (U.S. Department of Education, 2006, p.14). However, how this request is fulfilled is at the discretion of the institutions (e.g., SCHEV, 2007). Though accreditors have begun to develop a framework for student learning assessment (Ewell, 2009), the culture of accountability for accreditation still appears to predominantly drive assessment. When surveyed about the reasons why their institutions conducted outcomes assessment, university provosts consistently ranked accreditation as the most important reason for assessment (Kuh &

Ikenberry, 2009; 2013; Kuh, Jankowski, Ikenberry, & Kinzie, 2014). There also may be confusion as to what is sufficient evidence of student learning gain versus student performance. All regional accreditors mention student learning gains in their standards, but the standards are vague about what is sufficient evidence of student learning gains (Smith et al., 2015; Table 1).

It would be unfortunate if institutions only cared about student achievement to the extent that student achievement granted them accredited status. Fortunately, the locus of assessment for accreditation has shifted somewhat in recent years. Institutions have been moving toward a model of accountability where student learning, rather than accredited status, drives the need for assessment (Ewell, 2009; Gaston, 2013). The transition has not been smooth. Instead of a seamless shift from assessing and reporting on student competencies to student learning gains, this relatively newer line of thought has produced two assessment models: one for accreditation and one for learning improvement. The two models can operate together or independently. For example, institutions can report outcomes such as student competency and graduation rates for accreditation but internally assess student learning gains for their own purposes. Whether or not these institutions do assess student learning gains, however, is the question. The flaw in this two-model system is that one requires dissemination of information (accreditation model) whereas the other does not (learning improvement model). It is difficult to determine if institutions assess student learning gains without the type of information provided by the latter model.

Assessment practitioners seem to believe they assess student learning gains. According to the National Institute for Learning Outcomes Assessment (NILOA; Kuh et

al., 2015, p. 20), “Colleges and universities are collecting a broader range of information about student learning, and more of it, than even a few years ago... The practical challenge is to translate this growing body of information into evidence that answers pressing questions about student and institutional performance in ways that will inform pedagogical changes and policy going forward.” Nonetheless, there is little evidence that institutions actually are measuring student learning gains in addition to what is required of them. Evidence of student learning gains is necessary to make inferences about student learning; evidence of student performance does not afford the same inferences.

The Importance of Using Results for Improvement

It is disheartening that improvement of student learning is federally *recommended* but largely missing from actual institutional assessment. If student learning outcomes are not measured, or are measured but then not reported or acted upon, assessment devolves from a powerful mechanism employed to advance academic progress of students into a bureaucratic chore. Assessment is an intuitive process for progressing curricula, pedagogy, and, in turn, student learning (Fulcher al., 2014). What if the federal government or regional accreditors required institutions to report on student learning gains? Ostensibly, it cannot be assumed that *all* methods used by *every* American institution can capture student learning gains. The manner in which student learning outcomes are assessed directly affects the validity of the inferences made about curriculum effectiveness (SCHEV, 2007). By measuring and reporting estimates of student learning gains, practitioners have necessary (yet not sufficient) data to both identify weaknesses in the curriculum and enact solutions to strengthen these flaws (Ewell, 2009; Fulcher et al., 2014; Kuh & Ikenberry, 2009).

It is likely that assessment data presently are not collected nor analyzed in a way that supports the measurement or use of student learning gain data. For instance, an institution may collect critical thinking data from its graduating class. This information conveys little about how students *developed* their critical thinking skills during their tenure at the university. On the other hand, tracking this group of students throughout their years of study would allow the institution to see the progression of critical thinking. As outlined in Table 2, however, learning gains can be conceptualized in a myriad of ways. Different data collection designs and methods of measurement correspond with certain conceptualizations of “learning gain”. Thus, the way in which “learning gain” is defined dictates the appropriate research design and method of measurement.

Research Designs Used to Assess Learning

Generally speaking, multiple research, or experimental, designs are available to collect data. However, not all designs are appropriate for higher education settings. For instance, the pretest/posttest control group design, a “true experimental” design, is considered to be one of the more methodologically sound experimental designs. Though practitioners may hope to employ this design in order to make valid inferences about student learning, it is not well-suited for applied settings for reasons elaborated in the sections below. Data collected to make inferences about student learning can be measured using one of several other designs: a nonequivalent comparison group design, a separate sample pretest/posttest design, one-group posttest-only design, a one-group pretest/posttest design, and a static-group comparison design. In that vein, the type of design applied determines whether inferences can be made about student performance, student learning gain, or both. The designs listed above are conceptually distinct and

provide different estimates of “learning” outcomes. Inferences about student outcomes are tied to these estimates and are therefore tied to the research design employed. In the sections below, I describe best practices to control for validity threats. I then describe each design, the research questions each answers, and, if applicable, what can be inferred about student learning based on results. I also discuss the pros and cons associated with each design.

Best practice for good designs. Assessment practitioners must understand which experimental designs enable correct inferences about program or curriculum effectiveness; only certain designs afford causal inferences about how the curriculum affects student learning gains. Best practice necessitates that threats to both external and internal validity are controlled (see Table 3 for descriptions of these threats). *External validity* refers to the accuracy of generalizations made from results (Dawson, 1997). In higher education, one may aim to generalize assessment results from the measured sample of students to all students at the university. To achieve some degree of external validity, the researcher must obtain representative samples of the population. Random sampling is the best method of achieving this outcome. When sampling is random, each individual in the population has an equal chance of being selected for participation in the study (Shadish et al., 2002). Thus, responses from the sample should reflect those from the population. If the sample represents the population, the outcome likely reflects what occurs in the population. In other words, the inferences made from these responses about the population are externally valid. In the context of higher education, if a practitioner randomly samples from students at the university and assesses that sample, the

distribution of test scores from these students will be similar to the distribution of test scores from all students at the university.

When a sample is not representative of the population, the estimate derived from the sample is biased. This bias is termed ‘sampling error’, because the estimate is “off” from what it would have been if the sample was representative of the population.

Analyzing unrepresentative samples can lead to less externally valid inferences (Shadish et al., 2002). For example, if an assessment practitioner administers a science test to a group of males at a predominantly female institution, she may make less valid inferences about students’ science knowledge at the institution (assuming, of course, that males and females are from different populations). However, random sampling alone is not enough to create experimental conditions appropriate for making such desired inferences. Several common threats to external validity are described in Table 3; the researcher should try to minimize these threats as much as possible.

Internal validity refers to the accuracy of inferences made about the causal effects of a treatment on an outcome (Shadish et al., 2002). Extending the example from above, suppose the assessment practitioner is interested in whether or not the students’ science coursework increases their science knowledge. Thus, she will need to ensure that coursework is the only experience that would affect students’ scores on the test. Random assignment of participants to experimental groups is used to improve internal validity. When random assignment is used to place participants in either the treatment or the control group, each individual has an equal chance of being assigned to either group. Random assignment distributes individuals between the groups in such a way that each group should be evenly matched on all variables (e.g., gender, ability, personality),

including those related to the outcome that may not be assessed during the experiment. In other words, equivalent groups are formed by the dispersion of preexisting differences; this dispersion is why cross-sectional estimates can approximate longitudinal estimates. Thus, researchers are able prevent, to an extent, confounding variables from differentially influencing the outcome of a study (e.g., from influencing student learning gains). By evenly dispersing individual differences, researchers can then infer that differences in the outcome across groups are driven by treatment (e.g., curriculum, programming, pedagogy) and not by other variables.

Despite these approaches, random sampling and random assignment cannot account for other threats that may compromise either external or internal validity (see Tables 2 and 3). However, the data collection design that is chosen determines which of these other threats affects desired inferences. Below, I describe the designs available and their strengths and weaknesses with respect to validity.

True experimental and quasi-experimental designs. Experimental designs tend to fall into one of two categories: true experimental or quasi-experimental (Shadish et al., 2002). True experimental designs isolate the treatment effect by controlling for all alternative explanatory variables through random assignment of students to treatment and occasionally through random sampling. Further controls may be employed through the use of a control group, where students are randomly assigned to *not* receive the treatment. Results obtained from this control group can then be compared to the results from the treatment group. Quasi-experimental designs, in contrast, do not involve random assignment of students to treatment. In the experimental design literature, quasi-experimental designs are described as “experiments that lack random assignment of units

to conditions but that otherwise have similar purposes and structural attributes to randomized experiments.” (Shadish et al., 2002, p. 104). Thus, these designs control for some, but not all, alternative explanatory variables. Quasi-experimental designs are common in applied settings where not all explanatory variables can be controlled or manipulated. To control for the explanatory variables that the researcher can manipulate in quasi-experimental designs, control groups are usually (although not always) formed.

The section below describes common true experimental (pretest/posttest control group design) and quasi-experimental (nonequivalent comparison group design, separate sample pretest/posttest design, one-group posttest-only design, one-group pretest/posttest design, posttest only design with nonequivalent groups), and how they relate to higher education assessment.

True experimental: Pretest/posttest control group design. One particularly powerful data collection design for making desired inferences is the pretest/posttest control group design. This design is longitudinal in nature, and is also referred to as a within-subjects design or repeated-measures design. To make inferences about the effectiveness of curriculum or educational programming, “pretest” scores on the outcome of interest are often gathered prior to experiencing the programming and “posttest” scores are often gathered upon completion of the programming (Campbell & Stanley, 1963; Shadish et al., 2002). The validity of inferences is further improved when pretest and posttest scores are collected and compared for a sample that experienced the programming (treatment group) and a sample that did not (control group). In general, the measurement of an individual or a sample of students at two time points allows researchers to estimate the learning gain for that particular individual or sample. Thus, in

higher education contexts, a researcher who employs a longitudinal or pretest/posttest design can answer the question, “How much do students change, or gain, from time one to time two?” With two groups, the design also answers, “Do students who experience the curriculum learn more than students who do not?”

The design can be conceptualized as follows:

R: X1pre T X1post

R: X2pre X2post

“R” designates that the samples were randomly assigned to receive or not receive the treatment. “X1pre” is the measurement of Group 1 before receiving treatment or curriculum “T”. “X1post” is the measurement of Group 1 after receiving treatment “T.” Note that Group 2 (i.e., control group) is assessed twice with a pretest (“X2pre”) and posttest (“X2post”) but does not actually receive the treatment.

Pros. True experimental designs such as the pretest/posttest control group design are ideal because they suffer relatively few threats to internal validity. Thus, researchers are in a position to isolate the treatment effect from possible confounds. Random assignment makes this design powerful with respect to valid inference about curriculum effectiveness. By randomly assigning students to groups, practitioners are able to produce two groups of students that are equivalent, or balanced, on the variables that affect the studied outcome. By gathering data at multiple time points for both samples, practitioners are able to empirically demonstrate change in ability over time and compare change in ability across groups (Shadish et al., 2002).

Another strength of longitudinal designs in general is that each individual serves as her own control within each sample or group (Porter, 2012). That is, variations that naturally occur across groups (i.e., background characteristics, differences in academic experience) do not exist within groups. The aforementioned variations are held constant over time for each individual in each group (Zumbo, Wu, & Lui, 2012). Therefore, this design eliminates variability associated with individual differences and increases statistical power, which is the capability of detecting an effect that exists (Shadish et al., 2002). According to Witte (1993), "...the variability within groups reflects only random error, that is, the combined effects (on the scores of individual subjects), of all uncontrolled factors, such as individual differences among subjects, slight variations in experimental conditions, and errors in measurement." (p. 339). Consequently, the practitioner who employs this design may more accurately assess students' learning gains.

Cons. The pretest/posttest control group design *theoretically* can be used to compute learning gain across two time points (e.g., before and after experiencing curriculum) in higher education contexts. This design, however, requires random assignment of students to specific courses or course sequences, which can be unethical if students are unaware of this practice or do not consent. Unsurprisingly, this kind of random assignment is not done in practice. For example, higher education administrators cannot randomly assign students to complete certain courses or course sequences; students complete coursework based on their interests and academic schedules. Thus, true experimental designs are difficult to implement in university settings. Consequently, higher education practitioners and researchers may rely on quasi-experimental designs.

Although unaffected by internal validity threats, the pretest/posttest control group design is subject to several external validity threats (see Table 3 for examples of validity threats in higher education contexts). One such threat is the interaction of testing and treatment, where the pretest that the participant completes affects how he respond to the treatment (Campbell & Stanley, 1963; Dawson, 1997). Another threat is *reactive arrangements*, where participants attempt to produce behavior they believe the researcher wants to see. Lastly, the interaction of selection bias and treatment may also affect external validity. If participants in the treatment group differ from those in the control group even after random assignment, there is a chance that participants in one group will react differently to the treatment than the other.

Quasi-experimental: Nonequivalent comparison group design. This design is also longitudinal because the same sample of students is measured at “pretest” and at “posttest” (Liu, 2011b). Practitioners can use the nonequivalent comparison group design to compare student performance estimates and student learning gain estimates, the latter of which can be computed across months in college or prior to and after coursework. In educational contexts, such designs may also be referred to as gain score models because they produce an estimate of a student’s learning gain (Castellano & Ho, 2013). The design can be conceptualized as follows:

$$X_{1pre} \quad T \quad X_{1post}$$

$$X_{2pre} \quad X_{2post}$$

“X_{1pre}” is the measurement of Group 1 before receiving treatment, or curriculum, “T”. “X_{1post}” is the measurement of Group 1 after receiving treatment “T”. Group 2 is also

assessed with both a pretest (“X2pre”) and posttest (“X2post”) but does not actually receive the treatment. Notice that individuals are not randomly assigned to groups; the researcher is measuring groups that are already formed.

Pros. By gathering data at multiple time points for both samples, or groups, practitioners have some empirical evidence to demonstrate change in ability and compare the change between groups (Shadish et al., 2002). These learning gain estimates provide more information about student progress than performance estimates (Gong, 2004; Zvoch & Stevens, 2006). For instance, students may gain at above average rates even if average performance levels are low at posttest (Zvoch & Stevens, 2006). A student who raises her class grade from an F to a C- over the course of a semester may not be considered proficient in the subject matter but has grown substantially. Clearly, this student is learning, even if she is not performing well at posttest; longitudinal designs enable researchers to observe this effect.

By having a control group, practitioners can compare the learning gain estimates from both samples to make inferences about the effectiveness of the curriculum. Though practitioners cannot eliminate maturation from affecting either sample’s results, comparison of the two groups prevents maturation from affecting conclusions made about the curriculum. The effect of maturation on both samples’ learning gains estimates should be equivalent because both samples are maturing at the same rate. This design also enables practitioners to calculate the relationship between learning gain estimates and curriculum (e.g., Pieper et al., 2008). For instance, practitioners can collect learning gain data from samples that have taken one course, two courses, three courses, etc. in the

curriculum. The correlation between coursework and learning gain estimates can then be calculated to determine the relationship between the two variables.

Cons. Quasi-experimental designs sacrifice some evidence of internal validity evidence for experimental feasibility (see Table 3). Regression to the mean, where participants who initially score highly on a pretest achieve a lower score on the posttest, is a concern in quasi-experimental settings, especially when participants are selected based on extreme scores. This outcome, though, is natural and not due to a negative treatment effect. For example, students who score highly on a math placement pretest complete a posttest after their coursework. However, the posttest scores of the high-scoring students are closer to the posttest scores of their peers than before. These students' coursework did not negatively affect their learning gains, though one might try to make such a claim; the decrease is merely a statistical artifact. The interactions of typical internal validity threats (see Table 3) are also likely. Additionally, lack of random assignment to groups limits the inferences practitioners can make about learning gains and curriculum effectiveness. Individuals in each group are likely unequal on all variables if individuals are not randomly assigned to groups. The difference between the estimated average learning gains across groups may be driven by variables that affect learning gain other than curriculum. If a practitioner is able to randomly assign individuals to either receive the curriculum or not *and* measure the outcome both before and after the experiencing the curriculum, she has powerful evidence about student learning and the curriculum's value. Without random assignment, it is especially important that this evidence is interpreted in context.

External validity threats are a concern as well. As with true experimental designs, an interaction effect of testing is one threat that may affect generalization. Another limitation is that the nonequivalent comparison group design, like most applied longitudinal designs, is susceptible to attrition, or mortality (Campbell & Stanley, 1963; Klein, 2010; Pieper et al., 2008; Shadish et al., 2002). Students who complete the pretest do not always complete the posttest (e.g., students drop out of the school). The sample size is reduced if the researcher decides to analyze data only from students who have completed both tests. Analysis of the smaller sample is not problematic if the reduced sample is representative of the student body, but attrition hardly scales down samples so favorably. Differential attrition negatively affects the principle of balanced groups that is inherent in random assignment (Shadish et al., 2002). More often, students who have completed the pretest and posttest are stronger academically and have higher test scores. Learning gain estimates produced from this sample are upwardly biased; results are therefore sample dependent and would not generalize to all university students.

Beyond validity threats, several other limitations exist. Longitudinal models used for estimating student learning gain can quickly become complex for practitioners without a statistical background (Gong, 2004). Pretest/posttest designs are also less frequently employed than cross-sectional designs, in part because they can be costly to implement (e.g., collecting data over time for multiple groups or employing sophisticated analyses that require consultation; Seifert et al., 2010).

Quasi-experimental: Separate sample pretest/posttest design. Another quasi-experimental design is the separate sample pretest/posttest design. The name of the separate sample pretest/posttest design is slightly misleading. It is not a longitudinal

design but is a cross-sectional, or posttest, design. Cross-sectional designs (i.e., independent-samples designs, between-subjects designs) enable comparisons of performance estimates from two different samples of students (Castellano & Ho, 2013; Gravetter & Wallnau, 2009). A cross-sectional design serves to answer the question: “What is the average *difference* between Sample 1 and Sample 2?” When measuring performance for higher education accountability purposes, this question can be reframed as, “How does student performance differ, on average, between incoming students who have yet to complete the curriculum and upper-class students who have completed the curriculum?”

The design can be conceptualized as follows:

$$\begin{array}{cc} X1_{pre} & T \\ & T \quad X2_{post} \end{array}$$

“X1pre” is the measurement of Group 1 (the comparison group) before experiencing treatment “T”. “X2post” is the measurement of Group 2 (the treatment group) after experiencing treatment “T”. In higher education contexts, “X1pre” often refers to the measurement of first-year students and “X2post” often refers to the measurement of upper-class students. Because the two groups are measured at the same time, entering or first-year students who complete the “pretest” are not the same as the upper-class students who complete the “posttest” (Liu, 2011b). The assumption underlying this design is that if students are learning at an institution due to the curriculum they complete, average performance for students who have completed the curriculum should be greater than average performance for students who have not.

Pros. If differences in background characteristics are controlled through random assignment, the difference between the two performance estimates can approximate a learning gain estimate. That is, the two samples are likely equivalent on all variables related to the outcome. Moreover, with the separate samples pretest/posttest design, a test that measures desired student learning outcomes (e.g., quantitative reasoning) can be administered to both samples in the same academic year (Liu, 2011b). Administrators can then make relatively immediate comparisons between first-year and upper-class students.

This design is frequently employed to investigate learning outcomes and to make institutional comparisons (Klein et al., 2007; Klein, 2010); it is an easy and relatively cheap design that can be used by any institution (Liu, 2011; SCHEV, 2007). Similar to actual pretest/posttest designs (i.e., longitudinal designs), practitioners can calculate the relationship between curriculum and performance. Specifically, practitioners can calculate the correlation between the number of courses completed and performance estimates. The outcome of interest, the performance difference between the two cohorts, is simple to compute; the average performance score or estimate of one cohort is subtracted from the average performance of another cohort (Gong, 2004). However, it is important to keep in mind that this design produces performance estimates, *not* a learning gain estimate.

Cons. This design is subject to multiple internal validity threats. In higher education, it is expected that performance estimates for the two groups (e.g., first-year students vs. upper-class students) are different because one sample of students experienced the curriculum and the other did not. However, differences between the samples in other constructs related to the outcome of interest, such as intelligence or

motivation, may drive the difference in performance. Students opt into courses and other activities; the institution does not assign students to these academic experiences. If students are not randomly assigned to groups, the difference between the performance estimates is difficult to interpret (Porter, 2012). Nevertheless, the separate samples pretest/posttest design does not use randomization. Consequently, the two groups have not similar distributions of background characteristics. Though possible that performance is affected by academic experiences, the conclusion that curriculum exposure caused the difference is not sound.

Maturation effects are also a concern. Students who experience the curriculum (i.e., the upper-class students) will be systematically older than the students who have yet to experience the curriculum. Attrition affects are another concern; these upper-class students are likely more academically adept than the cohort of first-year students due to attrition. To elaborate, upper-class students may appear to have higher performance than first-year students because students with lower ability drop out of the university before achieving upper-class status. Thus, the upper-class performance estimate is based upon only those retained students and is therefore not representative of the student population. In contrast, the sample of first-year students analyzed includes both the students who will persist through college to their final year as well as the students who will not, thus more accurately reflecting the college student population. History effects, where events prior to participation impact the outcome, are an additional problem. For example, students who complete AP Calculus prior to being tested on college math proficiency and completing math courses at college will likely perform better on the test and in the classes. Instrumentation effects may be a problem if the pretest differs from the posttest. If the

test scores are not on the same metric (e.g., the pretest is more difficult than the posttest), incorrect inferences may be drawn about student ability and student learning gain.

Other limitations with this design are also present. It may be tempting to use terminology such as ‘pretest’ and posttest’ to describe the tests administered to the pre-treatment and post-treatment samples, respectively. It is equally appealing to refer to the difference between samples as an estimate of gain in knowledge or ability (e.g., U.S. Department of Education, 2006). Referring to the measurement time points by these terms, though, implies the data were measured longitudinally. Thus, this terminology is inappropriate. Most importantly, the students in one sample differ from the students in the other sample. This is the most important limitation of the separate samples pretest/posttest design (and cross-sectional designs in general) because it does not answer the question of how much students are gaining from their education.

Quasi-experimental: One-group posttest-only design. The one-group posttest-only design is the simplest quasi-experimental design. One group or sample, non-randomly formed, is measured after experiencing a treatment. Accordingly, the group completes a posttest but does not complete a pretest. In higher education, this design can be used to answer the question, “After experiencing the curriculum, are students meeting a standard of academic proficiency?”

The design can be conceptualized as follows:

T $X_{1_{\text{post}}}$

Again, “T” refers to the treatment and “ $X_{1_{\text{post}}}$ ” refers to the measurement of the sample after receiving the posttest.

Pros. This design is useful if the researcher already has a pre-formed group of interest. The one-group posttest-only design is convenient when a measure is only available after the group has received the treatment. This design is relatively cheap to implement, and the researcher – if somehow not concerned about making causal inferences - does not have to worry about testing effects, reactive effects of experimental arrangements, instrumentation effects, or statistical regression effects.

Cons. The one-group posttest-only design, however, is extremely limited with respect to internal validity. This design is subject to history effects, maturation effects, selection bias, and attrition. Subsequently, one cannot infer much from the posttest scores about the effect of the treatment. Results are likely sample-dependent and, as a consequence, inferences about the results unlikely to generalize to the student body. This design also suffers from effects due to the interaction of selection bias and treatment. For instance, a researcher may measure students in a particular math course to understand math learning gains at the university. If these students opted to take this course due to interest in the material, they may be more likely to learn from the course. Researchers who use the one-group posttest-only design may therefore make less externally valid inferences about the desired outcome. Of primary concern, however, is whether this design can be used to assess student learning gains. Perhaps expectedly, it cannot. The one-group posttest-only design only provides an estimate of student performance because students are only measured once.

Quasi-experimental:: One-group pretest/posttest design. The one-group pretest/posttest design is considered a quasi-experimental design because students are not randomly assigned to the treatment group. This design is another longitudinal design. Similar to the non-equivalent comparison group design, a pre-formed group is measured before and after experiencing a treatment. Only one group, though, is measured.

The design can be conceptualized as follows:

$$X_{\text{pre}} \quad T \quad X_{\text{post}}$$

“ X_{pre} ” is the measurement of the sample before receiving treatment, or curriculum, “T” and “ X_{post} ” is the measurement of the sample after receiving treatment “T”. This design addresses the question, “How much do students gain from time one to time two?”

Pros. Because the one-group pretest/posttest design is a longitudinal design, it has several of the same benefits as the nonequivalent comparison group design. Similar to that design, the one-group pretest/posttest design produces student learning gain estimates. The relationship between learning gains and curriculum can be calculated to further investigate the curriculum’s effect.

Cons. This design has the same limitations as the nonequivalent comparison group design. In particular, inferences about student learning are affected by lack of random assignment, and other validity threats (see Table 2). Lack of a control group also means that history effects may influence how students receive the treatment. As well, this design can suffer from attrition effects. As explained above in the section on the nonequivalent comparison group design, attrition may upwardly bias learning gain estimates.

Additional limitations exist due to the measurement of only one sample. Unlike the nonequivalent comparison group design, this longitudinal design does not allow for comparisons of average gains between groups. The control group is needed to estimate the treatment effect. That is, it is impossible to parse out what effects can be attributed to curriculum and what effects occur naturally with time (i.e., maturation; Campbell & Stanley, 1963; Shadish et al., 2002). Though inferences can be made about student learning, they may be less valid. Lastly, the one-group pretest/posttest design can be expensive to use, even with one sample, due to the extra effort involved when tracking students over time.

Quasi-experimental: Posttest only design with nonequivalent groups. The posttest only design with nonequivalent groups is used to assess two groups at one time point. Therefore, it is similar to the separate samples pretest/posttest design. This design attempts to address the question: “Are the outcomes of students different depending on the curriculum each student experiences?”

This design can be conceptualized as follows:

	X1post
T	X2 post

“X1post” and “X2post” refers to the measurement of Group 1 and Group 2, respectively, after experiencing or not experiencing the treatment “T”. This design is often used to compare upper-class students who have yet to experience and who have experienced the treatment or curriculum (e.g., algebra test scores from upper-class students who have completed math coursework and from upper-class students who have not completed the math coursework; see Table 2).

Pros. The posttest only design with nonequivalent groups is easy to employ in higher education and is comparatively cheaper than longitudinal designs. It also provides relatively immediate results and can be used to effectively assess student proficiency in subject matter.

Cons. This design is subject to multiple validity threats. Similar to the separate samples pretest/posttest design, the posttest only design with nonequivalent groups is subject to selection, attrition, and threat interactions, as well as an interaction effect of selection bias and treatment. Thus, the researcher who uses this design cannot be sure he has removed all confounding influences and also cannot generalize his findings back to the population. If used for higher education assessment, it is hard to make valid inferences about student performance and impossible to make valid inferences about student learning gain.

How to determine the correct design for estimating learning gain. Assessment practitioners and institutions must obtain a valid estimate of student learning gain to demonstrate that their curricula facilitate student learning or, if not, to improve student learning gains. To obtain this estimate, students must be sampled and measured using an appropriate design.

The posttest-only designs described introduce construct-irrelevant variance (e.g., differences in personalities, demographics, motivation), which contaminates the performance estimates or inferences made about the estimates. When two groups are measured at posttest, the differences between the groups' performance estimates may stem from systematic differences in personal characteristics or curriculum characteristics (i.e., the treatment). In other words, the curriculum effects are confounded with the

differences in personal characteristics, thereby biasing the estimated effect of curriculum on performance. Furthermore, a pretest or posttest only provides an estimate of student ability at a particular point in time. Therefore, the separate samples pretest/posttest, posttest only design with nonequivalent groups, and one-group posttest-only design designs are inadequate for measuring learning gains.

As has been emphasized, learning gains must be measured longitudinally to make valid inferences about learning. A longitudinal design, though, is necessary but not sufficient to make these inferences. In order to make inferences about the quality or effectiveness of the curriculum, the learning gains of students who complete specific courses must be compared to the learning gains of students who have not completed these courses. Without this comparison group, it is difficult to gauge the magnitude of learning gain. Given that the one-group pretest/posttest design cannot be used to compare curriculum effects, it loses some efficacy for measuring student learning gains.

The pretest/posttest control group design can produce good estimates of learning gain because random assignment are employed. As elaborated, that the researcher can assume that confounding differences in background characteristics between groups are eliminated when random is employed. Random assignment, however, is difficult (if not impossible) to achieve in higher education contexts; it is not realistic to randomly assign students to courses.

In comparison, the nonequivalent comparison group design is better suited for applied settings. Curriculum effectiveness can be determined by comparing the learning gains of groups who have and have not experienced the curriculum. Furthermore, this design addresses the questions, “Are students learning” and “How much are students who

experience particular curriculum learning compared to students who are not experiencing the curriculum”. Researchers should use this design to measure student learning gain, especially if improvements are to be made to the curriculum.

Learning Gain Estimates

When investigating student learning, interest lies in the estimated learning gain. The estimated learning gain is “how much a student has learned on an absolute scale” (Castellano & Ho, 2013, p.35). That is to say, how much a student has learned is compared only to his past performance and not compared to a peer’s performance. The estimated learning gain is also described as the difference between posttest and pretest scores (Castellano & Ho, 2013; Liu, 2011b). In the sections below, I describe several methods used to compute this estimate: the residualized estimate, the raw mean difference, and Cohen’s *d*. I also discuss concerns regarding the raw mean difference and Cohen’s *d*.

Residualized estimate. The calculation of the residualized estimated learning gain is another approach to estimating student learning gain. This estimate can be used when performance is measured with different instruments. The residualized estimate is the difference between the observed score and the expected score that is predicted from prior performance (Castellano & Ho, 2013; Rogosa, 1995). It is computed by first predicting an individual’s posttest score from a pretest score via linear regression (Castellano & Ho, 2013). This predicted score is then subtracted from the observed posttest score. To illustrate, a researcher interested in quantitative ability collects data from a sample of incoming students at a college. He predicts the students’ senior quantitative GRE scores based on the students’ quantitative SAT scores. Once the

students complete the GRE, the researcher subtracts the actual GRE scores from the predicted GRE scores.

