2014-11-01

Self-Reported Mastery: Moving on from Self-Reported Gains in Assessing Learning Outcomes

Michael S. Thompson
Brigham Young University - Provo

Self-Reported Mastery: Moving on from Self-Reported Gains in Assessing Learning Outcomes

Michael S. Thompson

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Russell T. Osguthorpe, Chair
Trav D. Johnson
Richard R Sudweeks
Stephen Yanchar
Andrew S. Gibbons

Department of Instructional Psychology and Technology

Brigham Young University

November 2014

ABSTRACT

Self-Reported Mastery: Moving on from Self-Reported Gains in Assessing Learning Outcomes

Michael S. Thompson
Department of Instructional Psychology and Technology, BYU
Doctor of Philosophy

As the learning outcomes movement gains strength, the need to effectively measure learning outcomes becomes more important. This study looked at the effectiveness of self-reported mastery in measuring learning outcomes by examining the correlations between (a) self-reported mastery, (b) self-reported gains, and (c) objective measures of learning outcomes. The objective measures of learning outcomes were final exams for two classes, Calculus (consisting of two forms) and Statistics. The self-reported mastery and self-reported gains items were taken from the pilot student ratings form and the old student ratings form. A total of 848 undergraduate students completed the final exam and the two student ratings forms. The summed total of the self-reported mastery items correlated at a medium strength with objective measures of learning outcomes (Calculus Form A: r = .436; Calculus Form B: r = .361; Statistics: r = .416). The relationship between self-reported gains and objective measures of learning outcomes was weaker than that of self-reported mastery and objective measures of learning outcomes (a difference of .276 for Calculus Form A, .138 for Calculus Form B, .110 for Statistics). The relationship between self-reported gains and self-reported mastery was stronger than the other two relationships (Calculus Form A: r = .473, Calculus Form B: r = .500, Statistics: r = .628). A confirmatory factor analysis produced even stronger relationships between the three latent variables, including differences between the two forms of the Calculus exam. Self-reported mastery may be more effective at measuring objective measures of learning outcomes than self-reported gains, but self-reported mastery cannot completely serve as a proxy for objective measures of learning outcomes. Administrators or researchers measuring learning outcomes on a large scale may benefit by administering self-reported mastery items instead of self-reported gains items.

Keywords: learning outcomes, self-reported mastery, self-reported gains, self-assessment, subjective measurement, measuring ability

ACKNOWLEDGEMENTS

I could not have completed this dissertation without the help of a few people. My wife and kids supported me through late nights and lots of time at the library. (I owe it to all of you.) Russ Osguthorpe and Trav Johnson started me on my path and believed in me through the whole process. Dr. Sudweeks and Dr. Joseph Olsen helped me with the needed statistical knowledge to finish the project. And I thank my Heavenly Father for more support than I am sure I recognized. Thanks to all for helping me through this whole process. I appreciate each one of you.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

Statement of Purpose 3

Research Questions 3

CHAPTER 2: REVIEW OF LITERATURE 4

Concerns with Self-Reported Gains 4

Validity of Self-Reported Gains 5

Self-reported gains and longitudinal gains. 6

Affective measure or cognitive measure? 7

Social desirability 9

Halo effect. 9

Correcting for bias. 10

Theories about the problems with convergent validity and bias. 10

Moderators 15

Differences across universities. 15

Differences between students. 16

Differences in content. 17

Differences in measured content. 17

Feedback. 18

Self-Reported Mastery 18

Single-Item Indicators 22

Reasons to Use Single-Item Indicators 22

Reliability 23

Validity 24
Convergent validity. 24
Content validity. 24
CHAPTER 3: METHOD 26
Participants 26
Instruments 27
Pilot Student Ratings Form 27
Final Exams for Statistics and Calculus 28
Self-Reported Gains Item 28
Procedures 29
Data Analysis 30
Correlations 30
Confirmatory Factor Analysis 30
Psychometric Properties 31
CHAPTER 4: RESULTS 33
Correlations 33
Statistics Course 33
Calculus Course 34
Confirmatory Factor Analysis 34
Test of Invariance for the Calculus Test Forms 34
Goodness-of-Fit 35
Correlations between Latent Variables 36
Statistics course 36

Calculus course. 36
Influences of Measurement Error 37
Psychometric Properties 39
CHAPTER 5: DISCUSSION 41
Correlations 41
Measurement Error Concerns 44
Psychometric Properties 45
Recommendations 47
Strengths 51
Weaknesses 51
References 54
Appendix A 58
Appendix B 66
Appendix C 74

LIST OF TABLES

Table 1. Participants 27

Table 2. Correlations between the Various Self-Reported Measures and Objective Measures of Learning (Final Exam) for the Statistics Course 33

Table 3. Correlations between the Various Self-Reported Measures and Score on the Final Exam for the Calculus Course 34

Table 4. Goodness-of-Fit Results for the Test of Invariance for Calculus Forms A and B 35