The difference between the estimates can be interpreted as how well the students actually performed on the GRE versus how well the researcher thought they would, given the SAT scores. Therefore, the residualized estimate is technically not an estimate of learning gain. This estimate is better suited to answer the question “How much did ability differ from what was expected?” and not “How much did ability change?” Nonetheless, it is often calculated by researchers attempting to measure learning gain (e.g., Herzog, 2011).

Raw mean difference. The first method produces the raw mean difference, or gain score, between posttest and pretest scores. The gain score is easy to calculate (posttest group mean minus pretest group mean) and it is comprehensible (e.g., the student gained X number of points on the measure from her first year to her last year of college). However, this mean difference loses interpretability if the pretest and posttest measures are on different scales. A common example is when researchers use SAT scores to measure ability when students are freshmen and GRE scores when students are seniors.

Concerns regarding reliability of raw mean difference. A misconception is that these gain scores are unreliable and therefore should not be used to estimate learning gain. This is an unfortunate misjudgment that begs clarification.

To explicate, the reliability of the raw mean difference is the ability of the measure to detect distinct rates of change. It is a function of the pretest and posttest reliabilities, correlations, and standard deviations (Bandalos, 2016; Rogosa, 1995; Williams & Zimmerman, 1996). There are several reasons for assuming that difference

scores do not yield reliable learning gain estimates. One such reason is that there is a low pretest/posttest correlation. Another, more prominent reason is that, although there is a high pretest/posttest correlation, the reliability of the gain scores is low.

That is, the reliability for the gain scores will be low when, holding the reliabilities of the pretest and posttest constant, the pretest/posttest correlation is high and there is little variability in pretest and posttest scores (Bandalos, 2016; Rogosa, 1995; Williams & Zimmerman, 1996). If there is little variability in the pretest and posttest scores (i.e., pretest scores are similar and posttest scores are similar), the gain scores will be similar. Because the students change similarly, there will be little variability in the gain scores. Therefore, the change or gain rates will be nearly equivalent across all participants. One cannot detect differences in individual gain in this scenario because, for all practical purposes, there are no differences to detect (Bandalos, 2016; Rogosa, 1995). It follows that these learning gain estimates are reliable when there are actual variations in learning gains to be detected (i.e., not all students have the same gain scores). Additionally, holding the pretest/posttest correlation constant, the reliability of the difference scores will increase as the reliabilities of both pretest and posttest measures increase (Williams & Zimmerman, 1996).

Although one might expect to see a strong relationship between pretest and posttest scores, a high correlation between the scores is not always desirable when measuring learning gains. For example, suppose a university assesses all students' academic abilities with a pretest and a posttest. A group of students on academic probation participates in an academic intervention after receiving pretest results. After completing the intervention, these students score higher on the posttest than

nonparticipants. On one hand, this result speaks to the success of the intervention; students once lower in ability are now higher than their peers. On the other hand, the pretest and posttest scores of the entire sample will be less correlated and students' rank-order will differ.

Fortunately, most institutions do not care about rank-ordering students by gain. Though there are situations where it is necessary to identify students who gain more or less, preoccupation with gain score reliability diverts attention from the biggest concern – whether or not students are learning. Thus, the researcher who is not interested in rank ordering individuals needs not be concerned with low gain score reliability.

Cohen's d . A third method of estimating learning gains is the computation of Cohen's d (e.g., Hathcoat et al., 2015; Pastor et al., 2007; Roohr et al., 2016). As a standardized effect size, Cohen's d can be used for institutional comparisons (i.e., comparing learning gain estimates of institutions that employ measures with different metrics). This standardized effect size can be calculated by dividing the raw mean difference by the sample standard deviation of the difference scores (Cohen, 1992):

$$d = \frac{\bar{x}_{\text{post}} - \bar{x}_{\text{pre}}}{s_d}$$

In the above equation, \bar{x}_{post} refers to the posttest group average, \bar{x}_{pre} refers to the pretest group average, and s_d refers to the sample standard deviation of the difference scores. The resulting statistic d is an average learning gain estimate on the standardized gain metric and is interpreted in terms of standard deviations of the gain or difference scores. For example, $d = 0.3$ would be interpreted as a gain of 0.3 standard deviations on the standardized gain metric.

Although this metric is always standardized, the type of metric (e.g., gain) can change based on the standard deviation used in the denominator. As an aside, Cohen's d can be used to compare performance estimates computed from a cross-sectional design. When performance estimates are compared, the pooled standard deviation of the groups is typically used as the denominator (Dunst & Hamby, 2012). When computing an estimate of learning gain, a variety of standard deviations can be used. Alternative standard deviations, such as the standard deviations of the pretest (e.g., Pastor et al., 2007) or posttest scores (Morris & DeShon, 2002), can be substituted in the denominator of the equation above. Using different standard deviations places the estimated learning gain on different standardized metrics and affects interpretation. For instance, if the standard deviation of the posttest scores is used, Cohen's d would then be interpreted as the standardized learning gain estimate on the standardized posttest metric.

Concerns regarding choice of denominator for Cohen's d . The standard deviation of the gain scores, as illustrated above, can also be used. Using the standard deviation of the gain scores as the denominator, though, is said to produce an *overestimate* of the effect (Lakens, 2013). This concern is most prominent in meta-analytic studies, where results from both between-groups (e.g., cross-sectional) and within-subjects (e.g., longitudinal) studies are combined (Morris & DeShon, 2002). Researchers who aim to generalize their effect sizes want the cross-sectional Cohen's d estimates to be of similar magnitude to the longitudinal Cohen's d estimates. Generally speaking, the type of design used (cross-sectional versus longitudinal) should not greatly affect the magnitude of the effect size. Subsequently, the effect size should be largely independent from the design used.

However, effect sizes computed with the standard deviation of the difference scores are not independent from the research design. This standard deviation tends to be small because it accounts for the correlation between the measurements present in longitudinal designs. Because the standard deviation of the difference scores is smaller, it produces a larger effect size than if other denominators were used. Denominators have been developed that do account for the correlation in longitudinal designs (see Cohen, 1988) or ignore it entirely (average of the measurement standard deviations; Lakens, 2013). The benefit of the latter is that it produces a similar effect size to that produced from a cross-sectional design, which enables the researcher who uses it to generalize his effect. On the other hand, some phenomena cannot be measured using cross-sectional designs, which makes the need for equivalent design effect sizes moot (Lakens, 2013). The obvious example here is student learning gain, which should only be measured longitudinally. In this scenario, the standard deviation of the difference scores will not produce an overestimate of the true effect and is an appropriate denominator.

Beyond the computation of Cohen's d , other misconceptions about standardized and unstandardized effect sizes abound. In 1989, Cohen reluctantly recommended benchmarks of $d = 0.2$ (small effect), 0.5 (medium effect), and 0.8 (large effect). These benchmarks, still used today, were defined arbitrarily. The classifications were made to distinguish effects that were easily visible (medium effect) and correspondingly smaller or larger (Cohen, 1992). Thus, the numerical estimates of 0.2 , 0.5 , and 0.8 were *not* intended to be permanent benchmarks. Newer guidelines suggest interpreting one's computed effects relative to effect sizes already reported in the literature, as "large"

effects may not be substantial and “small” effects may have great importance (Thompson, 2007).

Personal and Curriculum Characteristics Related to Learning Gains

Multiple personal and curriculum characteristics affect how much students learn. Additionally, many personal and test characteristics affect *estimates* of academic learning gain. It should be noted that most published studies supposedly examining these personal and curriculum characteristics are not evaluating how these factors affect learning gain. Surprisingly little literature discusses factors that affect *changes* in ability or performance. This study will empirically investigate if and how these factors affect college students’ quantitative and scientific learning gains. Understanding how these factors affect student performance may help to better understand how these factors potentially impact student learning gain. In the sections below, I review the factors related to student performance, as well as some research on how these factors may relate to learning gains.

Gender. Research has found that gender both predicts and moderates student performance. Bray and colleagues (2004) investigated how reading comprehension and attitudes toward literacy develop from the first year to the third year of college. Regressing gender, among other predictors, on third year scores, the researchers found a conditional effect of gender: male students who took professional or technical courses had significantly lower scores in reading than female or other male students. As well, female students had significantly higher attitudes toward literacy than males. However, females did not have significantly higher scores in reading comprehension. Differences in math performance have also been documented. Males and females may differ on tests if

the items assess male-dominant cognitive skills (e.g., math word problems) or if economic and social differences are not included in analyses (Buchmann, DiPrete & McDaniel, 2008).

Pascarella and Blaich (2013) discovered a conditional effect of gender and high-impact learning practices on critical thinking learning gains. Specifically, males increased their learning gains significantly more when they interacted with faculty whereas females did not benefit more from interacting with faculty. Toutkoushian and Smart (2001) assessed the effect of gender on student learning gains. The researchers used self-reported gains to gauge student learning gain in six outcomes: learning/knowledge, tolerance/awareness, grad school preparation, communication skills, and miscellaneous achievements. Results suggested that female students have significantly greater gains in communication skills than males after controlling for ethnicity, prior academic ability, and other various personal characteristics. On the other hand, males and females did not appear to differ in their self-reported learning/ knowledge gains after controlling for personal characteristics. In contrast, some work has shown that females have smaller learning gains in math and science than males (Finney et al., 2016; Hagedorn, Siadat, Nora, & Pascarella, 1996). In sum, males and females may develop their math skills at different rates. Research investigating the effect of gender on math gains is remarkably slim; much of the research investigates the effect of gender on performance rather than learning gains (e.g., Bray et al., 2004). The current study will address this issue by investigating the predictive power of gender on quantitative and scientific learning gain.

Prior academic ability. Although it is desirable that all students leave college equally skilled, this outcome is not typical. In general, students with greater intellectual

abilities consistently outperform their less-adept peers (Seifert et al., 2007; Wholuba, 2014). With respect to college-level learning gains, prior academic ability does not appear to affect *self-reported* learning gains (Toutkoushian & Smart, 2001). In contrast, empirical research has found that more academically adept students demonstrate smaller learning gains than their peers in high school (Grigorenko, Jarvin, Diffley, Goodyear, Shanahan, & Sternberg, 2009) and in college (Pastor et al., 2007). Because these students are already performing highly, this result might stem from a ceiling effect. That is, these small gains may occur because these students have less to master during college or because the measures employed are not sensitive to learning gain. Both linear (Grigorenko et al., 2009) and nonlinear (Ryoo et al., 2014) models have been fit to rates of learning gain. To address this issue, the current study will examine the effect of prior academic ability on math and science learning gains. Linear and nonlinear predictors of academic ability will be included in the model.

Coursework. Course content affects student learning gains both in that domain and beyond (Pascarella & Terezini, 2005). Generally, a diverse curriculum appears to encourage development of diverse skills. A wide-spread investigation of college general education curricula found that students who had under 40% of their total coursework from general education courses and an unequal distribution of content matter (e.g., more math general education courses than literature general education courses) had greater gains on the ACT COMP objective test (Knight, 1993).

After controlling for prior academic ability, exposure to math and science courses is associated with higher scores in reading (Bray et al., 2004). Additionally, exposure to math and science courses is associated with higher critical thinking after controlling for

prior critical thinking ability (Terezini et al., 1995). If higher education truly causes learning gain, one would expect that a student's learning gain would increase as the student completes more coursework. Existing research supports this claim (e.g., Hathcoat et al., 2015; Pastor et al., 2007).

Major. Students' chosen field of study is also linked to student learning gain. Students in particular fields demonstrate increased learning gain in content matter relevant to their declared majors; this effect is particularly pronounced for students in STEM majors (Pascarella & Terezini, 2005). However, these findings were not replicated in studies focused on verbal skills (Pascarella & Terezini, 2005).

Additionally, student major tends to moderate learning gain in general skills (Pike, 1992) and domain-specific skills (Herzog, 2011). To be clear, general skills concern overall performance whereas domain-specific skills concern performance in a particular field of study. On measures of academic aptitude, business students have been shown to have the greatest gain in both general skills (Pike, 1992) and domain-specific skills (Herzog, 2011) than other majors. After business students, students majoring in physical sciences (e.g., physics, math) exhibit greater gain in domain specific skills than other majors (Herzog, 2011). It should come as no surprise that students who take courses related to their major tend to exhibit greater gains in that field. Students who are interested in the material tend to learn more (Wigfield & Eccles, 2002). One would expect that these students are interested in and willing to learn the material from these relevant courses.

Test-Taking Motivation and Learning Gains

Practitioners are interested in how personal characteristics can affect student performance and learning gain. However, personal characteristics can also affect how accurately researchers estimate learning gain. One such characteristic of particular concern is motivation. According to Expectancy-Value theory (E-V theory; Wigfield and Eccles, 2000; 2002), motivation (or expended effort) is a function of two domain-specific components: expectancy and value. E-V theory can be applied to test-taking behavior (Sundre & Moore, 2002; Wolf & Smith, 1995). E-V theory is particularly useful for explaining test-taking behavior on low-stakes tests, which will be the focus of the remainder of this literature review. In low-stakes testing contexts, performance on the test is not associated with consequences for students. A student who does poorly on the test will not receive reprimands, and a student who does well will not receive rewards. However, scores from these low-stakes tests are often used by administration in high-stakes situations (e.g., curriculum modifications and higher education accreditation). Because there are no consequences, students tend to put forth little effort on these tests.

This amotivation can be described in terms of expectancy and value. Expectancy concerns students' perceptions of their capabilities to complete the test; value concerns the significance of the test to the students. The value component can be further divided into four subcomponents: interest, usefulness, importance, and cost (Wigfield & Eccles, 2002). Expectancy is often dropped or disregarded in applications of E-V theory as it is not as closely associated with expended effort as test value (Eklof, 2010; Wigfield & Eccles, 2000). Expectancy is also much more difficult to manipulate than value, as students may not be able to accurately judge their capabilities on tests. Some research has

shown a weak relationship between expectancies and effort (Barry & Finney, 2016; Eklof, 2006). However, newer work suggests that there may be a stronger relationship between expectancy and effort than previously found (Penk & Richter, 2016).

The four value sub-components also take on their own meanings in low-stakes contexts. Interest is how much enjoyment examinees get out of taking the test; usefulness is how worthwhile the test is to achieving future goals; importance is how important examinees believe the test to be; and cost is what examinees had to give up in order to take the test (Eklof, 2010). Value tends to be positively associated with test-taking effort. Specifically, importance (Cole, Bergin, & Whittaker, 2008; Knekta & Eklof, 2014; Thelk et al., 2009) and usefulness (Penk, Pohlmann & Roppelt, 2014) have been shown to be positively correlated with effort. Students who place higher importance on the test or believe test scores can help them achieve their goals tend to try harder on the test. Most research investigating motivation focuses on the relationship between importance and effort. Work has been done to ensure these factors are distinct (Finney, Mathers & Myers, 2016; Thelk et al., 2009). Researchers have also developed measures of motivation that assess both perceived test importance and test-taking effort (e.g., Student Opinion Scale; Sundre & Moore, 2002).

What test-taking motivation affects. Test-taking motivation affects test-taking behavior. This behavior, in turn, affects test performance and, potentially, learning gain estimates. That is, learning gain estimates may be attenuated by low test-taking motivation. Thus, test-taking motivation can impact the validity of inferences about student performance and may impact the validity of inferences about learning gain. In the following sections, I discuss in detail how test-taking motivation affects estimates of

performance and learning gain, as well as the impact of perceived test importance on performance and learning gain.

Test-taking motivation is positively related to test performance (Knekta & Eklof, 2014). Students who put forth more effort on tests perform better than students who put forth less effort (Eklof, 2007; Penk et al., 2014; Sundre & Kitsantas, 2004; Wise & DeMars, 2005; Wise & DeMars, 2010; Wise & Smith, 2011; Wise, Wise & Bhola, 2006). In fact, motivated students can perform up to half a standard deviation better than unmotivated students (Wise & DeMars, 2005).

Researchers have empirically demonstrated that importance has an indirect effect on performance through effort (Cole et al., 2008; Mathers, Finney, & Myers, 2016; Myers, Finney, & Mathers, 2016; Zilberberg et al., 2014). That is, how highly a student values a test relates to how much effort the student puts forth on the test. Test-taking effort, in turn, relates to how well the student performs on the test. Thus, it would be expected that a student who believes a test to be important would put forth good effort and perform well, and a student who does not value a test would not try to do well and therefore perform poorly.

Given these relationships, it can be difficult to make valid inferences about students' abilities from test scores. As previously outlined, students demonstrate higher levels of test-taking motivation when test has meaning to students. That is, students could perform better on these tests if they were more motivated. It can reasonably be assumed that performance estimates of unmotivated students may be *underestimates* of these students' abilities. If so, low test-taking motivation has become construct-irrelevant variance.

Construct-irrelevant variance. Construct-irrelevant variance (CIV) is a predictable, quantifiable (i.e., systematic) error that clouds estimates of the construct of interest (Haladya & Downing, 2004). Consider the following scenario. A student with poor English skills is given a math test. However, the test consists mostly of word problems, and the student has a difficult time understanding what the problems require her to do. Although her score on this test is meant to be an indication of her math skills, it is more indicative of her reading comprehension. In this scenario, reading comprehension is CIV and undermines her estimated math ability. Low test-taking motivation functions the same way. The test is not meant to measure low test-taking motivation, yet low test-taking motivation still undermines test scores.

One can ascertain how much test-taking motivation may affect performance estimates by examining the relationship between test-taking motivation and performance. Hathcoat et al. (2015) found test-taking motivation to be moderately correlated with performance ($r = 0.47$). Myers et al. (2016) found that the indirect effect of perceived test importance on test performance through test-taking effort accounted for up to 30% of the variance in test scores. Wise and DeMars (2005) found that students' mean test performance increased by almost four points as they raised their desired level of effort on the SOS.

Test-taking motivation and learning gains. It is equally important to ensure low test-taking motivation does not affect estimates of learning gain. Low test-taking motivation can also account for the difference between seeing no gain in performance versus seeing a moderate gain in performance (Wise & DeMars, 2010). That is, low test-taking motivation may also attenuate learning gain estimates. This outcome can occur if a

student is unmotivated on a pretest, on the corresponding posttest, or on both of these measures. If a student is unmotivated on the pretest but motivated on the posttest, only her pretest score will be attenuated. Thus, the difference between her pretest and posttest scores will be artificially larger; it will appear that she has learned more than she has. If she is unmotivated at the posttest or on both measures, the difference between her two scores will be artificially smaller; it will appear as if she has learned less than she actually has. It is therefore critical that researchers investigate how test-taking motivation affects learning gain estimates in applied settings. Some work has been done in this area. Research has found that motivation is positively associated with change in performance (Gottfried et al., 2007; Taasoobshirazi & Sinatra, 2011). Furthermore, change in motivation has been found to relate to change in math performance (Gottfried et al., 2007). Finney et al. (2016) found that change in importance and change in effort were positively correlated with value-added estimates of quantitative and scientific reasoning. Corresponding research conducted by Williams (2016) corroborated the effect of changing importance on learning gains. She also found a stronger effect between change in effort than change in importance on learning gain.

Fortunately, researchers have developed a method to reduce the attenuating effects of low motivation on learning gains. When data from unmotivated students are removed from analyses, results computed from the remaining data are more indicative of student learning gain. This technique, motivation filtering, is described in the section below.

How to Address Low Test-Taking Motivation: Motivation Filtering

To produce trustworthy estimates of learning gains, it is critical to eliminate the attenuating effects of low motivation. Researchers have proposed statistical adjustment of test scores, where motivation would be included as a predictor in a regression analysis (Wise & DeMars, 2005). This technique, however, has not been put into practice as researchers are concerned about the implications of such artificial inflation of test scores. Motivation filtering, on the other hand, has garnered both positive attention and legitimacy in the struggle against low test-taking motivation.

Motivation filtering is a method of removing CIV in order to obtain better estimates of students' abilities (Wise & DeMars, 2005). It leads to more precise estimates of ability (i.e., decreased SDs; Wise et al., 2006). Motivation filtering also leads to increases in average test scores when scores have been attenuated by low motivation (Wise et al., 2006). There are several ways to conduct motivation filtering. In computer-based testing (CBT), response-time effort (RTE) is often used to identify unmotivated students (Wise & DeMars, 2010). RTE refers to the amount of time a student takes to answer an item. It is assumed that the amount of time spent corresponds to the student's effort. A lower time indicates that a student is not putting forth effort (i.e., exhibiting rapid guessing behavior). Typically, a threshold is set for examinee's rapid-guessing behavior. The assumption is that if students were providing valid responses, they would require more time to read the item and respond thoughtfully (Swerdzewski et al., 2011). To determine the boundary between rapid-guessing and effortful responding, a time threshold is set for each item (Wise & Kong, 2005). The threshold reflects the minimum amount of time a student will spend answering an item if he is motivated. Students who

fall below the threshold are removed from the analysis because it is assumed that they are not motivated to perform well (Swerdezowski et al., 2011; Wise & DeMars, 2010). For example, a researcher could set a threshold for an item at 4 seconds. Students who spend at least 4 seconds on the item are assumed to exhibit good effort. After the data have been collected, the researcher would filter out examinees who took less than 4 seconds to respond to the item.

Self-report measures can also be used for motivation filtering in either CBT or paper and pencil modalities. Motivation filtering via self-report measures is conceptually similar to motivation filtering via RTE. That is, both methods involve calculating a threshold of motivation and filtering out unmotivated students from the sample who do not meet that threshold. With self-report measures, a cutoff score (i.e., threshold) is used to identify unmotivated students. Students whose reported motivation falls below this score are removed from the sample. Some deprecate self-report measures for their sensitivity to response bias and inability to account for changes in effort during the test (Wise & Ma, 2012). On the contrary, self-report measures have been shown to have utility when conducting motivation filtering (Rios et al, 2014; Swerdezowski et al., 2011; Wise & Kong, 2005). The Student Opinion Scale (SOS), developed under E-V theory, is one such measure (Sundre & Moore, 2002). This scale demonstrates good psychometric properties (Thelk et al., 2009). As well, it can be used to identify unmotivated students (Sundre & Wise, 2003; Swerdezowski et al., 2011).

The SOS can be either test-specific (administered following the test) or test session-specific (administered following a battery of tests). Both measures have been used for motivation filtering (Hathcoat et al., 2015). Motivation filtering has been

conducted using the test session-specific total SOS score (Sundre & Wise, 2003). However, it is not recommended that examinees are filtered using their total motivation score because the total score confounds information about perceived test importance with expended effort. That is, an examinee who believes the test to be very important but who expends little effort (and therefore has little motivation) may achieve the same SOS total score as an examinee who does not believe the test to be important but who puts forth effort (and therefore is highly motivated). Thus, highly motivated examinees may inadvertently be filtered from the sample. Instead, examinees should be filtered based on effort scores. Only students who expend little effort will then be filtered from the sample.

However, the measures may classify different students as motivated or unmotivated. That is, filtering using the test-specific measure may produce different results than filtering using the test-session specific measure. Specifically, the two measures have been found to identify 78.7% of the same motivated students (Hathcoat et al., 2015). In the aforementioned study, however, 8.9% of students reported adequate effort on the test-specific measure but were unmotivated by the end of the battery. These students were therefore not retained when the test session-specific measure was used to filter data. Furthermore, the researchers found evidence to suggest that test-specific and test session-specific effort scores are not redundant (i.e., do not measure the same type of motivation). However, filtering using the two measures produces similar performance estimates (Hathcoat et al., 2015). Unfortunately, few studies compare motivation results from these two self-report measures. At the current author's institution, data from both the test-specific and test session-specific measures of motivation are collected. The

current study will assess how filtering using test-specific and test-session specific measures affects learning gain estimates.

Determining an Adequate Amount of Learning Gain

Learning gain estimates can shed light on how college coursework affects student learning, but only if context is provided for the estimated learning gains. Estimated learning gains that are reported without reference to a predetermined standard have little utility. Put simply, estimated learning gains that are reported without reference to a standard do not inform stakeholders of whether students are adequately learning.

Who should determine what is an adequate amount of learning gain, and how should they determine this standard? An adequate amount of learning gain should be determined by those who develop and administer the curriculum: faculty. Faculty involvement in student learning assessment is necessary to improve student learning (Banta & Blaich, 2009). In fact, their roles in student learning assessment extend far beyond the classroom. Faculty should be involved in selecting or developing measures to assess student learning gain (e.g., Ewell, 2009; Schmeiser & Welch, 2006) and determining desired scores (or level of ability) on those measures (Castellano & Ho, 2013). Faculty should also be able to use results to determine the amount of learning gain they would like or expect to observe as a result of their pedagogy.

Unfortunately, little has been done by higher education administrators or faculty to determine how much learning gain should be expected if students *are* learning from the curriculum. At the same time, there has been a push to make learning gains comparable across institutions (Roohr et al., 2016; U.S. Department of Education, 2006), which is valuable information for the higher education community. After all, in the current

student-as-consumer higher education climate, this information can affect where students enroll. Feasibly, students, parents, and other stakeholders are also eager to know how much students can expect to learn after attending a given institution. Yet, institutions themselves lack a standard of *absolute* learning gain. For an institution to be able to demonstrate its effectiveness, it is important that the institution provides evidence that students are learning and meeting learning gain expectations. Evidence of effectiveness can be provided in the form of a standard of learning gain.

Standard setting. Current practice for *performance* standard setting involves faculty setting a cut score for criterion-referenced tests to determine student proficiency at one time point (e.g., DeMars et al., 2002; Hathcoat et al., 2015). However, setting a performance cut score has limited utility for determining adequate learning gain. Knowing whether or not a student is minimally proficient does not assist in knowing how much that student changed over time. A student may grow substantially yet still fall below the performance cut score. Therefore, it may be more appropriate to set a learning gain standard rather than a performance standard in order to gauge curriculum impact. A learning gain standard can be set by referring to current learning gain estimates (Gong, 2004). Additionally, procedures for setting learning gain standards have been described; I discuss these procedures below.

There are three types of procedures for learning gain standard setting: scale-based, target-based, and norm-referenced (Castellano & Ho, 2013). Scale-based setting classifies learning gain into different categories (e.g., “low” v. “high”) based on cut points. A group of faculty determines these cut points by examining the institution’s distribution of student learning gains and basing categories on typical learning gains. Target-based

setting also classifies learning gain into categories but takes into account whether or not a student is on target to achieve a set standard (e.g., one group of students is “on track” to meet a college readiness standard by the time they are in 11th grade, whereas another group is not). Norm-referenced setting involves comparing the distribution of student learning gain estimates to the distribution of a control group. For example, suppose the learning gain estimates from the control group are normally distributed. Researchers can compare a score from the treatment group to this distribution to determine if the student’s gain is typical or atypical. This control group should come from the same or similar population. The scale-based approach is most appropriate for determining a standard of absolute learning gain. However, this method has a major limitation: if faculty are unaware of how much their students are learning, they cannot make any decisions about what would be an adequate (or inadequate) amount of gain.

What faculty expect with respect to learning gains. To the author’s knowledge, no research has been conducted on faculty expectations of learning gain. Extensive research has been conducted on teacher expectations in K-12 settings, which may provide some insight into how much college-level faculty expect of their students.

Though this body of literature may provide some insight, research in K-12 educational settings is mixed on whether teacher expectations align with student performance. Teacher expectations have been found to significantly overestimate reading performance of minority primary school students (Rubie-Davies et al., 2006). In a study on teacher perceptions of elementary school performance, however, teachers tended to have similar median expectations to students’ observed math performance; math performance was operationalized as students’ scores on the Woodcock Johnson

Applied Problems subtest (Hinnant et al., 2009). These researchers also found that, for students whose families were low income, teachers' expectations of math performance significantly and positively predicted their math performance in later grades. This result implies that how teachers expect students to perform may impact students' learning, and that teachers with high expectations may encourage greater learning in their students. A study on Dutch primary schools found that teacher expectations correlated highly with students' performance on high-stakes national test (Timmermans, de Boer, & van der Wer, 2016). However, it is important to keep in mind that the current study focuses on results from a low-stakes test.

Research also indicates that middle school teachers do believe their students can achieve relatively high performance-based standards (Harris, 2012). However, these teachers described challenges that might prevent their students from reaching their expectations, such as students' academic abilities, problems at home, and "lack of student responsibility for their own learning or motivation" (Harris, 2012, p. 138). A sample of high school teachers, when questioned about the decline in academic achievement of their African American students, also attributed the decline to family-influenced factors (e.g., "lack of parental support in the home"; Falconer-Medlin, 2014, p.88) and student-influenced factors (e.g., "lack of interest in school or low motivation", p.88). These high school teachers additionally attributed the decline to school-influenced factors (e.g., "curriculum is not engaging, relevant, or culturally-inclusive", p. 88).

At the college level, frameworks for student learning outcomes have been proposed. One such framework is the Degree Qualifications Program (DQP), a resource that describes what students should be able to know or do after obtaining an Associate

Bachelors', or Master's degree (Kuh et al., 2015; Lumina Foundation, 2011). For example, a student at the bachelor's level "translates verbal problems into mathematical algorithms and constructs valid mathematical arguments using the accepted symbolic system of mathematical reasoning." (Lumina Foundation, 2011). Though the DQP may be helpful in identifying what level of performance is expected, it does not illustrate what level of learning gain is expected.

Instead, faculty expectation research centers on why faculty believe students are or are not learning at college. In their work on student learning gains in higher education, Arum and Roksa (2009) gave the impression that faculty do not have faith in their students' motivation to learn. Leaning on research in sociology, the researchers warned that students' peer groups may affect their willingness to learn. Arum and Roksa furthered explained that "Many students come to college not only poorly prepared by prior schooling for highly demanding academic tasks that ideally lie in front of them, but - more troubling still - they enter college with attitudes, norms, values, and behaviors that are often at odds with academic commitment." (Arum & Roksa, 2009, p. 3). Chickering theorized that poor student learning stems from poor pedagogy (Chickering, 1999). In his seminal work, Chickering outlined the various academic and personal stages of development that college students move through to become intellectuals. He argued that lecture-based coursework and conventional examinations only moved students through 'simpler' stages of development, and did not support student learning. Although the author did not elaborate on whether or not college students learn at their schools, his stance seemed to imply that students are not learning as much as they could.

However, other work has shown that faculty do believe in their students' academic capabilities. Darby and Newman (2014) conducted a study on faculty who taught service-learning courses. The researchers asked these faculty their opinions on questions ranging from what they perceived were the benefits of service-learning coursework to what affected their motivation to teach such courses. Faculty elaborated that they were motivated by student-based outcomes, such as integration of knowledge and connection of course material to real-world experiences. These faculty believe that their pedagogy is effective, and that their students can both retain and apply the material learned in their courses.

Although a substantial body of research exists in the K-12 education domain, there is little literature regarding faculty expectation of how much students should be learning. Instead, the faculty expectation literature focuses on whether faculty believe students can learn and what affects student learning. Given this gap in the literature, the current study will investigate how much faculty expect students to learn from their coursework. Faculty will be asked to estimate how much they think students at the institution learn, as well as how much they would like students to learn.

CHAPTER THREE

Methods

This study employs a mixed methods design. That is, I employed quantitative analyses and then used results from the quantitative analyses to inform the qualitative analyses. Mixed methods research, however, constitutes more than use of quantitative and qualitative research methodologies, or *strands*. One of the primary features of mixed methods research is that the researcher articulates her paradigms, or her views on what knowledge is and how knowledge is gathered (Creswell & Plano Clark, 2011; Merriam & Tisdell, 2016)⁸. In this study, I adopt a post-positivist paradigm for the quantitative strand and a constructivist paradigm for the qualitative strand (Creswell & Plano Clark, 2011; Merriam & Tisdell, 2016). The post-positivist paradigm acknowledges that knowledge or reality is not always adequately captured, but still posits that there is one reality and that it can be measured. In adopting this paradigm, I assert student learning gains are real phenomena to be assessed and predicted. With respect to weighting, I prioritized the quantitative strand (QUAN⁹). In contrast, the constructivist paradigm asserts that knowledge and reality are socially constructed. In adopting this paradigm, I assert that the opinions of faculty at this institution, with respect to their expectations and desires of student learning gains, are constructions that stem from each faculty's teaching experience. I weighted the qualitative strand less than the quantitative strand (qual).

⁸ To date, there are four paradigms a researcher may adopt: post-positivist, constructivist, critical research, and postmodern (Merriam & Tisdell, 2016).

⁹ For researchers unfamiliar with mixed methods terminology, please consult Creswell and Plano Clark (2011) for an in-depth description.

To adequately assess learning gains and faculty expectations, this study employs a multiphase *embedded design*. In an embedded design, a secondary strand is added to address a research question that cannot be answered by the primary strand (Creswell & Plano Clark, 2011). My qualitative strand is embedded within my quantitative strand; the qualitative hypothesis is distinct from the quantitative hypotheses but cannot be addressed without results from quantitative analyses. The current study begins with quantitative analyses followed by qualitative analyses. Below, I provide information on data collection for each strand.

Participants and Procedures for Estimating Growth (Phase 1)

At the public, Mid-Atlantic university where this study was conducted, the effectiveness of the general education curriculum has been assessed for over twenty years during the biannual Assessment Day. Assessment Day is held once before the start of the fall semester and again several weeks into the spring semester. Incoming first-year students are tested during the fall. Upper-class students are tested during the spring once they have accumulated between 45-70 credit hours. These longitudinal data allow for the computation of gain scores, which can be used for both accountability purposes and, just as importantly, the improvement of the general education curriculum.

All incoming students are assessed during the mandatory assessment day in the fall. Given time constraints, however, each student does not complete all tests. Students are randomly assigned to a testing room based on the last few digits of their student ID number. Each testing room corresponds to a specific battery of tests. Test batteries are comprised of both cognitive and noncognitive measures. A majority of these measures were developed by faculty to align with general education learning outcomes at the

university. Each test battery takes approximately two hours to complete. Assigning students to test configurations by their student ID enables university assessment experts to assign students to the same battery at both testing sessions (at the start of their college career and again a year and a half later after accumulating between 45-70 credit hours).

If a student fails to attend Assessment Day, a hold is placed on the student's account and the student must attend a makeup session. With the exception of this repercussion, no other consequences exist for students. Performance on the tests does not affect graduation or course grades. For example, if a student performs poorly on a math and science test administered during Assessment Day, it does not affect her Calculus course grade. Thus, the tests administered on Assessment Day are low stakes for students; they have no personal consequences to the student.

Data used in this study were collected from cohorts 2007-2009, 2008-2010, 2013-2015, 2014-2016, and 2015-2017 during the regular Assessment Day (i.e., not from makeup testing; see Table 4). I analyzed data from these five cohorts to gauge the stability of the estimates of student learning gains in quantitative and scientific reasoning¹⁰. For students in each of the five cohorts, I gathered the number of math and science courses completed at the time of the second testing (number of courses completed ranged from zero to seven). By computing the gains based on number of courses completed, I was able to evaluate if collapsing across coursework masks the effects of the curriculum (i.e., if increased coursework affects the magnitude of the learning gain). Due to few students having completed either zero or at least five courses by their sophomore

¹⁰ All datasets are distinct from the data analyzed in published studies by Hathcoat and colleagues (2015) and Finney and colleagues (2016).

year, I collapsed across the cohorts to determine how much students gain after completing or not completing quantitative and scientific reasoning coursework.

Measures for Estimating Growth (Phase 1)

Natural World, Version 9. Quantitative and scientific reasoning was assessed using the Natural World 9 (NW9), a 66-item quantitative and scientific reasoning test developed by faculty and university assessment consultants (Sundre, Thek, & Wigtil, 2008). In use since 2007, this test intentionally aligns with the general education quantitative and scientific reasoning curriculum. The test yields one total quantitative and scientific reasoning score (Sundre et al., 2008). In past studies, total scores have been shown to have good reliability (e.g., $\alpha = .77$, Finney et al, 2016). Adequate reliability was also evidenced across the five cohorts at both testing occasions, as shown in Table 5.

I subtracted students' quantitative and scientific reasoning pretest scores from their posttest scores to estimate individual learning gain on the metric of the NW9 test. I then computed the unstandardized average learning gain for the total sample (collapsing across the cohorts and number of quantitative and scientific courses) and for each cohort (collapsing across number of quantitative and scientific courses). I consider a 3-point gain on the NW9 a moderate unstandardized learning gain. I based this unstandardized learning gain value on prior quantitative and scientific reasoning studies (e.g., Hathcoat et al., 2015) and reports (e.g., Curtis, 2016) from this institution, where 3-point gains on this particular test are associated with moderate standardized learning gain estimates.

I then standardized these average unstandardized gain scores (i.e., Cohen's d estimate) using the standard deviation of the gain scores and again using the standard deviation of the pretest scores. Using the standard deviation of the gain scores allowed

comparisons to Roohr and colleagues' (2016) findings, whereas using the standard deviation of the pretest scores allowed comparisons to Pastor and colleagues' (2007) findings. In line with Cohen's benchmarks and findings from Pastor et al. (2007), I consider a standardized gain of 0.50 on the standardized pretest metric a moderate standardized learning gain. In their discussion on student learning gain estimates, Roohr et al. (2016) considered their standardized math gain estimate of $d = 0.41$ on the standardized gain metric to be moderate. Thus, I also consider a standardized gain of 0.40 SDs on the standardized gain metric a moderate standardized learning gain.

Number of courses completed. Given that coursework is predicted to have the greatest impact on learning gains, the number of relevant courses completed was gathered from university records. University faculty designed a set of math and science general education courses intended to increase quantitative and scientific reasoning. This math and science curriculum covers the three topics of "Quantitative Reasoning", "Physical Principles", and "Natural Systems", and includes a lab component. Example courses are "Calculus I" (Quantitative Reasoning course), "Concepts of Chemistry" (Physical Principles course), and "Biological Anthropology" (Natural Systems course). Students must complete a course in each of the three topics in addition to a lab. At minimum, these courses must amount to 10 credit hours. Three courses usually are enough to satisfy the 10-credit hour requirement (i.e., one course = 3 credit hours, one course with lab component = 4 credit hours), but some students may complete four courses if they complete the lab separately. In the current study, I gathered data on the exact number of relevant courses students completed upon the second testing occasion. Given that number of courses completed was collected from university's records, all students had complete

data. The number of courses completed ranged from zero to seven, excluding lab-only courses.

Academic ability. Academic ability estimates, as reflected via SAT or ACT, were gathered from university records to estimate the effect of academic ability on learning gains. Students' pre-college academic achievement tends to affect college performance (Seifert et al., 2007; Wholuba, 2015) and may affect learning gains (Grigorenko et al., 2009; Ryoo et al, 2014). Thus, regressing estimated learning gains on these scores allows for estimates of the effect of coursework on learning gains while controlling for academic ability.

SAT subscale scores range from 200 to 800 (Dorans, 1999). Both SAT Math and SAT verbal scores were summed to create one total SAT score. If a student completed the ACT instead of the SAT, and the ACT composite score was unavailable, ACT Math and ACT Reading scores were summed to create one ACT score. Most students in the five cohorts had SAT data. For those students that did not have SAT data but completed the ACT ($n = 25$), ACT scores were converted to the SAT metric using concordance tables made available by ACT and College Board (ACT, 2009). Students who did not have SAT or ACT data were deleted from the regression analyses ($n = 282$, unfiltered condition; $n = 48$, filtered condition).

Gender. Gender data were gathered from university records to determine how gender affects learning gains and if gender moderates relationships between learning gains and other predictors (i.e., number of courses, prior ability). Research has suggested differential performance between males and females on science and math tests (Buchmann et al., 2008), as well as differences in self-reported learning gains

(Toutkoushian & Smart, 2001). Therefore, learning gain estimates were regressed on gender and the interactions among gender, prior academic ability, and number of courses. I dummy coded gender (male = 0, female = 1). Gender data were available for all students in all cohorts.

Student Opinion Scale. To assess the impact of low effort on learning gain estimates, I removed NW9 data from examinees who reported low expended test-taking effort. Test-taking effort was assessed via the Student Opinion Scale (SOS; Thelk et al., 2009). Based on expectancy-value theory (Wigfield & Eccles, 2002), the 10-item SOS was created to measure examinees' perceived test importance (i.e., task value) and expended effort (i.e., motivation).

Two versions of the SOS are available: a *test session-specific* measure and a *test-specific* measure. The test session-specific SOS is administered at the end of a battery of tests to assess student motivation across *all tests in the session*. The test-specific SOS is administered at the end of a test to assess student motivation on *that particular test*. Instructions for the two measures differ slightly to distinguish the context (session or test) and the items on the measures are essentially identical (see Appendix A). Research supports the two-factor structure of perceived test importance and expended effort for the test session-specific SOS (Thelk et al., 2009) as well as the test-specific SOS (Finney et al., 2016). The test session-specific SOS ($\alpha = .80$, importance subscale, $\alpha = .83$ effort subscale; Thelk et al., 2009) as well as the test-specific SOS has been shown to have adequate reliability ($\alpha = .76$, importance subscale, $\alpha = .82$ effort subscale; Mathers et al., 2016). In this study, reliability estimates ranged from $\alpha = 0.63$ to $\alpha = 0.87$ for test

session-specific effort and from 0.71 to 0.84 for test-specific effort (see Table 5). Test session-specific and test-specific importance data were not collected for this study.

SOS effort scores from both versions were used for motivation filtering. In this study, I filtered using test session-specific effort scores and test-specific effort scores (see Tables 6, 7, 8 and 9). Cohort One did not complete either effort subscale; therefore, Cohort One data were not used in analyses investigating the impact of low test-taking effort on learning gains. Some students in the 2008-2010 cohort only completed the test-session specific SOS; other students in this cohort only completed the test-specific SOS. For this cohort, I filtered students using their scores on whichever measure they completed. For each cohort, I computed three gain estimates: unfiltered gain, test-session filtered gain, and test-specific filtered gain.

Researchers who employ motivation filtering must select a cut score to distinguish between students who are “motivated” and “unmotivated”. The cut score on the SOS effort subscale should not be too high nor too low (Wise et al., 2006). A suggested test for overfiltering (i.e., removing so many students that the resulting sample does not resemble the population) is to compare the SAT scores of the filtered sample to the unfiltered sample (Wise et al., 2006). That is, students’ level of motivation should not be related to students’ prior academic ability (Rios et al., 2014; Wise et al, 2016). If the cut score value is too high and too many students are removed, I would inflate the estimated learning gains (i.e., overestimate learning occurring on campus) and produce an artificial relationship between prior academic ability and motivation. If too low, few

unmotivated students would be removed and learning gain estimates would be attenuated by low motivation¹¹.

Researchers have recommended cut scores of 15 (Swerdzewski et al., 2011; Wise et al., 2006), 14 (Hathcoat et al., 2015) and 13 (Rios et al., 2014) on the SOS effort subscale which ranges from a possible low score of 5 to a possible high score of 25. However, these cut scores were determined using different techniques. Wise et al. (2006) and Hathcoat et al. (2015) selected the cut score where the SAT scores did not change by more than three points from the original sample. In contrast, Swerdzewski et al. (2011) and Rios et al. (2014) used the average, or slightly below the average, score of the effort subscale. Similar to Swerdzewski et al. (2011), I initially used the average of the effort subscale, a cut score of 15, and removed NW9 data associated with students who have an effort score below this value. Specifically, I filtered out students who had SOS effort scores lower than 15 at either the pretest or posttest.

For each person removed, I recorded the reason for removal (low effort at pretest, low effort at posttest, low effort at both time points; see Table 10). After removing data from students with scores below 15, I examined average SAT scores to ensure I did not overfilter. If the SAT scores from the filtered sample were at least three points higher than the SAT scores from the students who were removed, I would need to lower the cutoff score to a number that does not artificially produce a relationship between motivation and academic ability. When motivation filtering was applied to data from

¹¹ Although recent research has suggested there may be a relationship between motivation filtering and prior academic ability (Rios, Guo, Mao, & Liu, 2016), the study in question used RTE in lieu of self-reported motivation on scores from a high-stakes test administered at one institution.

Cohort 4, using a cut score of 15 for both the test-session specific and test-specific effort scores appeared to produce qualitatively different samples. Average SAT scores were at least six points higher than in the unfiltered sample; these initial SAT averages are shown in Tables 7-9. I conducted the analysis again using lower cut scores until the SAT scores of the filtered samples from Cohort 4 were roughly within three points of the original sample. Based on results from this process, I used cut scores of 12 on the test-session specific effort subscale and 13 on the test-specific SOS for Cohort 4.

Prior to deleting cases with missing motivation data, NW9 data were available for 1554 students (see Table 6). Of these students, 0.31% identified as American Indian; 5.32% as Asian; 3.76% as Black; 3.13% as Hispanic; 0.38% as Pacific Islander; 82.17% as White; and 4.94% were unspecified. Furthermore, 67.87% identified as female and 32.13% identified as male. The average student age at pretest was 18.44 years, and the average at posttest was 19.91 years. Although there were slight demographics differences among the cohorts, these demographics align with the university demographics. SAT scores varied among the samples, ranging from 1117.39 (Cohort One) to 1146.81 (Cohort Four).

Recall that Cohort One did not complete either SOS measure. Collapsing across Cohorts Two-Five and prior to filtering, 828 students had complete data on the test-specific SOS and 564 students had complete data on the test session-specific SOS. After filtering for low test-specific motivation, NW9 data were available for 737 students (see Table 7). Thus, I filtered 91 out of 828 students (10.99%) due to low test-specific effort. Sample demographics changed slightly after filtering. Again, collapsing across the cohorts, 0.68% identified as American Indian; 6.38% as Asian; 5.02% as Black; 3.39% as

Hispanic; 1.09% as Pacific Islander; 84.40% as White; and 5.43% were unspecified. Of these students, 66.49% identified as female and 33.51% identified as male. The average age at pretest was 18.44 years, and the average at posttest was 19.90 years.

After filtering for low test session-specific motivation, NW9 data were available for 511 students (see Table 8). Thus, I filtered 53 out of 564 students (9.40%) due to low test session-specific effort. Again, sample demographics differed slightly from the unfiltered sample. Of these students, 1.12% identified as American Indian; 7.61% as Asian; 6.49% as Black; 5.37% as Hispanic; 0.89% as Pacific Islander; 86.35% as White; and 2.24% were unspecified. Of these students, 65.75% identified as female and 34.35% identified as male. The average age at pretest was 18.45 years, and the average at posttest was 19.91 years.

Furthermore, 489 students completed both the test-specific and test session-specific SOS. After filtering, NW9 data were available for 413 students. Twenty eight students indicated both low test-specific and low test session-specific effort (see Table 10). In total, I filtered 76 unmotivated students from this sample.

Participants for Faculty Reactions

Four quantitative and scientific reasoning general education faculty participated in this study¹². To recruit faculty, I sent an email to nine faculty on the quantitative and scientific reasoning assessment committee informing them of the nature of my study and asking for participation. This email contained the following text:

¹² Prior to recruiting participants, the protocol for the qualitative strand was sent to and approved by the Internal Review Board (IRB). This protocol included methods of recruitment, interview procedure, Forms A and B, intended data analyses and storage, and an interview guide.

“I am looking for 3 to 10 faculty members to participate in one-on-one interviews. Each interview will take no more than 45 minutes of your time. In each interview, I will give a brief introduction to the NW9, the test used to assess Cluster 3’s student learning outcomes. I will then ask how much you expect students to learn as a function of completing Cluster 3 courses. You will then observe the alignment between your expectations and the empirical estimates of learning gains. You will not be asked to provide identifiable information and your responses will be kept confidential. Personal benefits of participating in this study may include additional perspective on student math and science learning gains, information on how much students learn with each Cluster 3 course completed, and the opportunity to participate in a relatively new area of research. This study will benefit the research area by contributing to the nonexistent literature on faculty opinions of student learning gains. Furthermore, this study has the potential benefits of highlighting the strengths of the Cluster 3 curriculum or improving the learning gains of students who complete Cluster 3 courses at JMU. Possible negative consequences of participation are anticipated to be minimal (e.g., personal expectations not being observed in the data).”

After sending this email, I also asked these 9 faculty to participate during their monthly assessment meeting. Three committee members agreed to participate. I also invited via email an acquaintance who teaches quantitative and scientific reasoning general education courses at the institution to participate. All participants had taught at least 1 quantitative and scientific reasoning general education course within the past 10 years and thus were relatively familiar with capabilities of the cohorts assessed in this study. However, two participants were not familiar with the general education assessment process at this institution. To alleviate this issue, I developed a presentation on the NW9 that I showed to all interviewed faculty. This presentation included students’ average pretest performance, examples of test questions, and score reliability. This presentation took no more than five minutes of the interview. I also discussed how quantitative and scientific reasoning faculty developed the test with assessment experts and that faculty mapped items to quantitative and scientific reasoning learning objectives to ensure adequate objective coverage.

Procedures and Materials for Faculty Reactions

I interviewed each faculty member one-on-one in his or her office. Each interview lasted no more than 45 minutes. Before the interview officially began, I gave faculty an IRB-approved consent form and asked them to read and sign it (see Appendix B). I then provided a brief presentation on the purpose of the study as well as on the NW9. After this presentation, I gave the faculty member a sheet of paper (Form A; see Appendix C) with several questions aimed at investigating faculty's expected learning gains (e.g., "How many points do you **expect** students who have completed 1 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?) and desired learning gains (e.g., "How many points **would you like** students who have completed 1 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?). I told the faculty member to answer these questions while keeping in mind the information about the NW9 as well as their own knowledge of and experience with the quantitative and scientific reasoning curriculum.

If faculty said they could not estimate how much they expect and/or desire students to learn after completing 1.5 years of coursework or that estimating their expected and/or desired learning gain is difficult, I asked him/her to write and verbally explain why it is difficult. Two faculty members engaged in this activity. If faculty indicated that they required more information to produce their estimates, I asked him/her to write and verbally explain what information was needed to do so. One faculty member engaged in this activity. After the faculty member wrote these responses, I asked him/her to verbally explain the responses. I took notes during this part of the interview to collect faculty member' verbal responses.

To analyze data, I employed an inductive content analysis. I developed *codes*, simple descriptive text categories, and *themes*, grouping of relevant codes (Charmaz, 2006; Merriam & Tisdell, 2015) from the verbal and written responses from the four faculty. Because there is little literature regarding faculty expectations of student learning gains, I derived these codes from the transcript. Specifically, I utilized a line-by-line approach, where I assigned a code to each line of the transcript; each line of the written and verbal responses was summarized according to a descriptor, or code (Charmaz, 2006; Creswell & Plano Clark, 2011). Related codes were grouped together to determine emergent themes. Only those responses concerning expectation/desire alignment were coded using a priori codes (i.e., ‘aligned’ or ‘not aligned’) to reflect whether the responses are aligned (high desire and high expectation, low desire and low expectation) or not aligned (high desire and low expectation, low desire and high expectation). To ensure the themes I produced accurately captured faculty’s beliefs, I coded responses within faculty to ensure each faculty’s thoughts were adequately represented.

Quantitative and qualitative strands were mixed during the dissemination of the results. Because only a few faculty were involved in this study and did not produce enough data points to conduct statistical tests, I report descriptive statistics. Additionally, the raw (not aggregate) data are reported. However, the raw data is not be attached to any identifying information.

Trustworthiness criteria. To ensure my codes and themes reflect faculty’s perspectives, I engaged in several processes oriented towards increasing *trustworthiness*, or the extent to which my results are unbiased, generalizable, and reliable. To increase *transferability* (i.e., generalizability of results) of faculty opinions, I recruited Cluster 3

Assessment Committee members and asked them to extend the invite to their non-committee colleagues; I also reached out specifically to one of these non-committee members. To increase the *credibility* (i.e., accuracy of interpretations) of my results, I sent my results and my transcripts to my faculty interviewees. To increase credibility (i.e., that my coding accurately represented faculty's beliefs), three of my colleagues, one of whom is external to the institution and area of study, reviewed my transcripts and codes.

CHAPTER FOUR

Results

Hypothesis 1: Collapsing Across Courses, Students Should Have Moderate Gains

Collapsing across the cohorts and number of courses, students, on average, gained 3.72 points on the 66-item NW9 test (see bottom of Table 11). On average, students scored 44.95 at pretest (about 68%) and 48.66 points at posttest (about 74%). This gain was statistically significant ($F(1,1153) = 682.86, p < 0.001$). The eta-squared (η^2) value indicated that 31% of the variance in NW9 scores could be explained by testing time point. Students gained 0.67 SDs on the standardized gain metric and 0.56 SDs on the standardized pretest metric. Thus, results supported Hypothesis 1; students had moderate gains, collapsing across number of courses.

Cohort-specific average pretest scores ranged from 43.92 to 47.26 points, and average posttest scores ranged from 48.37 to 49.30 points. The pretest and posttest scores have comparable variability across and within the cohorts (see Table 11). Across cohorts, students tended to score about 5.50 points above or below the average pretest score, and about 6.00 points above or below the average posttest score. Cohort-specific unstandardized estimates ranged from an average difference score of 1.43 to 3.67 points. The cohort-specific standardized estimates ranged from 0.28 SDs to 0.77 SDs using a standardized gain metric or 0.22 SDs to 0.62 SDs using a standardized pretest metric.

To test whether the variance in the gain scores was related to cohort membership (and hence if the aggregate gain score was masking between-cohort differences in gains), I conducted a between-subjects ANOVA on the gain scores. Results from this ANOVA indicated statistically significant but not practically

different gain scores among the cohorts ($F(4, 1549) = 5.851, p < .001, \eta^2 = 0.02$)¹³. In fact, only about 2% of the variance in gain scores could be explained by cohort membership (i.e., $\eta^2 = 0.02$). Tukey's post hoc tests indicated significant differences in gain scores between Cohorts One and Three, Cohorts Three and Five, and Cohorts Two and Three. However, the unstandardized effect sizes for the difference between the gains scores for Cohorts One and Three (unstandardized difference in gain scores = 3.02), Cohorts Three and Five (unstandardized difference in gain scores = -2.04), and Cohorts Two and Three (unstandardized difference in gain scores = 2.25) were small to moderate. Thus, students at this institution tend to demonstrate similar learning gain on this test across cohorts, which justifies the computation of the aggregate learning gain across cohorts.

Hypothesis 2: Gains Will Increase with Increased Coursework

It is hoped that, although students on average gain 3.72 points on the NW9, this average gain score differs across the levels of completed coursework. Students without any coursework may demonstrate gain scores smaller than 3.72 points, whereas students who have been exposed to multiple courses may demonstrate gain scores larger than this value. To assess the effect of coursework on learning gains, I disaggregated these gain scores by linking them to completed quantitative and scientific reasoning coursework. Specifically, I computed the unstandardized and standardized learning gain estimates for each number of classes collapsing across the cohorts (e.g., learning gain for students who completed one course) and within each cohort (e.g., learning gain for students who completed one course between the years 2013 and 2015). Few students completed zero, five, six, or seven quantitative and

¹³ Ordinary least squares assumptions were checked. Data were distributed normally with no heteroscedasticity across the five groups. Observations were assumed to be independent.

scientific reasoning courses within any of the cohorts; consequently, these gains may be unstable. To produce more stable estimates of these students' learning gains for each number of courses completed (zero through seven), I collapsed across cohorts to produce the average unstandardized and standardized gain estimate (see bottom of Table 11).

Contrary to expectations, gain scores increased after students completed one quantitative and scientific reasoning course but then leveled off after multiple courses were completed. This trend tended to be observed across and within cohorts (see Table 11). For example, collapsing across cohorts, students who did not complete any quantitative and scientific reasoning courses gained 2.69 points on the test; students who completed one course gained 3.85 points; and students who completed three courses gained 3.78 points on the NW9¹⁴.

In contrast, the standardized learning gain estimates increased with each additional course completed. For example, students who did not complete any coursework gained 0.48 SDs on the standardized gain metric or 0.42 SDs on the standardized pretest metric; students who completed three courses gained 0.68 SDs using a standardized gain metric or 0.55 SDs using standardized pretest metric; students who completed six courses gained 0.98 SDs on the standardized gain metric or 0.51 SDs on the standardized pretest metric. The 0.98 SD learning gain estimate is due to low variation in gain scores (i.e., students who completed six courses had similar gain scores). Thus, results did not support Hypothesis 2; learning gain estimates did not increase as number of courses increased.

¹⁴ Only one student completed seven courses. This student gained 2.00 points on the test and also had low pretest (40.00 points) and posttest (42.00) scores. Thus, this student is likely qualitatively different from the student population.

Hypothesis 3: Removing Unmotivated Students Will Increase Learning Gains

Although sample sizes were noticeably reduced after motivation filtering (see Methods and Tables 11, 12, and 13), gain scores did not increase. In the original unfiltered sample ($N = 1554$), students gained on average 3.72 points on the NW9. When I removed students who were unmotivated during the test battery, this estimate decreased (minimally) to 3.53 points ($N = 444$). Likewise, when I removed students who were unmotivated on the quantitative and scientific reasoning test, this estimate decreased (minimally) to 3.47 points ($N = 737$). When I removed students who were unmotivated on either the test or the test battery, the average estimate again decreased to 3.37 points ($N = 413$). The unexpected decrease in unstandardized learning gain estimates with the removal of unmotivated students is because the filtered samples have higher average pretest scores than the unfiltered sample. That is, students in the motivated samples scored higher at the pretest than students in the total sample (see Tables 11-14). Although students in the motivated samples also had higher posttest scores than students in the total sample, the difference between the pretest scores is larger than the difference between the posttest scores.

The standardized estimates filtered for low test session-specific motivation (0.66 SDs on the standardized gain metric; 0.55 SDs on the standardized pretest metric) and low test-specific motivation (0.66 SDs on the standardized gain metric; 0.55 SDs on the standardized pretest metric) were essentially identical to the unfiltered standardized estimates (0.67 SDs on the standardized gain metric; 0.56 SDs on the standardized pretest metric).

Hypothesis 4: The Effort Measure Will Not Affect the Magnitude of Gain Scores

I visually compared test session-specific filtered learning gain estimates to test-specific filtered learning gain estimates collapsing across cohorts with *both* test

session-specific and test-specific effort data (see Table 15). Students in Cohort Two either completed only the test-specific effort subscale or only the test session-specific effort subscale. Thus, I only inspected data from Cohorts Three, Four and Five to address this hypothesis. Given small frequencies in number of courses, I collapsed across Cohorts Three, Four, and Five to create one large sample (see Table 14). Due to the larger size of this aggregated sample, estimates produced from this sample are more stable than estimates produced from the individual cohorts.

For students who completed both the test-specific and test session-specific SOS, I examined if removed students were unmotivated on one or both of these subscales (see Table 10). By examining this agreement, I was able to understand why the two measures produce similar estimates. As well, this examination allowed me to investigate two important outcomes: 1) if one measure identified more students as being motivated than the other at either or both time points, and 2) if the same students who were motivated on the quantitative and scientific reasoning test were still motivated by the end of the testing session, and vice versa.

With respect to the number of students removed from the analyses, a total of 76 students were removed from Cohorts Three-Five due to low motivation on either the test-specific pretest, test-specific posttest, test session-specific pretest, or test session-specific posttest (see Table 10). An essentially equivalent number of students indicated low test-taking motivation on the test-specific SOS ($N = 25$ of the 76 total removed) as the test session-specific SOS ($N = 23$ of the 76 total removed). However, students who were motivated on the test rather than the test battery tended to have higher gain scores when gain scores were disaggregated by completed quantitative and scientific coursework. A small number of students in Cohorts Three, Four and Five indicated low test-taking motivation on *both* SOS versions ($N = 28$ out of the 76

removed using either test). As well, more students indicated low test-taking motivation at posttest than pretest.

Results indicated that filtering using the test-specific effort subscale does not produce different learning gain estimates from the test session-specific effort subscale (see Table 15). Thus, the hypothesis that the two measures would produce similar learning gain estimates was supported. Collapsing across the three cohorts, the two filtered samples had similar overall unstandardized and standardized learning gain estimates (see Table 15). When these average learning gain estimates were disaggregated by coursework, negligible differences appeared between the filtered estimates. For example, students who were motivated on the test and completed one quantitative and scientific reasoning course gained 2.86 points. In comparison, students who were motivated on the test battery and completed one course gained 2.91 points on the NW9. At most, the two filtered samples differed by 0.76 points in gain scores. This 0.76 differences corresponds to a standardized difference of 0.18 SDs on the standardized gain metric or 0.13 SDs on the standardized pretest metric.

Hypothesis 5: Coursework and Personal Characteristics Will Predict Gains

I conducted a multiple regression analysis to determine if coursework predicts learning gains after controlling for personal characteristics. I collapsed across Cohorts Two, Three, Four, and Five to produce an aggregate sample. I dummy coded gender (0 = male, 1 = female). Prior to conducting analyses, I checked Ordinary Least Squares assumptions and these assumptions were met.¹⁵ I retained cases from this sample if the cases did not have missing SAT data. Thus, data from 1001 cases were available for analysis.

¹⁵Results indicated normality and homoscedasticity. Furthermore, relationships between each predictor and the gain scores were linear and not moderated by other predictors.

Recall that there were minor increases in learning gains as students completed more courses (Hypothesis 2). This small effect will likely be further reduced after partitioning out the variance in gain scores shared with prior academic ability and gender. My intent in controlling for the effects of personal characteristics was to showcase the unique effect of coursework on gain scores. However, if coursework does not bivariately relate to gain scores, controlling for the effects of personal characteristics may be moot. Nevertheless, I present the results to test this hypothesis.

Descriptive statistics for unfiltered sample. Students, on average, scored 45.46 points on the pretest (SD of 6.51 points; see Figures 2-5 for distributions by cohort). By posttest, students on average scored 48.89 (SD of 6.85). Thus, students tended to gain 3.43 points (SD of 5.48 points). This distribution of gain scores indicated there is variability to be explained by number of courses, gender, and prior academic ability.

I computed bivariate correlations among gain scores and my predictors (see Table 16). Expectedly, given the results above, coursework did not significantly or practically relate to gain scores ($r = .03$). Gender did not significantly or practically relate to gain scores ($r = .02$), nor did prior academic ability ($r = -.03$). As well, prior academic ability significantly but not practically related to gender ($r = -.21$). Gender significantly but not practically related to coursework ($r = .10$)

In addition to examining the main effects of number of courses, gender and prior academic ability on gain scores, I also examined possible interactions between the three predictors. Before conducting the analysis, I mean-centered prior academic ability to reduce multicollinearity between prior academic ability and the interaction terms that involved prior academic ability (Aiken & West, 1991).

Regression. In the regression, I entered one block containing prior academic ability, gender, and coursework. I then entered a second block containing the three interaction terms (see Table 17). The full model explained a negligible amount of variance ($R^2 = .003$, 95% CI for R^2 : .00, .01, $F(6,994) = 0.47$, $p = 0.83$). I conducted an F_{change} test to determine if the interaction terms could explain significantly more variance in gain scores beyond the variance explained by coursework, prior academic ability, and gender. The interaction terms did not explain a significant amount of variance in gain scores ($R^2_{\text{change}} < .001$, $F_{\text{change}}(3,994) = 0.33$, $p = 0.80$). Thus, the relationship between gain scores and prior academic ability did not appear to be moderated by gender. Likewise, the relationship between gain scores and number of courses was not moderated by gender or prior academic ability.

The reduced model (the model including only coursework, prior academic ability, and gender) also did not explain a significant amount of variance in gain scores ($R^2 = .002$, 95% CI for R^2 : .00, .01, $F(3, 997) = 0.61$, $p = 0.61$). No individual predictors contributed to this reduced model (see Table 17).

I fit this model to the test-specific filtered gain scores to assess if the utility of the model improved after controlling for low test-taking effort. After I removed cases with missing SAT data, 689 cases were available for analysis. Assumptions were rechecked for the sample of students who were motivated; again these assumptions were met¹⁶. I used the same procedures for mean-centering prior academic ability and dummy coding gender.

¹⁶ Data were normal and homoscedastic. Relationships between each predictor and learning gain were linear. The interactions of gender and coursework, and coursework and mean-centered SAT scores, were statistically significant but of negligible magnitude.

Descriptive Statistics for filtered sample. On average, motivated students scored 46.06 on the pretest (SD of 6.34). By posttest, students on average scored 49.58 points (SD of 6.42 points). Students had an average gain score of 3.53 points (SD of 5.30).

Prior to conducting analyses, I examined the bivariate correlations among the variables (see Table 16). As in the unfiltered sample, coursework ($r = .04$) and gender ($r = .05$) did not significantly or practically relate to gain scores; prior academic ability did significantly but not practically relate to gain scores ($r = -.08$). Furthermore, these relationships did not greatly differ from the correlations computed in the unfiltered sample.

Regression. In the motivated sample, the full model did not explain a significant amount of variance in gain scores ($R^2 = .02$, 95% CI: .00, .03, $F(6, 682) = 1.69$, $p = 0.12$). I conducted an F_{change} test to determine if the interactions could explain a significant amount of variance above that explained by coursework, prior academic ability, and gender. As in the unfiltered sample, the three interaction terms did not explain a significant amount of variance in gain scores ($R^2_{\text{change}} = .007$, $F_{\text{change}}(3,682) = 1.23$, $p = 0.30$). That is, the interactions of gender and prior academic ability, the interaction of coursework and prior academic ability, and the interaction of gender and coursework were not statistically or practically significant.

The reduced model (including coursework, prior academic ability, and gender) also did not explain a significant amount of variance in gain scores ($R^2 = .01$, 95% CI: .00, .02, $F(3, 685) = 2.15$, $p = 0.09$). Note, this model explained (within rounding error) an equivalent amount of variance in the unfiltered and filtered samples.

Hypothesis 6: Faculty's Expectations Will Not Match Actual Gain Scores

Recall faculty were asked to state their expectations regarding learning gains. Expectations or predictions of learning gains were defined as the number of points on the quantitative and scientific reasoning test that faculty believed students would gain. Faculty were also asked to state their desired learning gains. Desired learning gains were defined as the number of points on the quantitative and scientific reasoning test faculty would like students to gain.

Faculty tended to have similar expectations of student learning gain. Faculty expected that, after a year and a half of any college coursework, students should gain 4 points on the NW9 (see 'Overall' row in Table 18). When asked to disaggregate the estimated gain scores by coursework, all interviewed faculty expected that students without any quantitative and scientific reasoning coursework should gain from 2 to 4 points on the test. Furthermore, faculty expected learning gains to increase with each additional course completed (see Table 18).

Contrary to the expected learning gain scores, faculty's desired learning gain scores varied greatly. For example, Faculty Two desired students with one and a half years of college coursework to gain 21 points on the test. In contrast, Faculty Three desired students to gain 4 points on the test. When asked to disaggregate desired gain scores by coursework, all but Faculty Two desired that learning gains should increase with coursework completed. Faculty Two desired large and equivalent learning gains no matter the amount of coursework completed.

For two of the four faculty interviewed, faculty's expected gain scores were misaligned with their desired gain scores (see Table 18). Specifically, Faculty One and Two's desired gain scores, collapsing across number of courses completed, exceeded their expected learning gain estimates. Additionally, Faculty One's desired gain scores tended to become larger than his expected learning gain estimates as

number of completed courses increased. Note that Faculty One and Two also orally expressed that they perceived their desired gains as high but their expected gains as low. Faculty Three and Four's desired learning gains aligned with their expected gain score estimates (i.e., they believed their expected and desired learning gains were both reasonable).

Themes regarding expected and desired gain scores. I employed an inductive coding scheme, where codes were derived from transcribed responses rather than from previous studies. Using the written and oral responses from the four faculty members, I coded each faculty's data to unearth why faculty's expected and desired gain scores aligned or did not align, and grouped similar codes to form themes. I conducted several iterations of this coding scheme to ensure accurate representation of faculty responses. Within each faculty, I derived themes regarding their explanations for their expected and desired gain scores, as well as the alignment between these two estimates. I derived these themes from the coded responses to responses from Form A, where faculty estimated their expected and desired gain scores and explained why these estimates aligned or did not align (see Appendix C). I then linked common themes across the faculty. These themes are described below.

Faculty One: Themes about expectations and desires. Within Faculty One, I derived the following themes: students will demonstrate learning gain in college, but learning gain is mostly facilitated by domain-specific coursework; unrealized high desires for student learning gain; expecting low gains but desiring high gains; and students completing different courses will have different learning gains. Prior to seeing the empirical learning gains, Faculty One elaborated that he believed the learning gains would increase with increased quantitative and scientific reasoning coursework, but warned that differences in faculty's instruction of students might lead

to differences in student learning gains. He wrote, “With the wide variety of scientific and quantitative coursework, I believe that the gains made will vary across the courses [taught by other instructors].” For example, a student who completed one biology course taught by Professor A might have greater science learning gains than another student who completed one biology course taught by Professor B. Faculty One was also concerned by what he perceived as differences in faculty expectations of students learning gain. That is, that faculty may teach more or less rigorously depending on how much they expect their students to be able to learn. Faculty One believed that these differences in expectation might lead to variation in student learning gain.

Faculty One explained that the learning gains he desired of students were higher than the learning gains he expected of students “in the real world”. In other words, the learning gains he perceived students are making were lower than what he desired students to make. Thus, he had low expectations for gain scores but still desired high learning gains. Also when describing the misalignment between his expected and desired learning gains, Faculty One further attributed the differences in course instruction to the difference between what he expected and what he desired. That is, that students would gain less than he desired and closer to what he expected due to inconsistent pedagogical practices.

Faculty Two: Themes about expectations and desires. Within Faculty Two, I derived the following themes: students do not have high learning gains, but should learn with increased coursework; high standards for student non-cognitive attributes; unrealized high desires for student learning gains; and expecting low gains but desiring high gains.

At the beginning of our interview, Faculty Two lamented that students did not appear to be learning from their classes. He gave an example from his own class,

where students who had completed a statistics course could not explain a p -value. However, he still desired that students learn as they complete more quantitative and scientific reasoning courses. He held the conviction that student improvement (i.e., learning) does not necessarily mean that students will perform highly on the quantitative and scientific reasoning test. He also explained how he expected student to have integrity (i.e., should not cheat on their tests) and a desire to learn material (i.e., student non-cognitive attributes). Similar to Faculty One, Faculty Two expressed unrealized high expectations for student learning gains. As he elaborated, “I keep the bar high because I think that’s where it belongs.”

With respect to gain scores, Faculty Two said that he had low expectations but high desired gain scores, a theme identical to that derived from Faculty One’s responses. He also desired high gain scores for all students, which is evidenced by his high quantitative estimates for students with any level of completed coursework. When writing these estimates, he positioned himself as an ‘idealist’ and explained that he would like students to answer all the items on the test correctly. I understood this to mean that the best possible scenario for Faculty Two is one where all students have high learning gains.

Faculty Three: Themes about expectations and desires. Within Faculty Three, I derived the following themes: difficult to estimate learning gains; belief that expectations are reasonable; and students should learn from general and domain-specific courses.

Faculty Three found it difficult to estimate students’ gain scores for each number of courses completed, especially for students with one or two courses, saying, “It’s so hard!” However, she did not explain why she found it difficult. Nonetheless, she explained that the amount of learning gain that she expected was also the amount

of learning gain that she desired. She further elaborated that, even though her expectations were aligned with her desires, her estimated gain scores were reasonable (i.e., attainable). Faculty Three did expect that students should have some learning gain without completing quantitative and scientific reasoning courses, as quantitative and scientific reasoning skills are taught in other general education courses (e.g., economics). She also believed and desired that quantitative and scientific reasoning skills would increase due to increased courses in quantitative and scientific reasoning *and* increased courses in other domains. In other words, gain scores should increase as number of courses increase.

Faculty Four: Themes about expectations and desires. Within Faculty Four, I derived the following themes: expectations framed through student familiarity; students will demonstrate learning gain in college, but learning is mostly facilitated by domain-specific coursework; and desire for students to learn from quantitative and scientific reasoning coursework.

Faculty Four explained that his expectations resulted from his experiences with his students' learning in his courses. When he had first started teaching, his expectations had been higher. Over time, however, his expectations had decreased due to his increased familiarity with how much his students were learning. Similar to Faculty One and Three, Faculty Four expected that, due to increased maturity, students without quantitative and scientific reasoning coursework should demonstrate some learning gains. Nevertheless, he explained he did not have an opinion on how much he desired students to learn. He expected and desired, though, that gains scores should increase with increased coursework.

Prior to providing his answers, Faculty Four stated that it was difficult to estimate student learning gains without knowing how many quantitative and scientific

reasoning courses the students completed. He specified that students do not learn everything they are taught. Thus, he did not think that it was realistic for students to gain 20 points on the test. He explained that, given his familiarity with students, his expectations of their learning gains were reasonable.

Common themes among the faculty. Though only three faculty verbally mentioned this belief, as evidenced by their written gain score estimates, all faculty believed that students without any quantitative and scientific reasoning coursework should demonstrate some learning gains. Furthermore, all four interviewed faculty expected to some extent that learning gains would increase with increased quantitative and scientific coursework. Faculty One and Two discussed their unrealized high expectations for student learning gain. These two faculty believed, given current faculty instruction and observed poor student learning, that they should expect low learning gains. Faculty Three and Four believed their expectations were reasonable and realistic. However, all faculty stated that they desired high learning gains for their students. In other words, all faculty believed that their desired gain scores were high.

Numerical alignment between expected and empirical gain scores.

Faculty's expected and desired learning gain estimates were mostly misaligned with the empirical learning gain estimates. Collapsing across the number of courses completed, faculty's expected learning gain estimates (median of 4 points) and desired learning gain estimates (median of 5 points) were slightly larger than the empirical learning gain estimates (3.47 points).

However, disaggregating these estimates by coursework revealed greater misalignment (see Table 18). Faculty One, Two, and Three's expected gain scores overestimated the empirical gain scores for students who completed at least one quantitative and scientific reasoning course. All faculty's expected gain scores

increasingly diverged from the empirical gains scores as the number of course completed increased. That is, the faculty expected a relationship between number of courses completed and learning gains yet there was no empirical relationship. Faculty One, Three, and Four's desired gain scores increasingly diverged from the empirical gain scores as the number of completed courses increased. Faculty Two's desired gain scores consistently did not align with the empirical gain scores.

CHAPTER FIVE

Discussion

In this study, I investigated the impact of college coursework on student learning gains, a call put forth years ago by the higher education research community and the federal government. Specifically, this study was meant to address how much students *change* in knowledge and capabilities (i.e., learning gain) rather than what knowledge and capabilities students have at a particular point in time (i.e., student competency). Although both concepts are important outcomes, they are relatively independent (e.g., a student who is competent may not have learned and a student who has learned may not be competent) and answer two distinct questions. This study focused on answering the question of how much students are learning from their college coursework.

Findings from this study imply that students' average quantitative and scientific reasoning learning gains over the first two years of college may be larger than what has been found in previous studies but still less than desired. Students gained 3.72 points on a 66-item test of quantitative and scientific reasoning, without taking into account the amount of completed quantitative and scientific reasoning coursework. Contrary to prediction, gain scores were unrelated to the number of quantitative and scientific reasoning courses completed. Moreover, and differing from the literature, the gain scores were also unrelated to students' personal characteristics.

Unexpectedly, learning gain estimates showed no discernable improvement when corrected for low test-specific *or* test session-specific effort. When the gain scores were disaggregated by completed coursework, these gain scores did not align with what quantitative and scientific reasoning faculty desired and expected. In sum,

although students appear to be making modest gains in quantitative and scientific reasoning, it does not seem that there is a link between these modest learning gains and students' quantitative and scientific reasoning coursework. Given this summary of results, below I discuss these findings with respect to theory and prior research, as well as implications for student learning assessment and learning improvement processes.

Collapsing Across Courses, Students Appear to Have Moderate Gains

Based on the limited previous research on student learning gains (Blaich & Wise, 2011; Pascarella & Terezini, 2005; Roohr et al., 2016), I hypothesized that students would have what I considered moderate learning gains in quantitative and scientific reasoning after experiencing one and a half years of any college coursework. Recall, students may or may not have completed courses in the domain of quantitative and scientific reasoning during the 1.5 years. Indeed, students demonstrated both unstandardized and standardized gain estimates that aligned with *my* standard of moderate gains. These moderate gains corresponded to an average of 3.72 points on a 66-item test. Additionally, students at this institution demonstrated greater aggregate learning gains than what has been found in prior studies (e.g., Blaich & Wise, 2011; Pascarella & Terezini, 2005; Roohr et al., 2016). As an aside, this gain score aggregated across course completion was similar to what most faculty expected and desired when averaging student learning gains across students with different amounts of course exposure.

The efficacy of coursework completed within the first two years of college had been called into question with learning gain results from Roohr et al. (2016). She and her colleagues found that students with one or two years of college coursework achieved statistically significant but practically small estimated learning gains (e.g.,

standardized average math gain of $d = 0.22$). The authors explained that one or two years of college coursework had also previously been linked to small estimated learning gains; thus, it appeared students were not making learning gains in the first half of their college careers. Roohr et al. (2016) believed students' acclimation to college may have led to this small effect: "At the beginning of their college career, students may need some time to get used to the environment (both academically and socially), so the learning gain during the first two years is comparatively low." (Roohr et al., 2016, p. 11).

Nonetheless, results from this study indicate that, whether or not they are acclimated to the college culture, students are demonstrating moderate learning gains. That is, this small learning gain in math after one/two years was not supported in the current study; students who had completed one and a half years of college coursework had average estimated standardized gains of $d = 0.67$ (standardized gain score metric) in quantitative and scientific reasoning. Second-year students in the current study, with the exception of one cohort, gained more than four/five-year students in the Roohr et al. (2016) study ($d = 0.41$ in Roohr et al., 2016).

Improved sampling techniques in the current study may account for the incongruity in findings. A large number of students at this institution were randomly assigned to complete the quantitative and scientific reasoning test. Roohr et al. (2016) did not employ these methods; they obtained their estimates from a small, conveniently sampled group of students. Thus, it is likely that the Roohr et al. (2016) sample had smaller gains than the population. It could also be that the curriculum completed by students in the Roohr et al. (2016) study was not as clearly tied to student learning outcomes or the instrument of measurement as both are at this institution. Furthermore, the small gains in the Roohr et al. (2016) study, compared to

those found in the current study, may be due to particular characteristics of this non-random sample. When comparing the one/two year sample of students in Roohr et al. (2016) to the sample from this study, the Roohr et al. (2016) sample had a higher percentage of female students and a higher average SAT score (the two samples had similar percentages of white students.)

The moderate estimated learning gains from this study suggest that students are learning in college. Given solely these aggregate learning gains, one may assume that learning does not need to be improved; thus, pedagogy or curriculum modifications do not need to be made. That is, one may believe that the current curriculum is adequately designed and structured to support student learning. As well, researchers who report aggregate gain estimates likely assume that these gains are due to college coursework. Nevertheless, not all students will complete courses in the specific domain on which they are tested. Thus, aggregate learning estimates do not adequately indicate how college affects student learning gains.

Gains Did Not Increase with Increased Coursework

Institutions must begin to assess the impact of coursework on student learning to ensure students are learning from their coursework. That is, it is not appropriate to assess overall student learning gains and infer these gains are due to coursework. Given previous research (Hathcoat et al., 2015; Pascarella & Terezini, 2005; Pastor et al., 2007) that indicated completing domain-specific coursework should lead to increased knowledge in that domain, I hypothesized that students' quantitative and scientific reasoning learning gains would increase with additional quantitative and scientific reasoning coursework.

Unexpectedly, estimated learning gains did not appear to increase after completing more than one quantitative and scientific reasoning course. Although students who did

not complete any courses gained less than students who completed coursework, students who completed coursework had similar estimated learning gains. For example, students who completed one quantitative and scientific reasoning course had similar learning gains to students who completed three quantitative and scientific reasoning courses.

Results from Hathcoat et al. (2015) foreshadowed these results (partial credit completers gained $d = 0.42$ or $d = 0.55$ depending on the cohort assessed, credit completers gained $d = 0.46$ or $d = 0.52$ depending on the cohort assessed). However, I dismissed these findings due to the credit hour coding scheme the authors employed. Based on findings from Pastor et al. (2007), I expected that students who completed one course would have moderate learning gain estimates ($d = 0.54$ or $d = 0.41$ depending on whether the history or political science course was completed, or 4 out of 81 points) and students who completed two or more courses would have large learning gain estimates ($d = 0.90$, or 7.52 out of 81 points).

Why were the results from Pastor and colleague (2007) not replicated in the current study? Though these analyses provide no explanation as to *why* students are not learning, they instead lead to possible hypotheses; several of these hypotheses were addressed in the current study through motivation filtering and faculty interviews. First, Pastor and colleagues (2007) investigated history/political science learning gains rather than math/science learning gains. It is plausible that students at this institution do not learn as much from their quantitative and scientific reasoning courses as they do from courses in other domains. Thus, the relationship between magnitude of learning gains and coursework may be moderated by course domain. A second explanation is that students' test-taking motivation augmented the estimated learning gains. That is, although Pastor et al. (2007) did not measure students' effort

on the history and political science test, these students might have expended greater test-taking effort than students in the current study. As explained below, I addressed this possibility by computing the learning gains of motivated students.

Third, the quantitative and scientific reasoning test may not align with the content taught in the quantitative and scientific reasoning courses. Items on the test are mapped to specific learning objectives of the quantitative and scientific reasoning curriculum (Curtis, 2016). However, the learning objectives do not appear to be mapped to the courses. Consequently, students may be learning quantitative and scientific reasoning concepts, but these concepts are not assessed on the test.

A fourth, weighty possibility is that these college courses may not be as efficacious as previously believed. If students are not learning from their coursework, then learning improvement processes must be implemented. Given these undesirable learning assessment results, faculty should modify curricula (e.g., different pedagogy, additional courses, better course sequencing) and then reassess to evaluate if the modified curricula engenders greater student learning. In order to understand if faculty believed poor coursework failed to increase learning gains, I interviewed the faculty who design and teach these courses. The faculty interviews, discussed in further sections below, supports the need for learning improvement assessment.

After Removing Unmotivated Students, Learning Gains Did Not Increase

Given that learning gains did not increase as much as expected as quantitative and scientific reasoning coursework increased, one may question the quality of the data. Are the disaggregated, estimated gains inaccurate estimates of actual gains? Could the estimated gains be invalid due to low test-taking motivation? Empirically, students had similar learning gain estimates regardless of their level of motivation or completed coursework. Although motivated students did not gain more than the total

unfiltered sample of students, pretest and posttest scores of motivated students tended to be higher than those from the total sample of students. In other words, performance estimates – but not gain scores – were attenuated by low test-taking effort.

Consequently, the lack of relationship between learning gains and coursework does not appear to a function of test-taking motivation.

Even though students' motivation did not appear to affect their learning gains, students' pretest and posttest scores were influenced by low test-taking motivation. The effect of test-taking effort on test performance is well-documented (e.g., Cole et al., 2008; Mathers et al., 2016; Myers et al., 2016, Finney et al., 2016) and was supported in this study. Specifically, after filtering students who were unmotivated at pretest or posttest, average pretest and posttest scores increased. Nonetheless, the focus of this research is not on performance estimates but on learning gain estimates, which did not substantially change post-filtering.

These results contrast with previous work on learning gains and test-taking motivation. DeMars and Wise (2010) found that low effort attenuated learning gain estimates (difference of $d = 0.30$). With the exception of students who completed four courses, these findings were not replicated in the current study. However, the current study used self-report scales to measure test-taking effort whereas the prior study used RTE. Although both types of measures are used for motivation filtering, perhaps the different conceptualizations of motivation (affect versus behavior) can account for this discrepancy.

Furthermore, researchers have demonstrated that pretest effort relates negatively to student learning gain, whereas posttest effort positively relates to student learning gain (Finney et al., 2016). This result indicates that a student who put forth good effort at the pretest but failed to put forth effort on the posttest would have an

attenuated gain score estimate. In line with these findings, gain scores computed in this study decreased (minimally) after removing unmotivated students. Unmotivated students were filtered at pretest and posttest, which led to a higher average pretest score after filtering than prior to filtering, as well as a higher average posttest score after filtering than prior to filtering. However, more unmotivated students (not including those who were unmotivated at both time points, $n = 30$) were filtered at pretest ($n = 91$) than posttest ($n = 60$). Subsequently, the difference between pre- and post-filtered scores and pre- and post-unfiltered scores was greater for the pretest than posttest. Because the average pretest score increased more after filtering than the posttest scores, estimated learning gains (minimally) decreased after filtering.

This small decrease provides better empirical evidence that researchers are underestimating performance estimates rather than misestimating learning gains. In other words, researchers who do not filter unmotivated students are unlikely to produce invalid learning gain estimates but are likely to produce invalid performance estimates. Consequently, these results necessitate that faculty and assessment practitioners work to increase students' test-taking effort in order to ensure valid student performance estimates. However, researchers who are only interested in estimating learning gains do not need to be as preoccupied with students' low test-taking effort.

Test-specific and Test Session-Specific Gain Scores Are Similar

To further explore if test-taking effort impacted learning gain estimates, unmotivated students were filtered using two different measures of effort: test-specific effort and test session-specific effort. I hypothesized that the two measures would produce similar learning gain estimates. An equivalent number of students were removed due to low effort on the test, on the battery, or on both the test and battery.

Furthermore, few students indicated being unmotivated at both time points (i.e., unmotivated at pretest and posttest). The average learning gain estimates from these three samples (i.e., test-specific filtered, test session-filtered, and test-specific and test session-specific filtered) were essentially equivalent, even when disaggregated by coursework.

Results from this study partially corroborated findings from Hathcoat et al. (2015). Specifically, the authors found that more students indicated low test-taking effort on the quantitative and scientific reasoning test than the test battery. This result was not supported in the current study. In spite of this disproportion, students in the Hathcoat et al. (2015) study who were motivated on the test had similar performance estimates to students who were motivated on the battery. This result was supported in the current study. Together, findings from Hathcoat et al. (2015) and the current study suggest that using either measure to remove unmotivated students will result in the same inferences regarding student performance or student learning gains.

Coursework and Personal Characteristics Did Not Predict Learning Gains

I hypothesized that, after accounting for the effects of students' personal characteristics, coursework would predict the quantitative and scientific reasoning gain scores. Results from hypotheses two through four indicated that coursework did not affect learning gains. In accordance with these results, coursework did not significantly predict gain scores when controlling for personal characteristics (whether predicting gains from unfiltered or filtered data).

Although coursework did not predict learning gains, it was worthwhile to explore the impact of personal characteristics on learning gain. Unexpectedly, gender and prior academic abilities did not predict gain scores. Prior research found that male students gain more in math than female students (Finney et al., 2016) and students

with higher academic abilities tend to gain less than their lower-ability peers (Pastor et al., 2007). On the other hand, Roohr and colleagues (2016) found similar results to the current study; gender, prior academic ability, and time spent in college (their proxy variable for coursework) did not affect learning gain estimates. The gain scores estimated by Finney and colleagues (2016) were of similar magnitude and variability to those found in the current study; thus the lack of prediction was not due to range restriction. Furthermore, the lack of a statistically significant relationship is evidenced by the small point-increase in mean gain score with each completed course. Nonetheless, it is surprising none of the theory-based variables in this study predicted gain scores given the adequate variability. However, the gain scores from Roohr et al. (2016) have much less variability than the gain scores in the current study. This lack of variation may explain the null results in the Roohr et al. (2016) study but does not assist in understanding the current study's results.

The null effects of personal characteristics on learning gains, if true, hold implications for theory and practice related to pedagogy/curriculum modifications as well as gain score modeling. A research question yet to be fully answered is the question of whether males are more adept at math and science than females. The insignificant effect of gender on learning gains suggests that there is not a math and science learning gap between male and female college students. Thus, pedagogy or curriculum modifications do not need to be made to increase the learning gains of one gender. The null effect of prior academic ability holds similar implications. If students of lower ability gained less, then remedial courses or modifications to pedagogy might have been called for. Given that higher and lower academic ability students have similar gains, the necessity of these interventions is moot. With respect to modeling, the effects of students' gender and prior academic abilities on learning

gains may not need to be controlled for to accurately estimate the impact of coursework on learning.

However, other variables not included in the investigated models may predict gain scores. Previous content exposure is one such characteristic that might affect student learning gains. In this study, I only included students who earned credit from this institution (i.e., did not have AP or IB credit), thus eliminating any covariance between previous content exposure and gain scores.

Two other potential predictors related to coursework are student interest and self-efficacy. Student interest might indirectly affect learning gain estimates through students' engagement in previous and current course material. Self-efficacy is analogous to the expectancy component in EV theory. To reiterate, expectancy, or efficacy, refers to a student's belief that he will be able to perform a given task. Thus, a student who believes he is able to learn in a course will likely have increased learning gains.

If these variables do have positive relationships with learning gains, then making course material relevant to students or bolstering students' confidence in their quantitative and scientific reasoning skills may increase learning gains. An academic intervention used by Hulleman, Kosivich, Barron, and Daniel (2016) shows promise with respect to increasing students' course interest. Hulleman et al. (2016) required students to make connections from course material to their lives while completing an introductory psychology course; this process was shown to increase students' interest in course material. However, the use of such interventions presents a thorny issue: is it the faculty's responsibility to increase students' interest in and engagement with the course material? Might this engagement be better assisted by allowing students to

complete the courses they are interested in? This conversation is best left to faculty during learning improvement assessment, which I discuss below.

Faculty's Desired Gains Scores Did Not Match Actual Gain Scores

As anticipated, faculty's expected (i.e., how much they expected students to gain on the test) and desired (i.e., how much they hoped students to gain on the test) gain scores were larger than the empirical gain scores. Interestingly, faculty had similar expectations of student learning gain yet differed on whether they believed their expected learning gains were low or reasonable. This disagreement about what is considered low or reasonable learning gain may indicate that faculty need to discuss how much students should gain from their courses.

Discrepancies in expected learning gains are problematic for other reasons, as well. Research has linked faculty expectations to magnitude of student performance (Timmermans et al., 2016). Consequently, a professor who has low expectations of student performance or student learning may inadvertently create a self-fulfilling prophecy. Faculty One, in fact, discussed this issue when explaining why he had low expectations for student learning gain even though he desired high learning gains. Another concern divulged by Faculty One relates to *implementation fidelity assessment*, the process of determining if a program or curriculum is taught and received in the intended manner (Gerstner & Finney, 2013). As Faculty One explained, students may have varying learning gains depending on the instruction they receive. It is possible that disagreement over how much learning should be expected may indicate that students are not equally instructed in curriculum learning objectives. For example, a professor who does not think students are capable of learning a particular math concept may not emphasize that concept when teaching her courses, even if that concept is meant to be covered in all quantitative and scientific reasoning

courses. Thus, implementation fidelity assessment, as a part of learning improvement assessment, could be necessary to establish if students are receiving the intended curriculum. This additional assessment is especially pertinent given concerns as to whether the concepts taught in the quantitative and scientific reasoning courses are those concepts specified in the learning objectives. Implementation fidelity assessment can additionally be used to pinpoint areas of weakness in the curriculum if students are not receiving the intended curriculum (Gerstner & Finney, 2015).

Given the misalignment between the empirical and expected/desired gain scores, pedagogy and curriculum modifications may be necessary. If the curriculum is not effective, which may be proved through implementation fidelity assessment, then faculty must modify the current curriculum to improve student learning. The need for learning improvement also relates to the misalignment between expected and empirical gain scores. Thus, after discussing the misalignment between expected and empirical gain scores below, I then describe what this learning improvement process would entail.

The implications of the misalignment between the empirical gain scores and faculty's expectations are threefold and speak to the metric one uses when reporting gain scores, engaging faculty in setting expectations of growth, and assisting faculty in making curriculum-related modifications for learning improvement. First, these findings call into question how learning gain estimates are reported and interpreted in the literature. Most researchers interpret their standardized estimates using Cohen's (1988) values (e.g., Blaich & Wise, 2001; Roohr et al., 2016), likely for ease of comparisons with other studies as well as convention. As I have hopefully demonstrated in this study, solely interpreting standardized estimates does not provide a clear or accurate depiction of student learning gains. Recall that I aligned my

unstandardized gain score benchmarks with Cohen's arbitrary but widely-used effect sizes (Cohen, 1992). That is, my three-point gain benchmark corresponded to conventional, moderate standardized learning gain estimates. Without interviewing faculty, I concluded that students at this institution demonstrated moderate learning gains. In contrast, two of the four faculty believed that their expected gain scores, which aligned with or were higher than my moderate benchmark, were low. As well, the two faculty with "reasonable" expectations also expected and desired gain scores larger than three points. Therefore, interpreting results on the *test* (i.e., unstandardized) metric provides a clearer understanding of student learning gain.

The discrepancy between *my* learning gain benchmarks and faculty's expected and desired learning gains, as well as the discrepancy between the empirical gain scores and faculty's expected and desired learning gains, speaks to the second implication. That is, faculty must be involved when setting expectations of student learning gains and evaluating whether these standards are met. When describing best practices for student learning outcomes assessment, Banta and Blaich (2010) explicitly discussed the importance of involving faculty when conducting student learning outcomes assessment and interpreting assessment findings. The authors state, "If faculty do not participate in making sense of and interpreting assessment evidence, they are much more likely to focus solely on finding fault with the conclusions than on considering ways that the evidence might be related to their teaching." (Banta & Blaich, 2010, p. 24).

I both disagree and agree with this statement. The faculty I interviewed were not defensive nor antagonistic when discussing the efficacy of the quantitative and scientific reasoning curriculum. I do, however, agree with Banta and Blaich's (2010) comment that faculty must participate in interpreting assessment results. Again, I

considered student learning gains to be moderate (based on relatively arbitrary values). Faculty, on the other hand, did not consider the learning gains moderate. Moreover, if faculty participate in setting expectations related to student learning gain, they may be more likely to use these learning gain assessment results for program improvement; this possibility leads to the next implication.

The third implication of the misalignment between empirical and faculty-estimated gain scores is the need to assist faculty during learning improvement assessment processes. This assistance is paramount in order to improve either assessment of learning gains (if measure does not align with course content) or the quantitative and scientific learning gains at this institution (if the curriculum is not effective). As Fulcher et al. (2014) have explained, faculty often do not receive assistance on how to use assessment results to improve student learning. At the most basic level, using results requires faculty to implement modifications to pedagogy or curriculum after determining learning gains (as was done in this study).

My interviews with the faculty indicate that, in order to facilitate student learning, faculty must first set an expectation of student learning gain as well as work with assessment experts to ensure the measure aligns with course content. Faculty at this institution have set performance standards for students' quantitative and scientific reasoning abilities (Hathcoat et al., 2015). Students may meet performance standards, but it is possible for students to achieve competency yet gain little or less than expected. Furthermore, assessing competency answers a different question than assessing learning gains and can lead to different conclusions regarding students' abilities and the coursework meant to enhance those abilities. It is therefore necessary that faculty set a learning gain standard in addition to a performance standard.

Assuming that the measure is aligned with course content, informed changes to the quantitative and scientific reasoning curriculum may be necessary. What would this modified curriculum entail? In their discussion about quantitative and scientific reasoning learning gains, Hathcoat et al. (2015) made the point that students at this institution are exposed to a breadth of quantitative and scientific concepts, but may not have experienced much depth in content. Thus, a greater depth of content may be required. Additionally, research on service-learning faculty (i.e., faculty who require their students to apply course material in real world settings) suggests that these faculty tend to find that student learning improves when students are able to apply their knowledge beyond the classroom (Darby & Newman, 2014). It could be that the course curriculum needs to be modified to facilitate these experiences and thus engage students in coursework and facilitate student learning.

How would one know whether or not the modifications benefit students? In other words, how could faculty demonstrate learning improvement? First, faculty should come to a consensus on what aspects of the curriculum (e.g., content, structure, pedagogy) influence learning gains through use of implementation fidelity assessment and, moving forward, implement one or several modifications. As incoming and second-year students are both assessed during the academic year at this institution, assessment experts will be able to compute the learning gains of the first cohort of students to receive this modified curriculum. With the assistance of these assessment experts, faculty can compare the learning gains computed from this study to those learning gains from the cohort who experienced the modified curriculum. In other words, faculty and assessment experts, together, must *re-assess* student learning gains in determine whether student learning gains were improved by the modified curriculum.

Limitations

As with most applied research, this study has several limitations. A doctoral candidate at this institution conducted a missing data study on Cohorts Three and Four. In these cohorts, only a small - albeit random - section completed the full NW9. Consequently, there is a chance that the learning gain estimates computed from these smaller subsections do not represent the learning gains of the students in these cohorts. Additionally, this study investigated learning gains at just one institution that also has an extensive history and strong culture of student learning assessment. More research on student learning gains is needed across different institutions.

Several threats to validity were also present, due to the quasi-experimental nature of the study. Within each cohort, only a small number of students completed five or more quantitative and scientific reasoning courses. Even after collapsing across the cohorts, the total number of students who completed at least five quantitative and scientific reasoning courses remained relatively small. As well, students self-select to either complete or not complete these courses based on interests or what fits their academic schedules. A last threat to validity is attrition; the students in my sample may be more academically adept than students who are no longer enrolled at this institution. Findings based on students with these amounts of quantitative and scientific reasoning courses (magnitude of learning gain estimates, coursework as a non-significant predictor of learning gain) thus may be unstable or sample-dependent.

A similar issue is the need to assess coursework effects over a greater period of time. That is, students may demonstrate larger learning gains after completing three or more courses in a given domain. Nonetheless, most students are assessed before they have completed their quantitative and scientific reasoning curriculum

requirements (i.e. completed 10 credit hours of quantitative and scientific reasoning coursework).

To prevent academic ability from confounding results, I used different effort cut scores for Cohort Four. Thus, I retained more “unmotivated” students in this cohort. Only three of the five cohorts had data on both measures of test-taking effort. Consequently, not all available learning gain data could be used in this study (i.e., reduced sample size). As mentioned above, sample sizes also decreased when unmotivated students were removed from the sample. Thus, the estimated gain scores, especially for students who completed five or six courses, may be unstable.

Total ACT scores can be computed using ACT Math, ACT Reading, and ACT English scores (Dorans, 1999). However, the samples in this study tended to have data on either ACT Math and ACT Reading or ACT Math and ACT English. Thus, the total ACT scores computed in this study may not be accurate. As well, the ACT scores from students without SAT scores were converted to the SAT metric to compute one total prior academic ability indicator. As this transformation is not exact, there may have been loss of precision with respect to prior academic ability estimates.

Half of the interviewed faculty were unaccustomed with how quantitative and scientific reasoning is assessed at this institution (e.g., unfamiliar with the NW9). I provided a brief overview of the data collection design and measure in order to assist faculty in developing their expectations. However, this overview may not have been sufficient training. Faculty’s gain score expectations may change with better understanding of the measure and the standard setting procedure. In this initial study, setting an expectation of student learning gain may have been more difficult than anticipated (as Faculty Three indicated). At this institution, faculty have worked with assessment experts to set performance (i.e., competency) standards. Being able to

shift from performance framework to student learning framework may require more than a 45 minute interview.

Lastly, when coding qualitative data, researchers may bring their own biases into the data analysis. Although I hope that my position as a former student and assessment consultant at this institution has not clouded my data interpretations, this risk is still likely.

Future Research

Although this study adds to the literature on student learning gains, the field would benefit from continued applications of longitudinal methods. It is important to reemphasize that faculty considered the gain scores to be small. As this study should demonstrate, standardized learning gain estimates may misrepresent how much students are learning and confuse faculty. Likewise, other indices of “learning”, such as residualized gain scores or cross-sectional difference scores, may also prove difficult for faculty to interpret. This confusion could likely prevent use of assessment results, as faculty may draw erroneous conclusions about student learning from these indices. Therefore, I recommend researchers evaluating learning in higher education estimate and interpreting both the unstandardized and standardized learning gain estimates. Future studies could also examine faculty reactions to the empirical gain scores after faculty provide their expected and desired gain scores. Another powerful study would be an investigation of how well faculty are able to interpret common indices of “learning”. That is, an investigation of how well faculty are able to interpret assessment results such as unstandardized gain estimates, standardized estimates, residualized gain scores, and cross-sectional difference scores. The current study was an initial exploration into setting expectations of learning gain. A next step would be a

formal standard setting study, where faculty set standards of learning gain rather than competency.

Future studies could easily address the limitations described above. For instance, researchers should collect more data from students with at least five courses completed in a given domain. Researchers could also collect more precise estimates of prior academic ability. Although the two motivation measures used did not produce different learning gain estimates, this study did not investigate whether or not test session and test-specific effort are truly distinct constructs. An invariance study would easily provide insight into this issue.

As a final recommendation, higher education would benefit from more research on faculty expectations of student learning *gain*. That is, more research is needed on whether students are gaining as much as faculty expect them to, rather than research on how many students are meeting competency standards at pretest and posttest. This research was a small section of the current study and thus was not fully explored. A phenomenological or grounded theory approach to investigating faculty expectations may be better suited to unpacking this phenomenon.

Conclusions

Results from this study provide a tenable answer to the U.S. Department of Education's question of why American college students are falling behind their international peers (U.S. Department of Education, 2006). That is, students are making modest learning gains that may not be related to their coursework. Higher education has been slow to assess student learning gains, and thus we have remained largely ignorant to the magnitude of student learning occurring on our college campuses.

Table 1. *Mandates and recommendations regarding student learning data collection post-Spellings report.*

	Mandates	Recommendations
Federal	<p>“The institution evaluates success with respect to student achievement consistent with its mission. Criteria may include: enrollment data; retention, graduation, course completion, and job placement rates; state licensing examinations; student portfolios; or other means of demonstrating achievement of goals.” (SACSCOC, 2012, p.39)</p>	<p>“Higher education institutions should measure student learning using quality-assessment data...in order to improve the quality of instruction and learning” (U.S. Department of Education, 2006, p.33)</p> <p>“The results of student learning assessments, including value-added measurements that indicate how much students’ skills have improved over time, should be made available to students and reported in the aggregate publicly.” (U.S. Department of Education, 2006, p.33)</p> <p>“Accreditation agencies should make performance outcomes, including completion rates and student learning, the core of their assessment as a priority over inputs or processes. A framework that aligns and expands existing accreditation standards should be established to (i) allow comparisons among institutions regarding learning outcomes and other performance measures, (ii) encourage innovation and continuous improvement...” (U.S. Department of Education, 2006, p. 34)</p>
Regional	<p>“The institution engages in ongoing, integrated, and institution-wide research-based planning and evaluation processes that (1) incorporate a systematic review of institutional mission, goals, and outcomes; (2) result in continuing improvement in institutional quality; and (3) demonstrate the institution is effectively accomplishing its mission” (SACSCOC, 2012, p. 18)</p> <p>“The institution has developed an acceptable Quality Enhancement Plan (QEP) that includes an institutional</p>	

process for identifying key issues emerging from institutional assessment and focuses on learning outcomes and/or the environment supporting student learning and accomplishing the mission of the institution.” (SACSCOC, 2012, p. 21)

“The 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 areas: 3.3.1.1 educational programs, to include student learning outcomes” (SACSCOC, 2012, p.27)

“The institution identifies college-level general education competencies and the extent to which students have attained them” (SACSCOC, 2012, p.29)

State	<p>“Each college or university may choose to employ either absolute assessment measures or those that demonstrate the value-added ‘contribution the institution has made to the student’s development.’” (SCHEV, 2007, p.2)</p> <p>“The Commission further identified six areas of knowledge and skills that cross the bounds of academic discipline, degree major, and institutional mission to comprise basic competencies that should be achieved by all students completing a degree program at a Commonwealth institution of higher education—namely, Information Technology Literacy, Written Communication, Quantitative Reasoning, Scientific Reasoning, Critical Thinking, and Oral Communication.” (SCHEV, 2007, p.2)</p>	<p>“Each institution should continue to be responsible for implementing an assessment program that is congruent with its mission and goals; provides the kind of data needed for informed decision-making about curricula; and offers both policymakers and the general public useful information on student learning.” (SCHEV, 2007, p.3)</p> <p>“Assessment should continue to <i>fit</i>, rather than <i>drive</i>, the institution. It should be reasonable in its requirements for time, resources, and personnel and should, ideally, be integrated with the institution’s larger framework for continuous improvement and public accountability. It should also employ both valid and reliable measurements of educational experiences and student learning.” (SCHEV, 2007, p.3)</p>
-------	---	---

“The Code of Virginia, §23-9.6:1, charges the State Council of Higher Education for Virginia (SCHEV) with various duties and accords Council the authority to carry out those duties.

Duty #6

- To review and require the discontinuance of any academic program which is presently offered by any public institution of higher education when the Council determines that such academic program is (i) nonproductive in terms of the number of degrees granted, the number of students served by the program, evidence of program effectiveness, or budgetary considerations, or (ii) supported by state funds and is unnecessarily duplicative of academic programs offered at other public institutions of higher education in the Commonwealth...” (SCHEV, 2013, p.1)

“Following completion of the fifth year enrollment data collection, SCHEV will provide official notice to four-year public institutions and Richard Bland College of academic degree programs that fail to meet quantitative standards for FTES enrollment and numbers of graduates.” (SCHEV, 2013, p.2)

“Assessment should continue to focus on the improvement of learning while providing meaningful demonstration of accountability. It should continue to employ the six core areas and explore options to address the Council’s preferred ‘value-added’ approach that speaks to demonstrable changes as a result of a student’s collegiate experience.” (SCHEV, 2007, p.4)

“Institutions can and, perhaps, should continue to define, set, and measure standards of performance for their students within a competency framework—incorporating into it a value-added component that builds on what is already a quite strong assessment foundation.”(SCHEV, 2007, p.6)

“Terming them “areas of core competency,” [Information Technology Literacy, Written Communication, Quantitative Reasoning, Scientific Reasoning, Critical Thinking, and Oral Communication] the group recommended that institutions conduct regular assessments of these areas, the results of which would be shared with the general public.” (SCHEV, 2007, p.2)

Note. The first federal mandate included in a section of the SACSCOC report that describes the federal mandates institutions also have to assess.

Table 2. *Designs used to measure student “learning” outcomes and the inferences each affords due to validity threats.*

Evidence of Student “Learning”	Research Design	Validity Threats	Inference
Student meets a set performance standard or cutoff score on a measure	<p>One-group posttest-only design. One sample, one time point. Students in the sample are measured only after completing the relevant coursework.</p> <p>Example Sample. A sample of senior math majors complete a department-wide math test in a capstone course.</p>	<p>Internal. None.</p> <p>External. Interaction of testing and treatment</p>	<p>Desired: Students have achieved mastery of a skill after experiencing the curriculum.</p> <p>Example. As a function of completing the math major at Lord University, senior math students are capable of performing matrix algebra.</p> <p>Actual: Students have achieved mastery of a skill. The cause of mastery is unknown. Students could have mastered the skill from experiencing the curriculum, or the students could have mastered the skill prior to college.</p>
The average performance of a group of students that has experienced the institution’s curriculum compared against the average performance of a group that has not experienced the curriculum.	<p>Separate Sample Pretest/Posttest Design. Two samples, one time point. One sample is measured after completing the relevant coursework and the other sample is measured but did not complete the relevant</p>	<p>Internal. History, maturation, mortality, and threat interactions. Possibly instrumentation.</p> <p>External. None.</p>	<p>Desired: Students perform better after experiencing the curriculum.</p> <p>Example 1. Senior math students are better at matrix algebra than the psychology majors because the math students completed the math courses at Lord University.</p>

coursework. The samples can either be two groups of students at the same academic level (e.g., seniors), or one group of upperclassmen and one group of first-year students.

Example Sample(s) 1. A sample of senior math majors and a sample of senior psychology majors complete an institution-wide math test in their respective capstone courses.

Example Sample(s) 2. A sample of senior math majors and a sample of first-year math majors complete a department-wide math test on the first day of the semester.

Example 2. Senior math students are better at matrix algebra than first-year math majors because the senior students completed the math courses at Lord University.

Actual: There is a difference in matrix algebra ability between students who completed the coursework and the students who did not complete the coursework. The cause of the difference in matrix algebra ability is unknown. The difference could be due to the curriculum, student background characteristics, other differences in college experience, etc.

Example 1. Senior math students are better at matrix algebra than the psychology majors.

Example 2. Senior math students are better at matrix algebra than first-year math majors.

<p>Estimating the learning gains of a group of students after they have experienced the curriculum, estimating the learning gains of a group of students who have not experienced the curriculum, and comparing the gain estimates.</p>	<p>Nonequivalent comparison group design. Two samples, two time points. One sample is measured before and after completing the relevant coursework; the other sample is measured at the same times as the first sample. Both samples can be measured as first-year students and again as upperclassmen. If students</p>	<p>Internal. Threat interactions. Possibly regression.</p> <p>External. Interaction of testing and treatment. Possibly interaction effect of selection bias</p>	<p>Desired. Students are learning from the curriculum above and beyond that which can be explained by other effects (e.g., maturation).</p> <p>Example. Students who completed the matrix algebra course have increased in math skills, especially in comparison to students who completed the biology course; this greater increase in math proficiency is due to the assigned coursework.</p>
---	--	---	---

in one sample are randomly assigned to coursework, the researcher can infer that coursework caused the difference in learning gains.

Example Sample(s) 1. At the beginning of their first year, students are randomly assigned to complete either a matrix algebra or introductory biology course. These first-year students complete an institution-wide math test on the first day of the semester and again on the last day of the semester.

Example Sample(s) 2. At the beginning of their first year, students elect to complete either a matrix algebra or introductory biology course. These first-year students complete an institution-wide math test on the first day of the semester and again on the last day of the semester.

and treatment, and reactive arrangements.

Actual. There is a difference in how the two groups of students change over time.

Example 1. Students randomly assigned to the matrix algebra course have increased in math skills, in comparison to students randomly assigned to the biology course. This greater increase in math proficiency is due to the assigned coursework.

Example 2. Students who opted to complete the matrix algebra course have increased in matrix algebra skills, in comparison to students who opted to complete the biology course.

<p>Estimating the learning gains of a group of students after they have experienced the curriculum.</p>	<p>One group pretest/posttest design. One sample, two time points. The sample of students is measured before</p>	<p>Internal. History, maturation, testing, instrumentation,</p>	<p>Desired: Students are learning from the curriculum. Example. Graduating students in the math major are more adept at matrix algebra than they were during their first year;</p>
---	---	--	---

and after completing the relevant coursework. The sample can be measured as first-year students and again as upperclassmen.

Example Sample. A sample of first-year math majors complete a department-wide math test on the first day of the semester. These students complete the math test again on the first day of their senior year.

and threat interactions. Possibly regression.

External. Interaction of testing and treatment, interaction effect of selection bias and treatment. Possibly reactive arrangements.

increase in proficiency is due to their multivariate math coursework.

Actual. Student performance has changed over time. The change could be due to the curriculum, maturation, other college experiences, etc.

Example. Graduating students in the math major are more adept at matrix algebra than they were during their first year.

Table 3. *Description of internal and external validity threats to student learning inferences.*

Internal	External
<p><i>History.</i> Events that occurred before any testing (i.e., pretests or posttests) or before the treatment may influence the experiment's outcome. For example, completing AP Calculus prior to being tested on college math proficiency and completing math courses at college.</p>	<p><i>Interaction effect of testing.</i> A pretest affects how well a participant responds to the treatment. For example, students complete a calculus pretest before completing a calculus course. The pretest, however, reinforced the calculus concepts.</p>
<p><i>Maturation.</i> Participants' aging may influence the experiment's outcome. For example, a college senior having better proficiency in math than he did during his freshman year because his math skills increased as he aged.</p>	<p><i>Interaction effect of selection bias and treatment.</i> Participants in the control group would react differently to the treatment than the treatment group. For example, female students might learn more in a calculus course than male students and thus perform better on a math posttest.</p>
<p><i>Testing.</i> Completing a test affects how the participant completes all subsequent tests. For example, a student completes a math posttest comprised of the same questions as a math pretest that he completed. The student recalls the correct answers from the pretest.</p>	<p><i>Reactive effects of experimental arrangements.</i> Participants try to produce the behavior they believe the experimenters want. For example, students who are asked how much effort they put forth on a math test may indicate that they put a great deal of effort into the test even if they did not.</p>
<p><i>Instrumentation.</i> Changes in the choice of instrument may affect measurement. For example, a student completes a fairly difficult math test before completing college math courses. After the math courses, this student completes a fairly easy math test.</p>	<p><i>Multiple treatment interference.</i> Participants are exposed to multiple treatments, making it difficult to parse out the effects of one treatment from another. For example, a group of students completes a new math course but also receives one-on-one tutoring. The students' performance on a math test cannot be attributed to solely the math course or solely the tutoring.</p>
<p><i>Statistical regression.</i> Selecting participants on the basis of extreme pretest scores, when these scores regress to the mean at the posttest. For example, students who score highly on a math placement pretest are enrolled in an advanced math</p>	

course. All students then complete a posttest after their coursework. However, the posttest scores of the high-scoring students are closer to the posttest scores of their peers than before.

Selection. The control, or comparison, group is comprised of participants who do not resemble the treatment group. For example, the math performance of a group of students who have completed a calculus course is compared to the performance of a group who has not. However, the group that did not complete the course consisted solely of female students, whereas the group that did consist solely of male students.

Experimental mortality. Also known as *attrition*; some participants drop out of the experiment. For example, college seniors have higher average SAT scores than college freshmen because academically struggling students drop out before reaching senior year.

Threat interactions. The threats mentioned above may combine to produce interactive or additive threats. For example, females becoming more adept at math than males (i.e., selection threat example from above) as time progresses (i.e., maturation threat).

Note. Information in this table borrows heavily from Campbell and Stanley (1963) and Shadish, Cook, and Campbell (2002).

Table 4. *Total NW9 data available in each cohort per semester.*

	Cohort One Fall 2007-Spring 2009	Cohort Two Fall 2008-Spring 2010	Cohort Three Fall 2013-Spring 2015	Cohort Four Fall 2014-Spring 2016	Cohort Five Fall 2015-Spring 2017
Fall	1177	1592	1269	384	704
Spring	1113	1174	163	289	576

Note. Counts are only comprised of students with no missing data on the NW9 and who do not have AP/IB or transfer credit.

Table 5. *Cronbach's alpha reliability estimates for the NW9 and SOS-effort subscale.*

	Cohort One	Cohort Two	Cohort Three	Cohort Four	Cohort Five
NW9 Pretest	0.76	0.73	0.73	0.73	0.70
NW9 Posttest	0.79	0.79	0.73	0.79	0.77
SOS Effort Pretest Test-specific	-	0.84	0.71	0.79	0.80
SOS Effort Posttest Test-specific	-	0.80	0.83	0.81	0.79
SOS Effort Pretest Test session-specific	-	0.81	0.83	0.84	0.78
SOS Effort Posttest Test session-specific	-	0.63	0.87	0.84	0.83

Table 6. *Ethnicity, age, gender, and SAT data for students in each unfiltered cohort.*

	Cohort One	Cohort Two	Cohort Three	Cohort Four	Cohort Five
American Indian	0.00%	5.73%	1.25%	1.14%	0.81%
Asian	2.85%	2.51%	1.25%	6.82%	10.48%
Black	2.85%	0.00%	6.25%	7.95%	5.24%
Hispanic	2.64%	1.97%	5.00%	5.11%	5.24%
Not specified	3.25%	9.32%	2.50%	0.57%	3.23%
Pacific Islander	0.20%	0.00%	0.00%	1.70%	0.81%
White	88.01%	79.39%	88.75%	88.07%	85.08%
Age at pretest	18.46	18.43	18.41	18.44	18.46
Age at posttest	19.93	19.92	19.87	19.91	19.91
Female	68.50%	68.46%	70.00 %	64.77%	66.53%
Male	31.30%	31.54%	30.00%	35.23%	33.47%
SAT	1117.39	1126.50	1135.00	1146.81	1136.40
<i>N</i>	492	558	80	176	248

Table 7. Demographic information for students with adequate test-specific motivation.

	Cohort Two	Cohort Three	Cohort Four	Cohort Five
American Indian	0.00%	1.49%	1.22%	0.96%
Asian	5.37%	0.00%	6.71%	9.62%
Black	3.02%	5.97%	7.32%	5.77%
Hispanic	1.01%	4.48%	4.88%	5.29%
Not specified	9.73%	2.99%	0.61%	3.85%
Pacific Islander	1.01%	0.00%	1.83%	0.96%
White	79.87%	91.04%	87.80%	86.06%
Age at pretest	18.43	18.41	18.43	18.47
Age at posttest	19.92	19.87	19.90	19.92
Female	66.11%	70.15%	65.24%	66.83%
Male	33.89%	29.85%	34.76%	33.17%
SAT	1124.91	1135.97	1130.61	1138.38
<i>N</i>	298	67	164	208

Note. Demographics were computed without students who were unmotivated on the test. For Cohorts Two, Three, and Five, students were removed if their test-specific effort scores were below 15. For Cohort Four, students were removed if their test-specific effort scores were below 13.

Table 8. *Ethnicity, age, gender, and SAT data for students in each test session-specific filtered cohort.*

	Cohort Two	Cohort Three	Cohort Four	Cohort Five
American Indian	0.00%	1.49%	1.19%	0.94%
Asian	6.25%	1.49%	6.55%	10.38%
Black	1.56%	7.46%	7.74%	5.19%
Hispanic	0.00%	5.97%	5.36%	5.19%
Not specified	7.81%	2.99%	0.60%	3.30%
Pacific Islander	1.56%	0.00%	1.79%	0.47%
White	82.81%	86.57%	87.50%	85.38%
Age at pretest	18.56	18.40	18.43	18.44
Age at posttest	20.03	19.87	19.90	19.89
Female	67.19%	1134.76	1149.87	33.96%
Male	32.81%	32.84%	35.71	66.04%
SAT	1125.00	67.16%	64.29	1138.00
<i>N</i>	64	67	168	212

Note. Demographics were computed without students who were unmotivated on the test battery. For Cohorts Two, Three, and Five, students were removed if their test session-specific effort scores were below 15. For Cohort Four, students were removed if their test session-specific effort scores were below 12.

Table 9. *Demographic information for students with adequate test session-specific and test-specific motivation.*

	Cohort Three	Cohort Four	Cohort Five
American Indian	1.67%	1.23	1.05
Asian	0.00%	6.79	9.95
Black	6.67%	6.79	5.24
Hispanic	5.00%	4.94	5.76
Not specified	3.33%	0.62	3.66
Pacific Islander	0.00%	1.85	0.52
White	90.00%	87.65	85.86
Age at pretest	18.40	18.43	18.45
Age at posttest	19.86	19.89	19.90
Female	68.33%	64.81%	67.02%
Male	31.67%	35.19%	32.98%
SAT	1135.97	1152.15	1141.41
<i>N</i>	60	162	191

Note. Demographics were computed without students who were unmotivated on the test and test battery. For Cohorts Three and Five, students were removed if their test-specific or test session-specific effort scores were below 15. For Cohort Four, students were removed if their test-specific effort scores were below 13 and if their test session-specific effort scores were below 12.

Table 10. *Number of students removed for low test-taking effort.*

Courses	Cohort Two		Cohort Three			Cohort Four			Cohort Five		
	Test	Session	Test	Session	Both	Test	Session	Both	Test	Session	Both
0	6	1	0	0	0	0	0	0	2	1	1
1	23	6	4	2	1	2	0	0	6	3	6
2	25	4	1	3	2	1	2	2	5	9	3
3	10	2	1	0	3	0	0	3	1	1	5
4	13	1	0	0	0	0	0	0	1	2	2
5	1	0	0	0	0	1	0	0	0	0	0
Overall	103	14	6	5	6	4	2	5	15	16	17

Note. ‘Test’ indicates low motivation on only the test-specific measure. Students in Cohorts Two, Three, and Five were removed if their test-specific effort scores were below 15; students in Cohort Four were removed if their test-specific effort scores were below 13. ‘Session’ indicates low motivation on only the test session-specific measure. Students in Cohorts Two, Three, and Five were removed if their test session-specific effort scores were below 15; students in Cohort Four were removed if their test session-specific effort scores were below 12. ‘Both’ indicates low motivation on the test-specific and test session-specific measures. Students in Cohorts Three and Five were removed if their test-specific and test session-specific effort scores were below 15; students in Cohort Four were removed if their test-specific effort scores were below 13 or test session-specific effort scores were below 12. Students in Cohort 2 did not complete both measures.

Table 11. *Descriptive statistics regarding the unfiltered learning gain estimates.*

Course	0	1	2	3	4	5	6	7	Overall
Cohort 1									
Mean									
Gain Score	5.40	4.66	3.48	5.04	5.72	3.36	2.78		4.45
SD _{gain}	6.41	5.92	5.85	5.92	4.66	6.07	3.42		5.80
Pretest	44.13	43.68	44.29	43.63	45.00	40.18	44.22		43.92
SD _{pretest}	8.98	7.02	6.86	7.63	7.03	3.95	6.48		7.11
Posttest	49.53	48.34	47.77	48.66	50.72	43.55	47.00		48.37
SD _{posttest}	8.06	7.41	6.93	6.65	7.78	5.66	4.18		7.13
Cohen's <i>d</i>									
<i>d</i> _{gain}	0.84	0.79	0.59	0.85	1.23	0.55	0.81		0.77
<i>d</i> _{pretest}	0.60	0.66	0.51	0.66	0.81	0.85	0.43		0.62
N	15	157	147	107	46	11	9		492
Cohort 2									
Mean									
Gain Score	1.70	3.81	3.85	3.27	4.34	4.92	2.73	2.00	3.67
SD _{gain}	5.70	5.49	6.12	5.56	4.68	2.81	3.00		5.55
Pretest	44.80	43.94	44.50	46.66	46.02	44.00	41.09	40.00	44.83
SD _{pretest}	5.35	6.43	6.95	6.47	6.97	4.51	5.20		6.62
Posttest	46.50	47.75	48.35	49.93	50.36	48.92	43.82	42.00	48.50
SD _{posttest}	7.98	6.70	7.53	6.95	6.29	4.86	5.19		7.06
Cohen's <i>d</i>									
<i>d</i> _{gain}	0.30	0.69	0.63	0.59	0.93	1.75	0.91		0.66
<i>d</i> _{pretest}	0.32	0.59	0.55	0.51	0.62	1.09	0.52		0.55
N	30	164	175	100	64	13	11	1	558
Cohort 3									
Mean									
Gain Score	0.50	1.24	-0.14	2.15	4.00		5.00		1.43
SD _{gain}	0.71	4.66	6.16	4.96	4.81		1.41		5.15

Pretest	51.50	49.28	47.09	44.85	46.25	49.00	47.00	47.26
SD _{pretest}	4.95	6.71	5.13	6.31	8.24		5.66	6.35
Posttest	52.00	50.52	46.95	47.00	50.25	49.00	52.00	48.69
SD _{posttest}	4.24	5.80	5.55	7.02	7.11		4.24	6.21
Cohen's <i>d</i>								
<i>d</i> _{gain}	0.71	0.27	-0.02	0.43	0.83		3.54	0.28
<i>d</i> _{pretest}	0.10	0.18	-0.03	0.34	0.49		0.88	0.22
N	2	25	22	20	8	1	2	80
Cohort 4								
Mean								
Gain Score	0.83	3.06	3.22	3.39	3.10	7.20		3.23
SD _{gain}	5.12	6.32	5.23	4.71	7.03	6.30		5.59
Pretest	41.83	46.75	47.24	43.00	45.70	44.00		46.07
SD _{pretest}	8.70	5.62	6.71	6.19	6.52	5.20		6.51
Posttest	42.67	49.81	50.46	46.39	48.80	51.20		49.30
SD _{posttest}	10.71	6.29	6.25	7.29	9.47	5.54		6.97
Cohen's <i>d</i>								
<i>d</i> _{gain}	0.16	0.48	0.61	0.72	0.44	1.14		0.58
<i>d</i> _{pretest}	0.10	0.55	0.48	0.55	0.48	1.39		0.50
N	6	48	79	28	10	5		176
Cohort 5								
Mean								
Gain Score	3.22	3.61	4.07	3.04	2.29	3.00		3.47
SD _{gain}	5.17	5.90	4.70	5.66	5.08	3.92		5.29
Pretest	46.33	45.52	46.15	45.33	45.71	42.75		45.70
SD _{pretest}	4.12	6.89	5.52	6.67	5.72	5.32		6.12
Posttest	49.56	49.12	50.22	48.36	48.00	45.75		49.17
SD _{posttest}	6.15	7.71	5.54	6.88	6.67	4.99		6.62
Cohen's <i>d</i>								
<i>d</i> _{gain}	0.62	0.61	0.87	0.54	0.45	0.77		0.66
<i>d</i> _{pretest}	0.78	0.52	0.74	0.46	0.40	0.56		0.57

N	9	66	86	55	28	4			248
Overall									
Mean									
Gain Score	2.69	3.85	3.51	3.78	4.28	4.38	2.95	2.00	3.72
SD _{gain}	5.58	5.73	5.66	5.58	4.90	4.43	3.03	0.00	5.57
Pretest	44.79	44.66	45.26	44.93	45.65	42.76	42.91	40.00	44.95
SD _{pretest}	6.36	6.63	6.57	6.87	6.80	4.39	5.77	0.00	6.67
Posttest	47.48	48.51	48.76	48.71	49.94	47.15	45.86	42.00	48.66
SD _{posttest}	7.88	6.99	6.74	6.87	7.04	5.09	4.69	0.00	6.96
Cohen's <i>d</i>									
<i>d</i> _{gain}	0.48	0.67	0.62	0.68	0.87	0.99	0.98		0.67
<i>d</i> _{pretest}	0.42	0.58	0.53	0.55	0.63	1.00	0.51		0.56
N	62	460	509	310	156	34	22	1	1554

Note. 'SD' indicates standard deviation. 'Gain Score' indicates the difference between the posttest and pretest scores. '*d*_{gain}' indicates that Cohen's *d* estimates were computed using the standard deviation of the difference scores; '*d*_{pretest}' indicates that Cohen's *d* estimates were computed using the standard deviation of the pretest scores. 'N' indicates the number of students in the cohort or sample. 'Overall' indicates that the values were computed collapsing across all the cohorts. Students could score at most 66 points on the NW9.

Table 12. Descriptive statistics regarding the test session-specific filtered learning gain estimates.

Course	0	1	2	3	4	5	6	7	Overall
Cohort 2									
Mean									
Gain Score	-1.40	4.19	5.42	4.53	1.29		1.50		3.80
SD _{gain}	4.28	4.35	4.46	7.10	3.04		2.12		5.24
Pretest	45.80	41.19	43.47	45.40	47.29		41.50		43.89
SD _{pretest}	3.83	6.82	7.62	6.54	5.06		4.95		6.72
Posttest	44.40	45.38	48.89	49.93	48.57		43.00		47.69
SD _{posttest}	5.98	7.33	6.86	6.89	4.54		2.83		6.77
Cohen's <i>d</i>									
<i>d</i> _{gain}	-0.33	0.96	1.21	0.64	0.42		0.71		0.72
<i>d</i> _{pretest}	-0.37	0.61	0.71	0.69	0.25		0.30		0.57
N	5	16	19	15	7		2		64
Cohort 3									
Mean									
Gain Score	0.50	1.10	1.12	3.50	4.00	0.00	5.00		2.10
SD _{gain}	0.71	4.77	5.89	4.08	4.81		1.41		4.85
Pretest	51.50	50.10	46.88	44.31	46.25	49.00	47.00		47.37
SD _{pretest}	4.95	6.81	5.40	6.30	8.24		5.66		6.60
Posttest	52.00	51.19	48.00	47.81	50.25	49.00	52.00		49.48
SD _{posttest}	4.24	5.97	5.43	6.70	7.11		4.24		6.06
Cohen's <i>d</i>									
<i>d</i> _{gain}	0.71	0.23	0.19	0.86	0.83		3.54		0.43
<i>d</i> _{pretest}	0.10	0.16	0.21	0.56	0.49		0.88		0.32
N	2	21	17	16	8	1	2		67
Cohort 4									
Mean									
Gain Score	0.83	3.06	3.13	3.08	3.10	7.20			3.14
SD _{gain}	5.12	6.32	5.08	4.89	7.03	6.30			5.57
Pretest	41.83	46.75	47.52	43.67	45.70	44.00			46.33
SD _{pretest}	8.70	5.62	6.58	6.42	6.52	5.20			6.45
Posttest	42.67	49.81	50.65	46.75	48.80	51.20			49.48
SD _{posttest}	10.71	6.29	6.16	7.75	9.47	5.54			6.98
Cohen's <i>d</i>									
<i>d</i> _{gain}	0.16	0.48	0.62	0.63	0.44	1.14			0.56
<i>d</i> _{pretest}	0.10	0.55	0.48	0.48	0.48	1.39			0.49
N	6	48	75	24	10	5			168
Cohort 5									
Mean									
Gain Score	3.57	4.02	4.01	3.69	2.75	3.00			3.76
SD _{gain}	5.88	5.58	4.84	5.59	5.02	3.92			5.21
Pretest	47.29	45.91	46.59	45.29	46.04	42.75			46.01

	SD _{pretest}	4.23	6.58	5.58	6.88	5.77	5.32		6.12
	Posttest	50.86	49.93	50.61	48.98	48.79	45.75		49.77
	SD _{posttest}	6.41	7.27	5.60	6.85	6.21	4.99		6.44
Cohen's <i>d</i>									
	<i>d</i> _{gain}	0.61	0.72	0.83	0.66	0.55	0.77		0.72
	<i>d</i> _{pretest}	0.84	0.61	0.72	0.54	0.48	0.56		0.62
	N	7	55	74	48	24	4		212
Overall									
Mean									
	Gain Score	1.20	3.27	3.54	3.64	2.82	4.80	3.25	3.35
	SD _{gain}	4.73	5.57	4.99	5.41	5.11	4.72	1.77	5.28
	Pretest	45.70	46.29	46.68	44.78	46.18	44.00	44.25	46.03
	SD _{pretest}	5.55	6.31	6.18	6.63	6.23	4.72	5.30	6.37
	Posttest	46.90	49.56	50.21	48.42	49.00	48.80	47.50	49.38
	SD _{posttest}	7.38	6.75	5.94	7.04	6.79	4.77	3.54	6.61
Cohen's <i>d</i>									
	<i>d</i> _{gain}	0.25	0.59	0.72	0.68	0.55	0.88	2.12	0.63
	<i>d</i> _{pretest}	0.24	0.52	0.57	0.55	0.45	0.92	0.59	0.53
	N	20	140	185	103	49	10	4 0	511

Note. 'SD' indicates standard deviation. 'Gain Score' indicates the difference between the posttest and pretest scores. '*d*_{gain}' indicates that Cohen's *d* estimates were computed using the standard deviation of the difference scores; '*d*_{pretest}' indicates that Cohen's *d* estimates were computed using the standard deviation of the pretest scores. 'N' indicates the number of students in the cohort or sample. 'Overall' indicates that the values were computed collapsing across all the cohorts. Students could score at most 66 points on the NW9.

Table 13. Descriptive statistics regarding the test-specific filtered learning gain estimates.

Course	0	1	2	3	4	5	6	7	Overall
Cohort 2									
Mean									
Gain Score	1.00	3.60	4.61	2.63	5.09	3.20	2.00		3.84
SD _{gain}	6.56	5.10	5.70	4.97	4.28	3.03	2.76		5.22
Pretest	42.86	45.45	45.01	47.19	45.51	43.20	42.50		45.46
SD _{pretest}	8.59	5.93	6.80	6.43	7.01	5.22	6.28		6.52
Posttest	43.86	49.04	49.62	49.83	50.60	46.40	44.50		49.30
SD _{posttest}	12.09	5.95	6.41	7.10	5.70	3.36	6.35		6.51
Cohen's <i>d</i>									
d _{gain}	0.15	0.71	0.81	0.53	1.19	1.06	0.73		0.74
d _{pretest}	0.12	0.61	0.68	0.41	0.73	0.61	0.32		0.59
N	7	94	99	52	35	5	6		298
Cohort 3									
Mean									
Gain Score	0.50	1.05	0.74	4.07	4.00	0.00	5.00		2.07
SD _{gain}	0.71	5.01	6.01	3.65	4.81		1.41		5.00
Pretest	51.50	50.20	46.53	44.73	46.25	49.00	47.00		47.39
SD _{pretest}	4.95	6.26	5.16	6.65	8.24		5.66		6.40
Posttest	52.00	51.25	47.26	48.80	50.25	49.00	52.00		49.46
SD _{posttest}	4.24	5.31	5.67	6.96	7.11		4.24		6.00
Cohen's <i>d</i>									
d _{gain}	0.71	0.21	0.12	1.11	0.83		3.54		0.42
d _{pretest}	0.10	0.17	0.14	0.61	0.49		0.88		0.32
N	2	20	19	15	8	1	2		67
Cohort 4									
Mean									
Gain Score	0.83	2.85	3.03	3.08	3.10	9.50			3.07
SD _{gain}	5.12	6.37	5.04	4.89	7.03	4.20			5.57
Pretest	41.83	46.91	47.28	43.67	45.70	43.00			46.25
SD _{pretest}	8.70	5.67	6.68	6.42	6.52	5.42			6.52
Posttest	42.67	49.76	50.31	46.75	48.80	52.50			49.32
SD _{posttest}	10.71	6.41	6.33	7.75	9.47	5.45			7.08
Cohen's <i>d</i>									
d _{gain}	0.16	0.45	0.60	0.63	0.44	2.26			0.55
d _{pretest}	0.10	0.50	0.45	0.48	0.48	1.75			0.47
N	6	46	74	24	10	4			164
Cohort 5									
Mean									
Gain Score	5.33	3.58	4.05	3.70	2.80	3.00			3.72

SD _{gain}	4.50	5.76	4.58	5.74	4.88	3.92		5.16
Pretest	46.50	46.54	46.55	45.23	46.52	42.75		46.17
SD _{pretest}	4.04	6.05	5.55	6.33	5.49	5.32		5.80
Posttest	51.83	50.12	50.61	48.94	49.32	45.75		49.89
SD _{posttest}	6.24	6.65	5.16	6.23	5.66	4.99		5.90
Cohen's <i>d</i>								
d _{gain}	1.18	0.62	0.88	0.65	0.57	0.77		0.72
d _{pretest}	1.32	0.59	0.73	0.59	0.51	0.56		0.64
N	6	52	74	47	25	4		208
Overall								
Mean								
Gain Score	2.14	3.19	3.74	3.23	3.99	4.71	2.75	3.47
SD _{gain}	5.37	5.56	5.32	5.09	4.95	4.58	2.76	5.28
Pretest	44.43	46.48	46.18	45.64	45.94	43.43	43.63	46.01
SD _{pretest}	7.45	6.05	6.37	6.48	6.51	4.93	6.09	6.33
Posttest	46.57	49.67	49.92	48.88	49.92	48.14	46.38	49.48
SD _{posttest}	10.08	6.17	6.04	6.93	6.32	4.99	6.59	6.43
Cohen's <i>d</i>								
d _{gain}	0.40	0.57	0.70	0.63	0.81	1.03	0.99	0.66
d _{pretest}	0.29	0.53	0.59	0.50	0.61	0.96	0.45	0.55
N	21	212	266	138	78	14	8	737

Note. 'SD' indicates standard deviation. 'Gain Score' indicates the difference between the posttest and pretest scores. '*d*_{gain}' indicates that Cohen's *d* estimates were computed using the standard deviation of the difference scores; '*d*_{pretest}' indicates that Cohen's *d* estimates were computed using the standard deviation of the pretest scores. 'N' indicates the number of students in the cohort or sample. 'Overall' indicates that the values were computed collapsing across all the cohorts. Students could score at most 66 points on the NW9.

Table 14. Descriptive statistics regarding the test session-specific and test-specific filtered learning gain estimates.

Course	0	1	2	3	4	5	6	Overall
Cohort 3								
Mean								
Gain Score	0.50	0.82	1.25	4.14	4.00	0.00	5.00	2.25
SD _{gain}	0.71	5.07	6.06	3.78	4.81		1.41	4.98
Pretest	51.50	50.94	46.81	44.21	46.25	49.00	47.00	47.50
SD _{pretest}	4.95	6.27	5.56	6.58	8.24		5.66	6.63
Posttest	52.00	51.76	48.06	48.36	50.25	49.00	52.00	49.75
SD _{posttest}	4.24	5.55	5.60	7.00	7.11		4.24	6.04
Cohen's <i>d</i>								
<i>d</i> _{gain}	0.71	0.16	0.21	1.10	0.83		3.54	0.45
<i>d</i> _{pretest}	0.10	0.13	0.22	0.63	0.49		0.88	0.34
N	2	17	16	14	8	1	2	60
Cohort 4								
Mean								
Gain Score	0.83	2.85	3.04	3.08	3.10	9.50		3.07
SD _{gain}	5.12	6.37	5.11	4.89	7.03	4.20		5.60
Pretest	41.83	46.91	47.50	43.67	45.70	43.00		46.33
SD _{pretest}	8.70	5.67	6.55	6.42	6.52	5.42		6.47
Posttest	42.67	49.76	50.54	46.75	48.80	52.50		49.41
SD _{posttest}	10.71	6.41	6.18	7.75	9.47	5.45		7.05
Cohen's <i>d</i>								
<i>d</i> _{gain}	0.16	0.45	0.60	0.63	0.44	2.26		0.55
<i>d</i> _{pretest}	0.10	0.50	0.46	0.48	0.48	1.75		0.47
N	6	46	72	24	10	4		162
Cohort 5								
Mean								
Gain Score	6.20	4.04	4.11	4.09	2.91	3.00		3.97
SD _{gain}	4.44	5.55	4.70	5.53	5.06	3.92		5.12
Pretest	47.20	46.35	46.94	45.11	46.35	42.75		46.21
SD _{pretest}	4.09	6.15	5.53	6.31	5.70	5.32		5.86
Posttest	53.40	50.40	51.05	49.20	49.26	45.75		50.18
SD _{posttest}	5.50	6.79	5.06	6.17	5.90	4.99		5.94
Cohen's <i>d</i>								
<i>d</i> _{gain}	1.40	0.73	0.87	0.74	0.58	0.77		0.78
<i>d</i> _{pretest}	1.52	0.66	0.74	0.65	0.51	0.56		0.68
N	5	48	66	45	23	4		191
Overall								
Mean								
Gain Score	2.85	3.05	3.31	3.81	3.17	5.56	5.00	3.37

SD_{gain}	4.18	5.81	5.03	5.05	5.49	3.61	1.41	5.29
Pretest	45.38	47.29	47.19	44.54	46.17	43.56	47.00	46.45
SD_{pretest}	6.35	5.97	6.01	6.39	6.40	4.77	5.66	6.21
Posttest	48.23	50.34	50.50	48.35	49.34	49.11	52.00	49.82
SD_{posttest}	7.71	6.44	5.64	6.77	7.01	4.64	4.24	6.39
Cohen's d								
d_{gain}	0.68	0.53	0.66	0.75	0.58	1.54	3.54	0.64
d_{pretest}	0.45	0.51	0.55	0.60	0.50	1.16	0.88	0.54
N	13	111	154	83	41	9	2	413

Note. 'SD' indicates standard deviation. 'Gain Score' indicates the difference between the posttest and pretest scores. ' d_{gain} ' indicates that Cohen's d estimates were computed using the standard deviation of the difference scores; ' d_{pretest} ' indicates that Cohen's d estimates were computed using the standard deviation of the pretest scores. 'N' indicates the number of students in the cohort or sample. 'Overall' indicates that the values were computed collapsing across all the cohorts. Students could score at most 66 points on the NW9.

Table 15. Comparison of unfiltered and filtered estimates collapsing across cohorts 3-5.

Courses	0	1	2	3	4	5	6	Overall
Unfiltered								
Gain Score	2.06	2.99	3.21	2.96	2.76	4.80	5.00	3.06
d_{gain}	0.43	0.51	0.61	0.56	0.51	0.88	3.54	0.57
d_{pretest}	0.31	0.46	0.54	0.46	0.44	0.96	0.88	0.49
N	17	139	187	103	56	10	2	514
Test-specific								
Gain Score	2.71	2.86	3.20	3.68	3.09	5.56	5.00	3.23
d_{gain}	0.56	0.48	0.64	0.71	0.58	1.06	3.54	0.61
d_{pretest}	0.39	0.47	0.53	0.58	0.50	1.09	0.88	0.52
N	14	116	168	84	43	9	2	436
Test session-specific								
Gain Score	2.71	2.91	3.20	3.72	3.09	4.80	5.00	3.24
d_{gain}	0.56	0.49	0.64	0.72	0.58	0.88	3.54	0.61
d_{pretest}	0.39	0.48	0.53	0.58	0.50	0.96	0.88	0.52
N	14	117	168	85	43	10	2	439

Note. ' d_{gain} ' indicates Cohen's d estimates, and that these estimates were computed using the standard deviation of the difference scores. ' d_{pretest} ' indicates Cohen's d estimates, and that these estimates were computed using the standard deviation of the pretest scores. 'N' indicates the number of students in the cohort or sample. 'Overall' indicates that the values were computed collapsing across all the courses. Students could score at most 66 points on the NW9.

Table 16. Correlations among gain scores and potential predictors in the unfiltered and test-specific-filtered samples.

	Course		Gender		SAT		GenderxCourse		GenderxSAT		CoursexSAT	
	UF	F	UF	F	UF	F	UF	F	UF	F	UF	F
Gain Score	.03	.04	.01	.05	-.03	-.08*	.04	.08*	-.02	-.07*	-.03	-.09*
Course			.10*	.11*	-.06	-.11*	.69*	.66*	-.05	-.09*	-.09*	-.13*
Gender					-.21*	-.23*	.80*	.82*	-.10*	-.10*	-.20*	-.12*
SAT							-.17*	-.22*	.82*	.84*	.88*	.88*
GenderxCourse									-.09*	-.13*	-.19*	-.25*
GenderxSAT											.74*	.74*

Note. 'x' denotes interaction between the predictors. * indicates significance at $p < 0.05$ 'UF' denotes correlation computed in the unfiltered sample. 'F' denotes correlation computed in the filtered sample. Filtered correlations have been corrected for low test-specific motivation.

Table 17. Regression results in both the unfiltered and test-specific filtered samples.

	<i>F</i>	<i>df</i>	<i>p</i>	<i>R</i> ²	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	95.0% CI for <i>b</i>		<i>sr</i>
									LB	UB	
Unfiltered Models											
Reduced	0.61	(3,997)	0.61	0.002							
Intercept					3.03	0.42	7.22	<0.001	2.21	3.85	
Prior ability					-0.001	0.001	0.66	0.51	-0.004	0.002	-.02
Gender					0.18	0.38	0.48	0.63	-0.57	0.93	.02
Coursework					0.13	0.14	0.90	0.37	-0.15	0.41	.03
Full	0.47	(6,994)	0.83	0.003							
Intercept					3.39	0.64	5.25	<0.001	2.12	4.65	
Prior ability					0.001	0.004	0.19	0.85	-0.01	0.01	.01
Gender					-0.31	0.78	0.39	0.69	-1.83	1.22	-.01
Course					-0.06	0.29	0.21	0.84	-0.62	0.50	-.01
Gender x Course					0.25	0.33	0.73	0.46	-0.41	0.90	.02
Gender x Prior ability					<0.001	0.003	0.05	0.95	-0.01	0.01	-.002
Course x Prior ability					-0.001	0.001	0.56	0.57	-0.003	0.002	-.02
Filtered Models											
Reduced	2.15	(3, 685)	0.09	0.01							
Prior ability					2.93	0.50	5.88	<0.001	1.95	3.91	
Gender					-0.003	0.002	-1.78	0.08	0.01	0.00	-.07
Coursework					0.43	0.44	0.98	0.33	-0.44	1.30	.04
Intercept					0.14	0.17	0.81	0.42	-0.20	0.48	.03
Full	1.69	(6, 682)	0.12	0.02							
Intercept					3.91	0.77	5.07	<0.001	2.39	5.43	
Prior ability					-0.001	0.004	-0.24	0.81	-0.01	0.01	-.01
Gender					-0.97	0.94	-1.03	0.30	-2.81	0.87	-.04
Course					-0.38	0.34	-1.11	0.27	-1.04	0.29	-.04
Gender x Course					0.70	0.40	1.75	0.08	-0.09	1.48	.07
Gender x Prior ability					-0.001	0.004	-0.31	0.76	-0.01	0.01	-.01
Course x Prior ability					-0.001	0.001	-0.37	0.71	-0.003	0.002	-.01

Note. ‘x’ denotes interaction between variables. ‘LB’ denotes the lower bound of the confidence interval; ‘UB’ denotes the upper bound of the confidence interval. ‘sr’ denotes the semipartial correlation.

Table 18. *Empirical learning gain estimates filtered for low test-specific motivation compared to faculty-based estimates, and the alignment of expected estimates and desired estimates.*

	Actual	Faculty One			Faculty Two			Faculty Three			Faculty Four		
		Expect	Desire	Aligned	Expect	Desire	Aligned	Expect	Desire	Aligned	Expect	Desire	Aligned
0 courses	2.14	2	2		3	21		4	4		2-3	?	
1 course	3.19	4	5		4	21		7	7		3-5	5	
2 courses	3.74	6	9	Not aligned	5	21	Not aligned	10	10	Aligned	5-7	7	Aligned
3 courses	3.23	7	14		5	21		15	15		7-10	10	
Overall	3.47	4	5		4	21		4	4		-	-	

Note. Gain scores refer to the point-gain on the NW9 for each number of quantitative and scientific reasoning courses. For example, students who did not complete any quantitative and scientific reasoning courses, on average, gained 2.14 points on the 66-item test (after controlling for low test-specific motivation) and students who completed three quantitative and scientific reasoning courses, on average, gained 3.23 points on the 66-item test (after controlling for low test-specific motivation). ‘Overall’ indicates that average learning gain collapsing across number of courses completed (i.e., after 1.5 years of any college coursework). ‘Aligned’ refers to the alignment between faculty’s expected and desired gain scores (i.e., whether or not the expected estimates matched the desired estimates). Faculty Four did not provide written estimates for students with zero courses because he did not have an opinion on how much these students should gain. Faculty Four also did not provide estimates collapsing across courses (i.e., overall). As he explained, it was difficult to produce these estimates without knowing how much relevant coursework students had completed.

Table 19. Themes derived from faculty interviews.

Faculty One	Faculty Two	Faculty Three	Faculty Four
Themes regarding alignment between expectations and desires			
Students will demonstrate learning gain in college, but learning gain is mostly facilitated by domain-specific coursework.	Students do not have high learning gains, but should learn with increased coursework.	Difficult to estimate learning gains	Expectations framed through student familiarity
Unrealized high desires for student learning gains	High standards for student non-cognitive attributes	Belief that expectations are reasonable	Students will demonstrate learning gain in college, but learning gain is mostly facilitated by domain-specific coursework.
Expecting low gains but desiring high gains	Unrealized high desires for student learning gains	Students should learn from general and domain-specific courses	Desire for students to learn from quantitative and scientific reasoning coursework
Students in different courses will have different learning gains	Expecting low gains but desiring high gains		

Note. Bolded themes indicate themes that were shared across faculty.

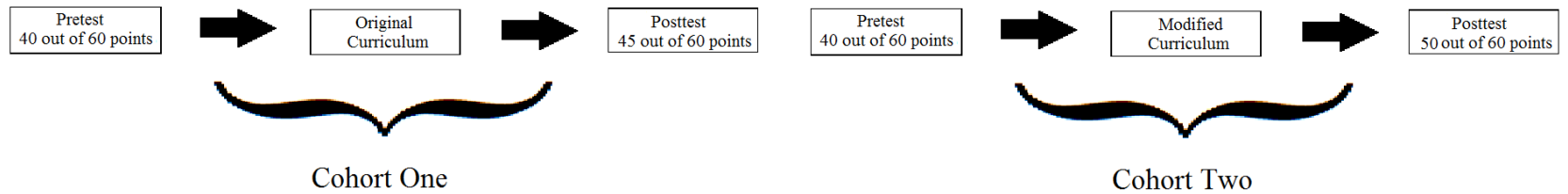


Figure 1. Illustration of learning gain versus learning improvement. As can be seen, Cohort One gains on average five points more on the pretest than on the posttest after completing the original curriculum. Thus, Cohort One has a *learning gain* of five points. Cohort Two gains on average ten points after completing the new, modified curriculum. Thus, Cohort Two has a *learning gain* of ten points. However, Cohort Two gained five points more after completing the modified curriculum than Cohort One gained after completing the original curriculum (i.e., ten versus five points). The positive difference between the gain scores is an indication of *learning improvement*.

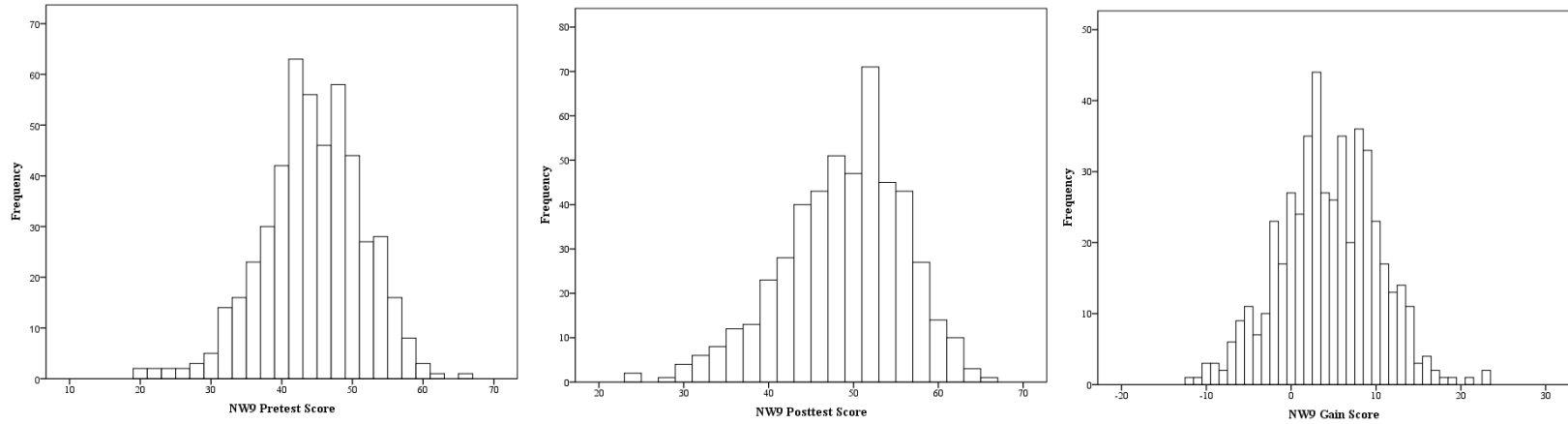


Figure 2. Cohort One unfiltered pretest, posttest, and difference scores (respectively). Bar widths represent intervals of the pretest, posttest, or gain scores. Smaller widths indicate smaller intervals. These latter two statements also apply to Figures 3-14).

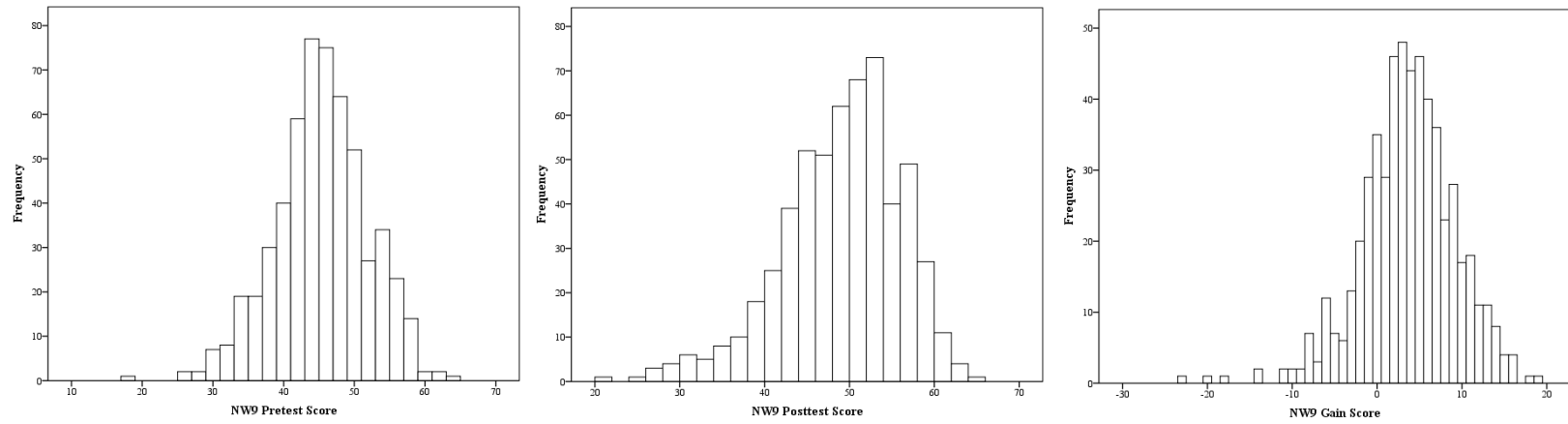


Figure 3. Cohort Two unfiltered pretest, posttest, and difference scores (respectively).

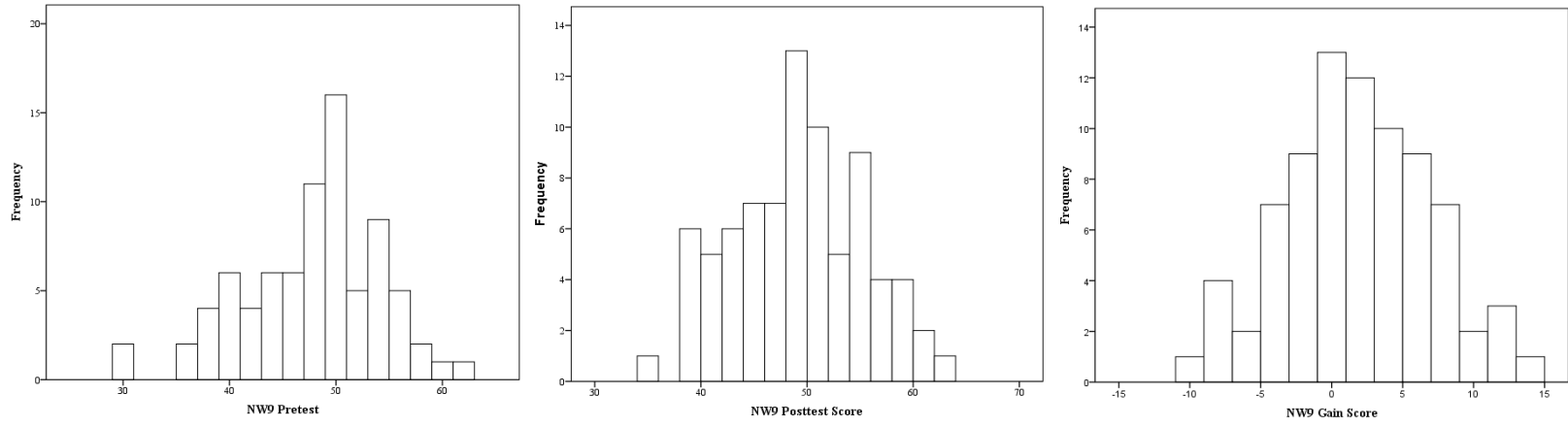


Figure 4. Cohort Three unfiltered pretest, posttest, and difference scores (respectively).

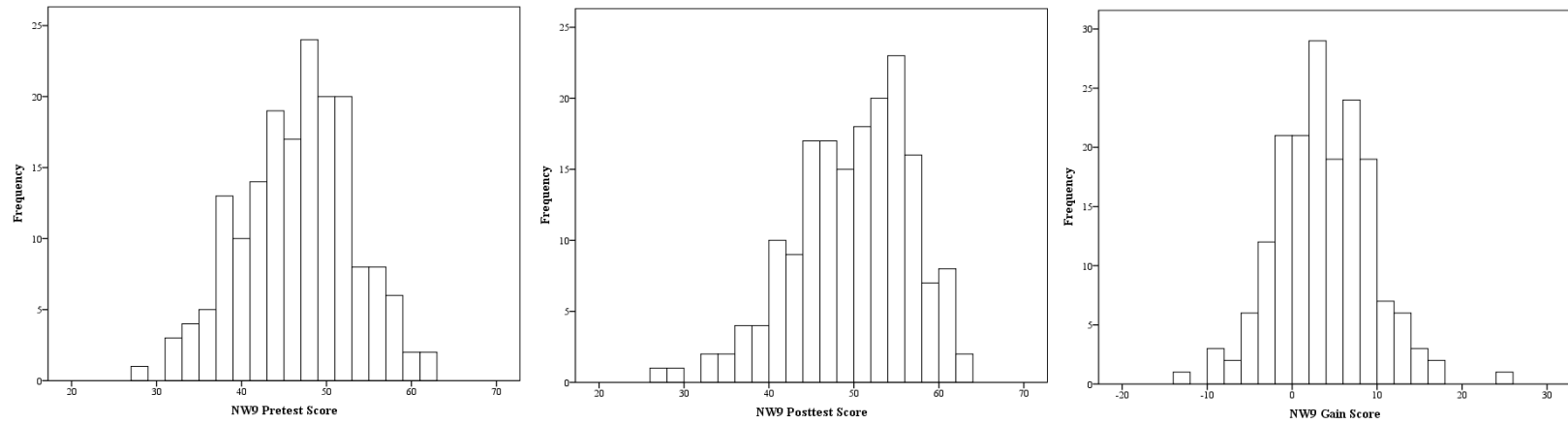


Figure 5. Cohort Four unfiltered pretest, posttest, and difference scores (respectively).

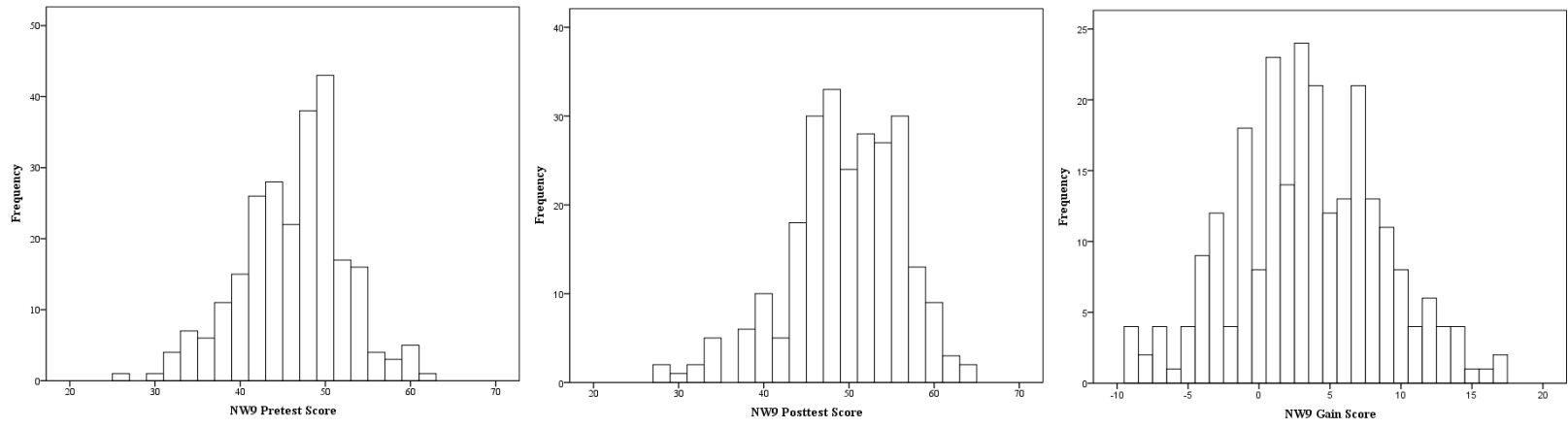


Figure 6. Cohort Five unfiltered pretest, posttest, and difference scores (respectively).

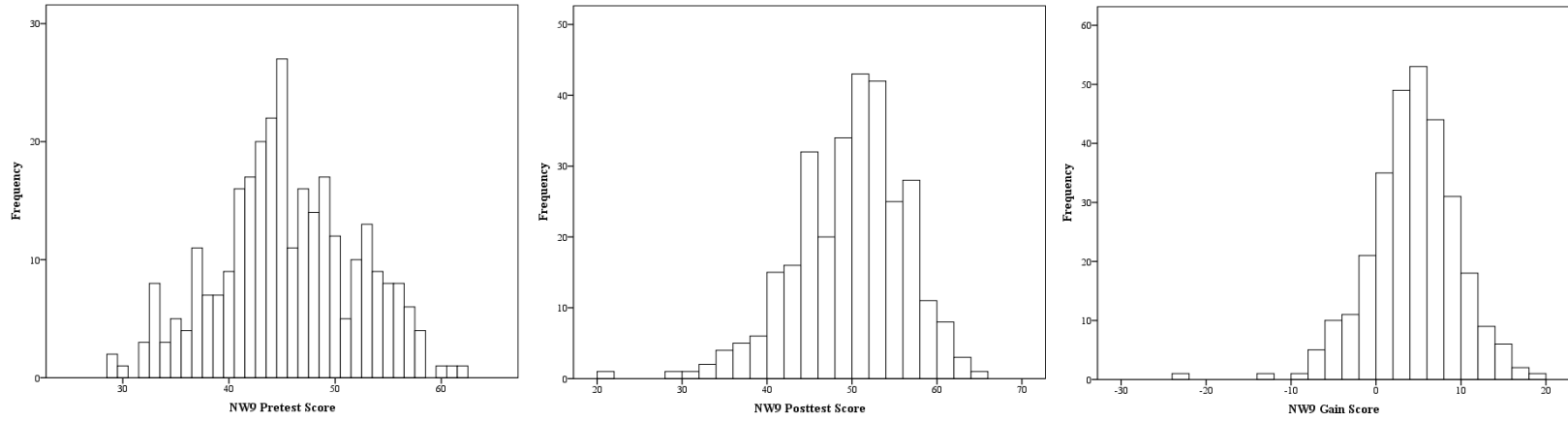


Figure 7. Pretest, posttest, and difference scores (respectively) filtered for low test-specific effort in Cohort Two.

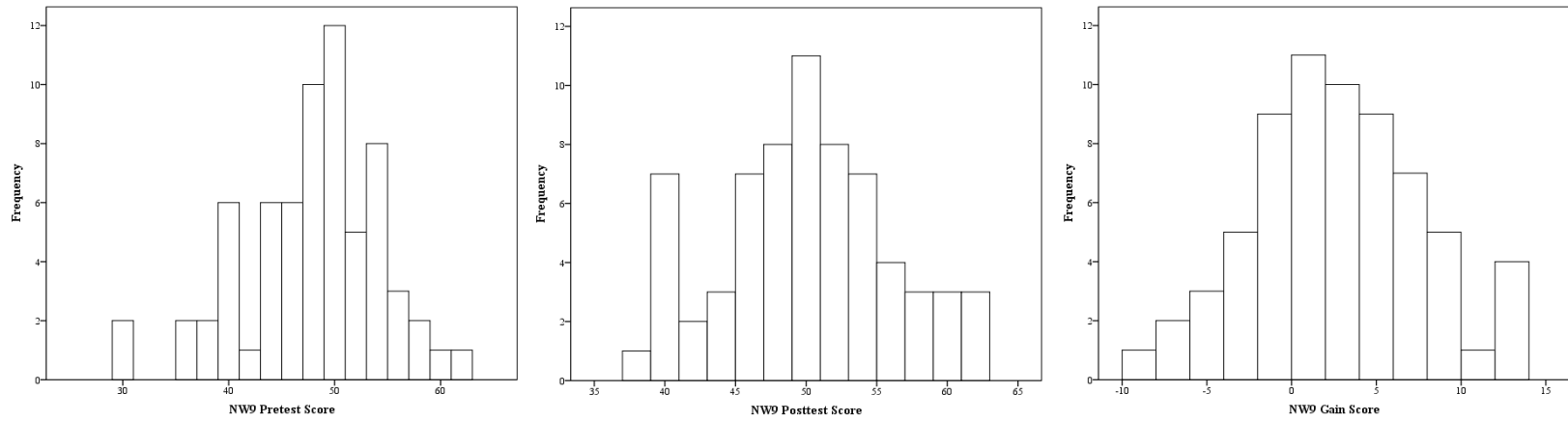


Figure 8. Pretest, posttest, and difference scores (respectively) filtered for low test-specific effort in Cohort Three.

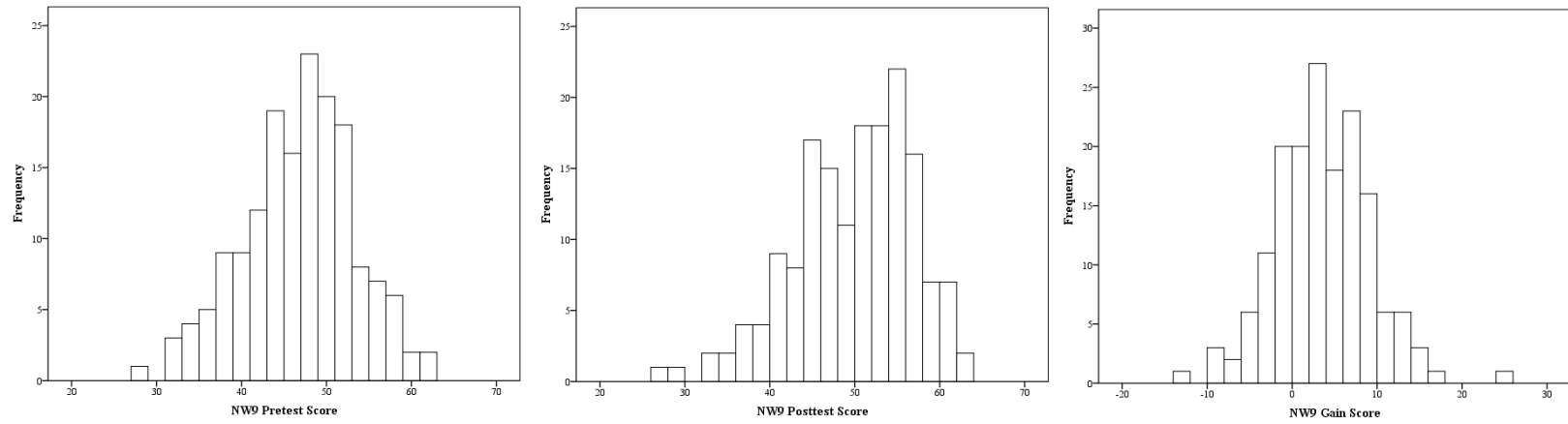


Figure 9. Pretest, posttest, and difference scores (respectively) filtered for low test-specific effort in Cohort Four.

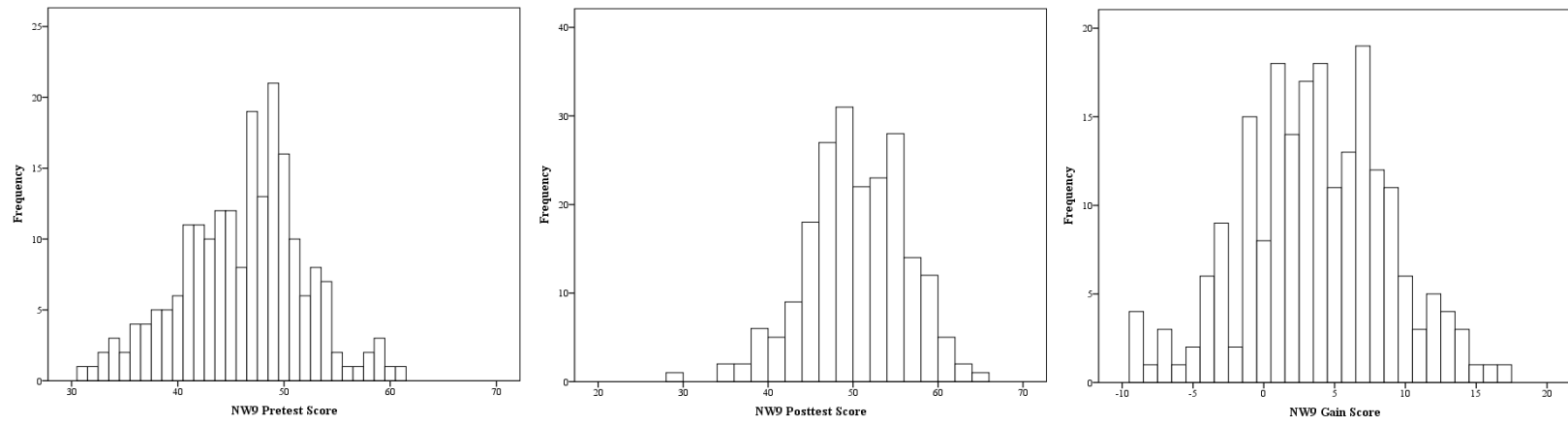


Figure 10. Pretest, posttest, and difference scores (respectively) filtered for low test-specific effort in Cohort Five.

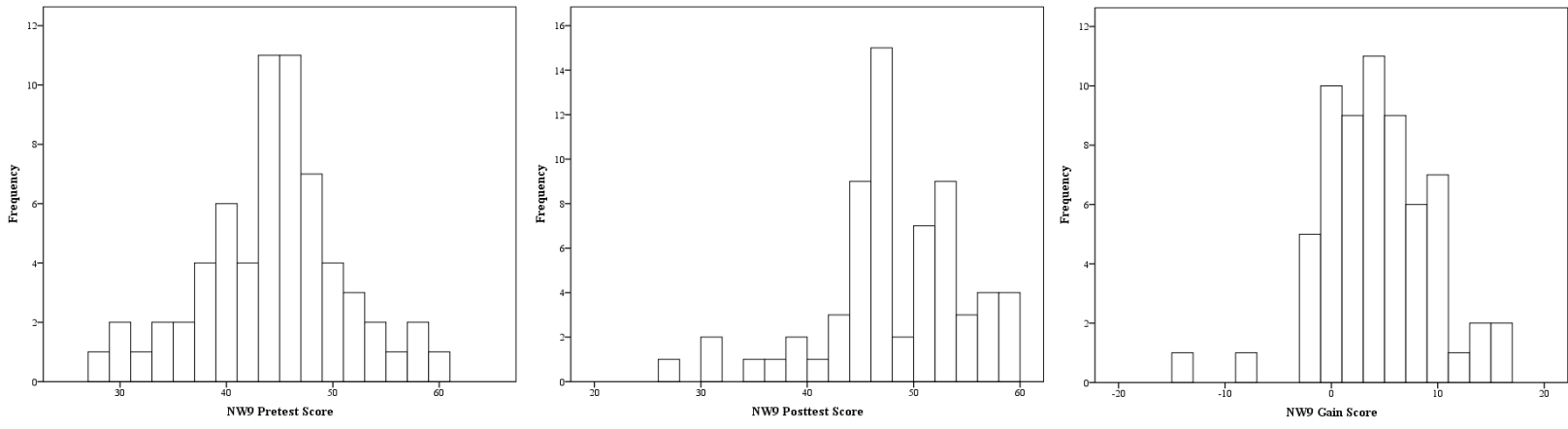


Figure 11. Pretest, posttest, and difference scores (respectively) filtered for low test session-specific effort in Cohort Two.

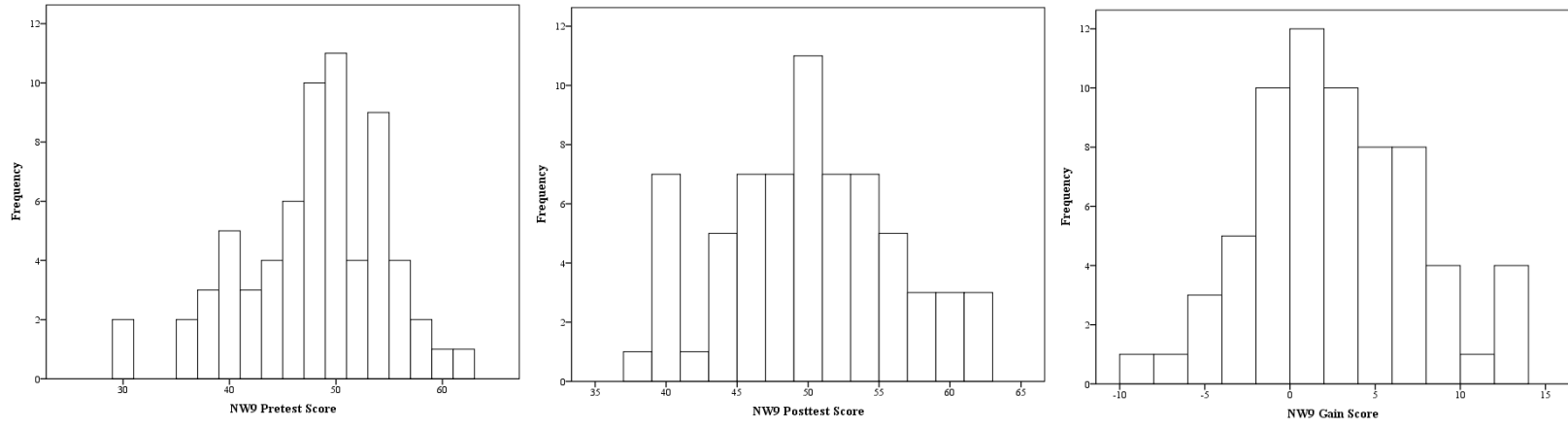


Figure 12. Pretest, posttest, and difference scores (respectively) filtered for low test-session specific effort in Cohort Three.

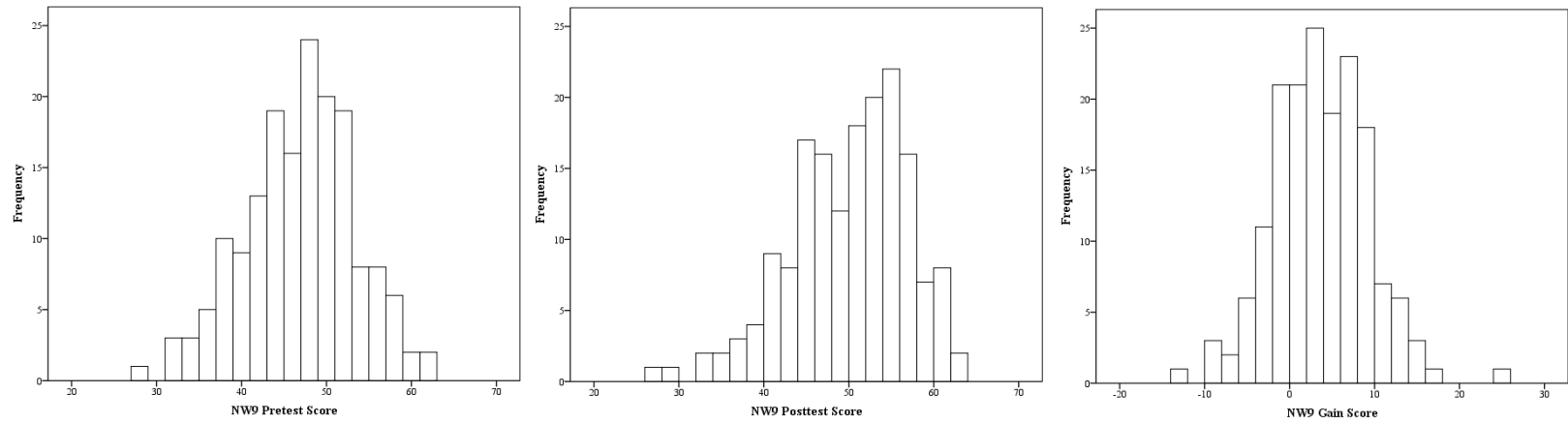


Figure 13. Pretest, posttest, and difference scores (respectively) filtered for low test-session specific effort in Cohort Four.

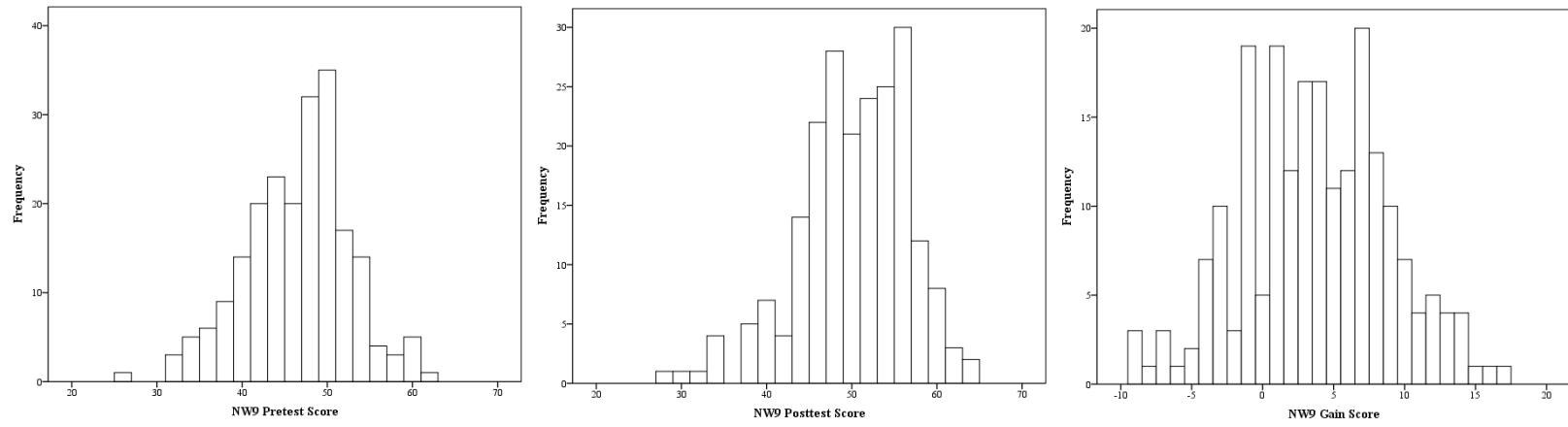


Figure 14. Pretest, posttest, and difference scores (respectively) filtered for low test-session specific effort in Cohort Five.

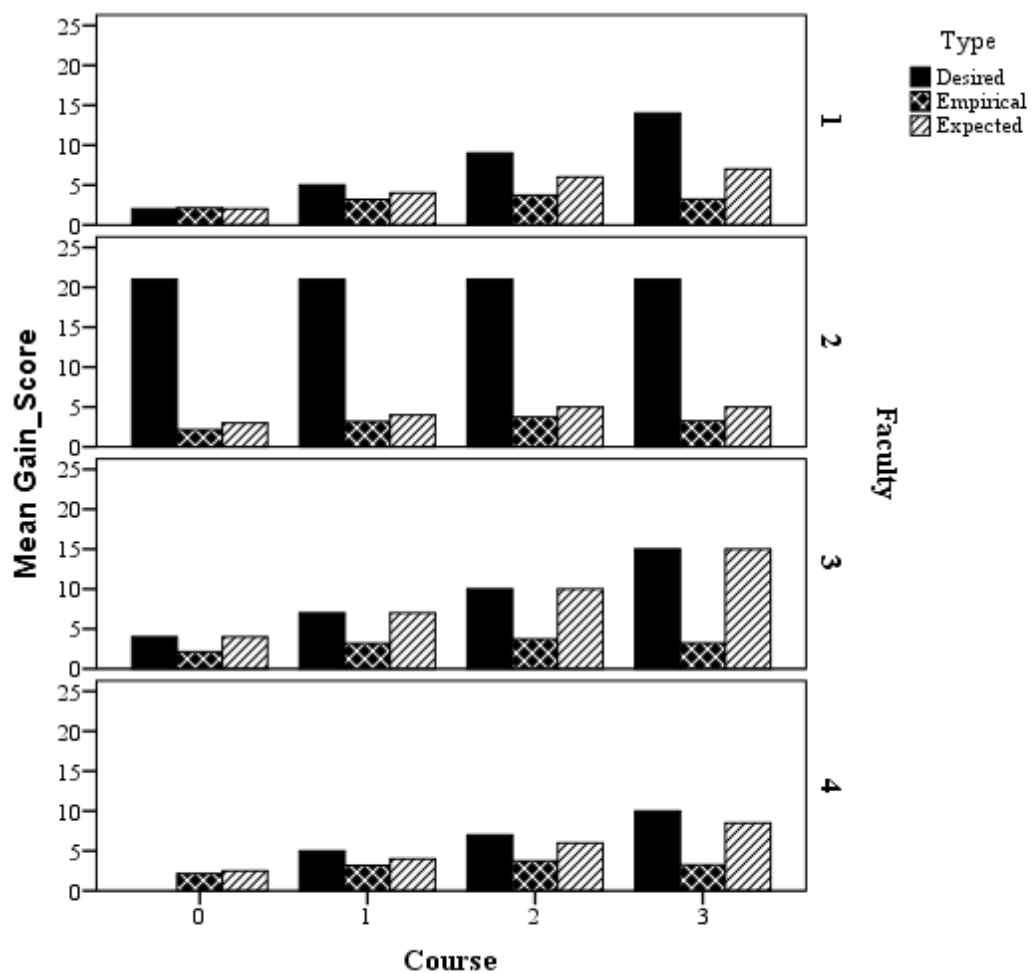


Figure 15. Empirical gain scores (filtered for low test-specific motivation) compared to the expected and desired gain scores of quantitative and scientific reasoning faculty. Estimated gain scores are located on the left y-axis; corresponding faculty member is located on the right y-axis. Number of completed courses are on the x-axis. Faculty Four did not provide a desired estimate for students who did not complete any courses. The empirical gain score is shown once in each faculty quadrant.

Appendix A

Test-Session Specific SOS

Please think about all the tests that you completed today. Mark the answer that best represents how you feel about each of the statements below.

1. Doing well on these tests was important to me.
2. I engaged in good effort throughout these tests.*
3. I am not curious about how I did on these tests relative to others.
4. I am not concerned about the scores I receive on these tests.
5. These were important tests to me.
6. I gave my best effort on these tests.*
7. While taking these tests, I could have worked harder on them.*
8. I would like to know how well I did on these tests.
9. I did not give these tests my full attention while completing them.*
10. While taking these tests, I was able to persist to completion of the tasks. *

Test-Specific SOS

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

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. *

*= item on the 'effort' subscale

Appendix B

Consent Form

Consent to Participate in Research

Identification of Investigators & Purpose of Study

You are being asked to participate in a research study conducted by Catherine Mathers and Dr. Sara Finney from James Madison University. The purpose of this study is to understand faculty expectations of student learning gains, and whether these expectations align with empirical student learning gains. This study will contribute to the researcher's completion of her Master's thesis.

Research Procedures

Should you decide to participate in this research study, you will be asked to sign this consent form once all your questions have been answered to your satisfaction. This study consists of an interview that will be administered to individual participants in Lakeview Hall. You will be asked to provide answers to a series of questions related to your opinions of student learning gains in math and science.

Time Required

Participation in this study will require 45 minutes of your time.

Risks

Breach of confidentiality is a minor risk. However, your anonymity will be preserved. The investigator does not perceive more than minimal other risks from your involvement in this study (that is, no risks beyond the risks associated with everyday life).

Benefits

Potential benefits from participation in this study include additional perspective on student math and science learning gains, information on how much students learn with each Cluster 3 course completed, and the opportunity to participate in a relatively new area of research. This study will benefit the research area by contributing to the nonexistent literature on faculty opinions of student learning gains. Furthermore, this study has the potential benefits of highlighting the strengths of the Cluster 3 curriculum or improving the learning gains of students who complete Cluster 3 courses at JMU.

Confidentiality

The results of this research will be presented at conferences. Identifying data (e.g., name, department) will not be collected. However, your verbal and written communications may be quoted to support qualitative analyses. The researcher retains the right to use and publish non-identifiable data. While individual responses are confidential, aggregate data will be presented representing averages or generalizations about the responses as a whole. All data will be stored in a secure location accessible only to the researcher and her advisor.

Participation & Withdrawal

Your participation is entirely voluntary. You are free to choose not to participate. Should you choose to participate, you can withdraw at any time without consequences of any kind.

Questions about the Study

If you have questions or concerns during the time of your participation in this study, or after its completion or you would like to receive a copy of the final aggregate results of this study, please contact:

Catherine E. Mathers

Graduate Psychology

James Madison University

matherce@dukes.jmu.edu

finneysj@jmu.edu

Sara J. Finney

Graduate Psychology

James Madison University

Telephone: 540-568-6757

Questions about Your Rights as a Research Subject

Dr. David Cockley

Chair, Institutional Review Board

James Madison University

(540) 568-2834

cocklede@jmu.edu

Giving of Consent

I have read this consent form and I understand what is being requested of me as a participant in this study. I freely consent to participate. I have been given satisfactory answers to my questions. The investigator provided me with a copy of this form. I certify that I am at least 18 years of age.

I give consent to have my verbal communication quoted in the researcher's Master's thesis and any subsequent scholarly articles.
_____ (initials)

I give consent to have my written communication quoted in the researcher's Master's thesis and any subsequent scholarly articles.
_____ (initials)

Name of Participant (Printed)

Name of Participant (Signed) _____
Date

Name of Researcher (Signed) _____
Date

Appendix C

Form A

Recall, students tend to score about 45 out of 66 points on the NW9 at the beginning of their first-year at JMU.

1. How many additional points do you **expect** students who have not completed any quantitative and scientific reasoning courses from Cluster 3 to gain on the NW9?
2. How many points do you **expect** students who have completed 1 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
3. How many points do you **expect** students who have completed 2 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
4. How many points do you **expect** students who have completed 3 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
5. How many points **would you like** students who have not completed quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
6. How many points **would you like** students who have completed 1 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
7. How many points **would you like** students who have completed 2 quantitative and scientific reasoning courses from Cluster 3 to gain on the NW9?
8. How many points **would you like** students who have completed 3 quantitative and scientific reasoning courses from Cluster 3 to gain on the NW9?

Recall, over their first 1.5 years of college, students can complete from 0 to 3 Cluster 3 courses.

9. How many points do you **expect** students who have completed 1.5 years of college coursework to gain on the NW9?

10. How many points **would you like** students who have completed 1.5 years of college coursework to gain on the NW9?
11. Please *explain why* your **expected** learning gain estimates match or do not match your **desired** learning gain estimates for each of the above questions.

Appendix D

Interview Guide

Part A.

Faculty participants are given the consent form. The researcher verbally explains the study and allows the interviewee to read and sign the consent form.

Part B.

After the consent form has been collected, the researcher provides background on the NW9 with respect to test development, average pretest scores, item difficulty, and test reliability. The researcher gives the interviewee the form shown in Appendix C; these questions are also listed below for easy reference. She then explains that she would like the interviewee to write down how much he/she expects students to gain on the NW9 and how much he/she would like students to gain on the NW9, taking into consideration the information just provided on the NW9 and his/her own familiarity with the Cluster 3 curriculum. She will also ask the interviewee to indicate when he/she has finished completing the form. After the interviewee has written his/her estimates, the researcher will ask the interviewee to verbally explain his/her estimates, and that at this time she will take notes to record the interviewee's response.

1. How many additional points do you expect students who have not completed any quantitative and scientific reasoning courses from Cluster 3 to gain on the NW9?
2. How many points do you expect students who have completed 1 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
3. How many points do you expect students who have completed 2 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
4. How many points do you expect students who have completed 3 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
5. How many points would you like students who have not completed quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
6. How many points would you like students who have completed 1 quantitative and scientific reasoning course from Cluster 3 to gain on the NW9?
7. How many points would you like students who have completed 2 quantitative and scientific reasoning courses from Cluster 3 to gain on the NW9?
8. How many points would you like students who have completed 3 quantitative and scientific reasoning courses from Cluster 3 to gain on the NW9?
9. How many points do you expect students who have completed 1.5 years of college coursework to gain on the NW9?
10. How many points would you like students who have completed 1.5 years of college coursework to gain on the NW9?
11. Please explain why your expected learning gain estimates match or do not match your desired learning gain estimates for each of the above questions.

If faculty say that it is too difficult to estimate the gains, or that they cannot estimate the gain, the researcher will ask the following questions:

1. What makes it difficult for you to estimate the gain?
2. What would you need to know in order to estimate the gain?

Part C.

After the faculty participant has completed the form in Appendix C, the researcher will conduct the debriefing session. The researcher will first collect the forms from the interviewee and thank him/her for participation. The researcher will allow for questions or comments. Afterward, the researchers will hand the interviewee a form that says the following:

“Thank you for participating in the study. As you know, the purpose of this study is to understand how much faculty expect and want students to learn from their coursework. There has been very little research to date on the subject. However, previous research on student learning gains has called for improved student learning in higher education. Student learning cannot be improved, unfortunately, if faculty do not have an understanding of how much their students are learning. What’s more, there may not be a need to improve student learning if students are learning as much as their professors want. Your participation in this study will help to clarify this area and is greatly appreciated. If you have any questions or concerns, or would like to request results of this study when they are available, please contact Dr. Finney or myself.”

Thank you,
Catie

Catherine E. Mathers
matherce@dukes.jmu.edu

Sara J. Finney
Telephone: 540-568-6757
finneysj@jmu.edu

References

- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.
- ACT & College Board (2009). *ACT-SAT Concordance Tables*. Retrieved from <http://www.act.org/content/dam/act/unsecured/documents/ACTCollegeBoardJointStatement.pdf>
- Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, Illinois: University of Chicago Press.
- Association of American Colleges & Universities (2004). *General education and the assessment reform agenda*. Washington, D.C.: Ewell, P.T.
- Astin, A. W. (1998). The changing American college student: Thirty-year trends, 1966-1996. *The Review of Higher Education*, 21, 115-135.
- Astin, A.W., Banta, T.W., Cross, K.P., El-Khawas, E., Ewell, P.T., Hutchings, P....Wright, B.D.(1996). 9 principles of good practice for student learning. *AAHE Bulletin*, 45, 1-3.
- Astin, A.W. & Lee, J.J. (2003). How risky are one-shot cross-sectional assessments of undergraduate students? *Research in Higher Education*, 44, 657-672.
- Atwell, R., Breneman, D.W., Edwards, V.B., Olson, L., Finn, C.E., Levine, A....Miller, C. (2006). *Reactions to the Spellings Commission Report*. Retrieved from <http://www.highereducation.org/crosstalk/ct0406/Spellings.pdf>.

- Auwarter, A.E., & Aruguete, M.S. (2008). Effects of student gender and socioeconomic status on teacher perceptions. *The Journal of Educational Research, 101*, 242-246.
- Babcock, P., & Marks, M. (2010). *Leisure college, USA*. Department of Economics, UCSB.
- Bandalos, D. (2016). Ways of assessing reliability. In Measurement theory and applications for the social sciences. Manuscript in preparation.
- Banta, T.W., & Blaich, C. (2010). Closing the assessment loop. *Change: The Magazine of Higher learning, 43*, 22-27.
- Banta, T. W., Jones, E. A., & Black, K. E. (2009). *Designing effective assessment: Principles and profiles of good practice*. San Francisco, California: John Wiley & Sons.
- Barry, C. & Finney, S.J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*, 46-64.
- Blaich, C., & Wise, K. (2011). *The Wabash National Study: The impact of teaching practices and institutional conditions on student growth*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans: Louisiana.
- Bray, G.B., Pascarella, E.T., & Pierson, C.T. (2004). Postsecondary education and some dimensions of literacy development: An exploration of longitudinal evidence. *Reading Research Quarterly, 39*, 306-330.
- Buchmann, C., DiPrete, T.A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology, 34*, 19-37.

- Campbell, D.T. & Stanley J.C. (1963). *Experimental and quasi-experimental designs for research*. Boston, Massachusetts: Houghton Mifflin Company.
- Castellano, K.E., & Ho, A.D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers.
- Council for Higher Education Accreditation (2002). *The fundamentals of accreditation*. Washington, D.C.
- Chickering, A. W. (1999). Personal qualities and human development in higher education: Assessment in the service of educational goals. *Assessment in higher education: Issues of access, quality, student development, and public policy*, 13-33.
- College Entrance Examination Board and Educational Testing Service (1999). *Correspondences between ACT and SAT I Scores (No. 99-1)*. New York, New York. Dorans, N.J.
- Cohen, J.(1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.
- Cohen (1992). A power primer. *Psychological Bulletin*, 112, 155.
- Cole, J.S., Bergin, D.A, & Whittaker, T.A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, 33, 609-624.
- Creswell, J.W. & Plano Clark, V.L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, California: SAGE Publications.
- Curtis, N. (2016). *Natural World Cluster 3 Assessment Report: Spring 2016*. Harrisonburg, VA

- Darby, A., & Newman, G. (2014). Exploring faculty members' motivation and persistence in academic service-learning pedagogy. *Journal of Higher Education Outreach and Engagement, 18*, 1-119.
- Dawson, T.E. (1997). *A primer on experimental and quasi-experimental design*. Paper presented at the annual meeting of the Southwest Educational Research Association. Austin: Texas.
- DeMars, C.E., Sundre, D.L., & Wise, S.L. (2002). Standard setting: A systematic approach to interpreting student learning. *The Journal of General Education, 51*, 1-20.
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). *Why people fail to recognize their own incompetence. Current directions in psychological science, 12*, 83-87.
- Dunst, C.J., & Hamby, D.W. (2012). Guide for calculating and interpreting effect sizes and confidence intervals in intellectual and developmental disability research studies. *Journal of Intellectual & Developmental Disability, 37*, 89-99.
- Eaton, J. S. (2009). Accreditation in the United States. *New Directions for Higher Education, 2009*, 79–86.
- Eklof, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement, 66*, 643-656.
- Eklof, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311-326.
- Eklof, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*, 345-356.

- Erwin, T.D., & DeFilippo, J.G. (2010). The evolution of assessment policy: A view from Virginia. *Change: The Magazine of Higher Learning*, 42, 40-44.
- Ewell, P.T. (2009). *Assessment, accountability and improvement: Revisiting the tension*. National Institute for Learning Outcomes Assessment.
- Falconer-Medlin, S.M. (2014). *Teacher and student perspectives of factors affecting the school performance of African American students at Suburbia High School* (Unpublished masters thesis). California State University at Sacramento, California.
- Finney, S.J., Mathers, C.E., & Myers, A.J. (2016). Investigating the dimensionality of examinee motivation across instruction conditions in low-stakes testing contexts. *Research & Practice in Assessment*, 11.
- Finney, S.J., Sundre, D.L., Swain, M.S., & Williams, L.M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, 21, 60-87.
- Fram, E.H. & Pearse, R. (2001). "Tough love" teaching generates student hostility. *College Teaching*, 48.
- 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 National Institute for Learning Outcomes Assessment*.
- Gaston, P. L. (2013). *Higher education accreditation: How it's changing, why it must*. Sterling, Virginia: Stylus Publishing, LLC.

- Gerstner, J.J. & Finney, S.J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research & Practice in Assessment, 8*, 15-28.
- Gong, B. (2004). *Models for using student growth measures in school accountability* [PowerPoint slides]. Retrieved from <http://nciea.org/publications/GongGrowthModels111504.pdf>
- Gottfried, A.E., Marcoulides, G.A., Gottfried, A.W., Oliver, P.H., & Guerin, D.W. (2007). Multivariate latent change modeling of developmental decline in academic intrinsic math motivation and achievement: Childhood through adolescence. *International Journal of Behavioral Development, 31*, 317-327.
- Gravetter, F.J. & Wallnau, L.B. (2009). *Statistics for the behavioral sciences*. Belmont, California: Wadworth, Cengage Learning
- Grigorenko, E.L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E.J., & Sternberg, R.J. (2009). Are SSATS and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101*, 964-981.
- Hagedorn, L.S., Siadat, M.V., Nora, A., & Pascarella, E.T. (1996). *Factors leading to gains in mathematics during the first year of college: An analysis by gender and ethnicity*. Paper presented at the annual meeting of the American Educational Research Association. New York: New York.

Haladya, T.M. & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing.

Educational Measurement: Issues and Practice, 23, 17-21.

Halpern, D.E. & Hakel, M.D. (2003) Applying the science of learning to the university and

beyond: Teaching for long-term retention and transfer. *Change: The Magazine*, 35, 36-41.

Harris, D.M. (2012). Varying teacher expectations and standards: Curriculum differentiation in

the age of standards-based reform. *Educational and Urban Society*, 44, 128-150.

Hathcoat, J.D., Sanders, C.B., & Miesen, C. (2015). *Motivation filtering: Comparison of a test-*

specific and global measure of student effort in low-stakes testing. Unpublished manuscript, James Madison University, Harrisonburg, Virginia.

Hathcoat, J.D., Sundre, D.L., & Johnston, M.M. (2015). Assessing college students' quantitative

and scientific reasoning: The James Madison University story. *Numeracy*, 18, 1-19.

HCM Strategists LLC (2012). *Using student learning as a measure of quality in higher*

education. Porter, S. R.

Herzog, S. (2011). Gauging academic growth of bachelor degree recipients: Longitudinal vs.

self-reported gains in general education. *New Directions for Institutional Research*, 150, 21-39.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content

analysis. *Qualitative health research*, 15, 1277-1288.

- Hulleman, C.S., Kosovich, J.J., Barron, K.E., & Daniel, D.B. (in press). Making connections: Replicating and extending the Utility Value Intervention in the classroom. *Journal of Educational Psychology*.
- Klein, S. (2010). *The Lumina Longitudinal Study: Summary of procedures and findings*. Unpublished manuscript.
- Klein, S., Benjamin, R., Shavelson, R. and Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*, 31, 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 Assessment*, 33, 662-673
- Knight, W.E. (1993). An examination of freshmen to senior general education gains across a national sample of institutions with different general education requirements using a mixed-effect structural equation model. *Research in Higher Education*, 34, 41-54.
- Kuh G.D. & Ikenberry, S.O. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. National Institute for Learning Outcomes Assessment.
- Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, Ewell, P.T., T. R., Hutchings, P., & Kinzie, J. (2014). *Using evidence of student learning to improve higher education*. San Francisco, California: Jossey-Bass.

- Kuh, G.D., Jankowski, N., Ikenberry, S.O., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of student learning in U.S. colleges and universities*. National Institute for Learning Outcomes Assessment.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1-12.
- Liu, O.L. (2011). Value-added assessment in higher education: *A comparison of two methods*. *Higher Education, 61*, 445-461.
- Lumina Foundation (2011). *The Degree Qualifications Profile*. Indianapolis, Indiana.
- Mathers, C.E., Finney, S.J., & Myers, A.J. (2016). “*While taking this test, I felt my heart beating fast.*” *The unintended effects of manipulating test instructions to increase examinee motivation in low-stakes testing*. Unpublished manuscript, James Madison University, Harrisonburg, Virginia.
- Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. San Francisco, California: John Wiley & Sons.
- Middaugh, M.F. (2010). *Planning and assessment in higher education: Demonstrating institutional effectiveness*. San Francisco, California: Jossey-Bass.
- Miller, M.A. & Ewell, P.T. (2005). *Measuring Up on college-level learning*. National Center Report #05-8. The National Center for Public Policy and Higher Education.
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological methods, 7*, 105.

- Myers, A.J., Finney, S.J., & Mathers, C.E. (2016). *A moderated mediation model of test importance, examinee effort, and test performance across test instruction conditions*. Unpublished manuscript. James Madison University, Harrisonburg, VA.
- National Center for Higher Education Management Systems (1983). *Information on student outcomes: How to get it and how to use it*. An NCHEMS Executive Overview. Boulder, Colorado: Ewell, P.T.
- National Center for Higher Education Management Systems (1985). *Transformation leadership for improving student outcomes*. NCHEMS Monograph 6. Boulder, Colorado: Ewell, P.T.
- National Center for Higher Education Management Systems (1987). *Assessment, accountability and improvement: Managing the contradiction*. Boulder, Colorado: Ewell, P.T.
- Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Pascarella, E. T., & Blaich, C. (2013). Lessons from the Wabash National Study of Liberal Arts Education. *Change: The Magazine of Higher Learning*, 45, 6-15.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students (Vol. 2)*. K. A. Feldman (Ed.). San Francisco, CA: Jossey-Bass.
- Pastor, D. A., Kaliski, P. K., & Weiss, B. A. (2007). Examining College Students' Gains in General Education. *Research & Practice in Assessment*, 2.

- Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*, 1-17.
- Penk, C., & Richter, D. (2016). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability, 1-25*.
- Pieper, S., Fulcher, K., Sundre, D. L., & Erwin, T. D. (2008). What do I do with the data now? Analyzing assessment information for accountability and improvement. *Research and Practice in Assessment, 2*.
- Pike, G.R. (1992). Lies, damn lies, and statistics revisited: A comparison of three methods of representing change. *Research in Higher Education, 33*, 71-84.
- Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 161*, 69-82.
- Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. *The Analysis of Change, 3*, 66.
- Roohr, K. C., Liu, H., & Liu, O. L. (2016). Investigating student learning gains in college: A longitudinal study. *Studies in Higher Education, 1-17*.

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, 76, 429-444.

Ryoo, J.H., Molfese, V.J., Heaton, R., Zhou, X., Brown, E.T., Prokasky, A., & Davis, E. (2014). Early mathematics skills from prekindergarten to first grade: Score changes and ability group differences in Kentucky, Nebraska, and Shanghai samples. *Journal of Advanced Academics*, 25, 162-188.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. *Educational measurement*, 4, 307-353.

Seifert, T.A., Pascarella, E.T., Colangelo, N., & Assouline, S.G. (2007). The effect of honors program participation on experiences of good practices and learning outcomes. *Journal of College Student Development*, 48, 57-74.

Seifert, T. A., Pascarella, E. T., Erkel, S. I., & Goodman, K. M. (2010). The importance of longitudinal pretest-posttest designs in estimating college impact. *New Directions for Institutional Research*, 2, 5-16.

Shadish, W.R., Campbell, T.D., & Cook, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, Massachusetts: Houghton Mifflin Company.

- Smith, K.L., Good, M.R., Sanchez, E.H., & Fulcher, K.H. (2015). Communication is key: Unpacking “Use of assessment results to improve student learning.” *Research & Practice in Assessment, 10*.
- Southern Association of Colleges and Schools Commission on Colleges (2012). *The principles of accreditation: Foundations for quality enhancement*. Decatur, GA: Southern Association of Colleges and Schools Commission on Colleges.
- State Council of Higher Education for Virginia (2007). *Guidelines for assessment of student learning*. Richmond, Virginia.
- State Council of Higher Education for Virginia (2013). *Virginia Public Higher Education Policy on Program Productivity*. Richmond, Virginia.
- Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* Unpublished manuscript, James Madison University. Harrisonburg, VA
- 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*, 6-26.
- Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*.

Sundre, D. L., Thelk, A., & Wigtil, C. (2008). *The Natural World Test, version 9: A measure of quantitative and scientific reasoning, test manual*. Unpublished manuscript, James Madison University, Harrisonburg, Virginia.

Sundre, D. L., & Wise, S. L. (2003). *'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Unpublished manuscript, James Madison University, Harrisonburg, Virginia.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *The Journal of General Education*, 167-195.

Taasoobshirazi, G., & Sinatra, G. M. (2011). A structural equation model of conceptual change in physics. *Journal of Research in Science Teaching*, 48, 901-918.

Terenzini, P. T., Springer, L., Pascarella, E. T., & Nora, A. (1995). Academic and out-of-class influences on students' intellectual orientations. *The Review of Higher Education*, 19, 23.

Terenzini, P. T., & Wright, T. M. (1987). Influences on students' academic growth during four years of college. *Research in higher education*, 26, 161-179.

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, 129-151.

The National Center for Public Policy and Higher Education (2006). *Measuring Up 2006: The national report card on higher education*. San Jose, California.

- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.
- Timmermans, A.C., de Boer, H., & van der Wer, M.P.C. (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. *Social Psychology of Education, 19*, 217-240.
- Toutkoushian, R. K., & Smart, J. C. (2001). Do institutional characteristics affect student gains from college? *The Review of Higher Education, 25*, 39-61.
- Spellings, M. (2006). *A test of leadership: Charting the future of US higher education*. US Department of Education. Washington, D.C.
- Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & health sciences, 15*, 398-405.
- Wholuba, B.H. (2014). *Examination of the motivation for learning of gifted and non-gifted students as it relates to academic performance* (Unpublished doctoral dissertation). Florida State University, Florida.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.
- Wigfield, A. & Eccles, J. (2002). *Development of achievement motivation*. San Diego, California: Academic Press.

- Williams, L.M. (2016). *The effect of examinee motivation on value-added estimates (Unpublished doctoral dissertation)*. James Madison University, Virginia.
- Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*, 59-69.
- Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*, 237-252.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.
- Wise, S. L., & DeMars, C. E. (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*, 163-183.
- Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada.
- Wise, S.L. & Smith, L.F. (2011). *A model of examinee test-taking effort*. In J.A. Bovaird, K.F., Geisinger, & C.W., Buckendahl, (Eds.), *High-stakes testing in education: Science and practice in K–12 settings*, 139-153. Washington, DC, US: American Psychological Association.

- Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*, 65-83.
- Witte, R.S. (1993) *Statistics (4th ed.)*. Orlando, Florida: Harcourt, Brace, Jovanovich.
- Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.
- Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*, 360-384.
- Zumbo, B. D., Wu, A. D., & Liu, Y. (2012). Measurement and statistical analysis issues with longitudinal assessment data. *Improving Large-Scale Assessment in Education: Theory, Issues, and Practice, 276*.
- Zvoch, K. & Stevens, J.J. (2006). Successive student cohorts and longitudinal growth models: An investigation of elementary school mathematics performance. *Education Policy Analysis Archives, 14*, 1-25.