Table 5. Goodness-of-Fit Results for the Confirmatory Factor Analysis of Statistics and Calculus Form A and B Assuming Weak Invariance 36

Table 6. Correlations between Latent Variables in the Confirmatory Factor Analysis of Calculus Form A and B, Assuming Weak Invariance, and Statistics 37

Table 7. Attenuated Correlations and Correlations between the Latent Variables for Statistics 38

Table 8. Attenuated Correlations and Correlations between the Latent Variables for Calculus Form A and Calculus Form B 38

Table 9. Distribution of Responses to the Self-Reported Mastery Items by Course and Expected Learning Outcome for Statistics 40

Table 10. Distribution of Responses to the Self-Reported Mastery Items by Course and Expected Learning Outcome for Calculus 40

LIST OF FIGURES

Figure 1. Path Diagram of the CFA Model for Self-Reported Mastery, Self-Reported Gain, and Course Achievement for Statistics and Calculus..... 32

CHAPTER 1: INTRODUCTION

The learning outcomes movement has burgeoned in the last six years. More than half the articles cited in the Educational Resources Information Center (ERIC) with the phrase *learning outcomes* in the title have been published since 2005. The term *learning outcomes* was first included in the title of a research article in 1961 and has increased in use consistently (almost on a yearly basis) since that time. The push to measure learning outcomes of university students may have increased with the publication of a report by the Commission on the Future of Higher Education, better known as the Spellings' Commission (Spellings' Commission on the Future of Higher Education, 2006). Their report challenged universities to improve four standards, "access, affordability, quality, and accountability" (p. xiii) and stated that the "lack of useful data and accountability hinders policymakers and the public from making informed decisions and prevents higher education from demonstrating its contribution to the public good." (p. 4). Following the challenge by the Spellings Commission, universities have been pressed to find effective methods for collecting "useful data" indicating the degree to which the specified learning outcomes for a course have been realized.

One method used to measure learning outcomes has been students' self-reported gain. An example of an item intended to measure a self-reported gain of a learning outcome is: "How much did you learn about the theory of relativity during this class?" Many large scale assessments, like the National Survey of Student Engagement, use self-reported gains to measure learning outcomes on the part of university students. The challenge is that many researchers believe that self-reported gains are not valid measures of learning outcomes. They claim that self-reported gains lack convergent validity (Bowman, 2009, 2011b; Gosen & Washburn, 1999;

Herzog, 2011; Pohlmann & Beggs, 1974) and are biased (Bowman & Hill, 2011; Gonyea & Miller, 2011; Pike, 1993; Pike, 1999).

These critics appear to be convinced that self-reported gains cannot be used as proxies for achievement measures, but do the problems with self-reported gains apply to all self-assessments? Some researchers appear to lump all self-reports into one category, but others have divided self-assessments into categories (Sitzmann, Ely, Brown, & Bauer, 2010).

One type of self-assessment, self-reported mastery, appears to avoid some of the claims about lack of validity. (Self-reported mastery has been given a name in this document because of its lack of a name in the literature). The main difference between self-reported gains and self-reported mastery is that instead of focusing on a gain in ability, self-reported mastery focuses on a student's assessment of their ability at a single point in time. An item measuring self-reported mastery may ask, "What was your level of mastery of the laws of thermodynamics at the end of this course?" Students no longer need to judge the change in their ability between two points in time, but can focus on estimating their ability at a single point in time. Therefore, self-reported mastery reflects a student's judgment of their status at the time of the assessment rather than being a judgment of their growth or progress.

The use of self-reported mastery may not resolve all of the problems inherent in self-assessment, but may be a more promising solution to measuring learning outcomes than self-reported gains. Self-reported mastery has been found to correlate stronger with cognitive measures of achievement than self-reported gains (Berdie, 1971; Sitzmann et al., 2010), and may be less susceptible to measurement error because of a reduction in the misjudgment of time that may be happening in self-reported gains (Bowman, 2009).

Statement of Purpose

The purpose of this study was to determine to what extent self-reported mastery can serve as a proxy for objective measures of course achievement when assessing learning outcomes.

Research Questions

The study focused on the following questions:

1. What are the correlations between students' (a) self-reported mastery of the expected learning outcomes in a course, (b) self-reported learning gain, and (c) an objective measure of their course achievement as indicated by their score on the final examination?
2. To what extent are the observed correlations influenced by measurement error?
3. What are the psychometric properties of the various items and tests in terms of traditional item and test analysis statistics (frequency distributions, point-biserial correlation coefficients, and Cronbach's alpha reliability coefficients)?

CHAPTER 2: REVIEW OF LITERATURE

Self-reported measures, mainly in the form of self-reported gains, have been used to assess learning for decades. Self-reported measures have become more common as universities have attempted to assess a wider variety of issues through cost effective means (Gonyea, 2005). Because of the simplicity and cost effectiveness of self-assessments, it is unlikely that self-assessments will be disregarded; but self-reported mastery may be preferable if it avoids some of the pitfalls of self-reported gains.

Self-reported gains have been criticized in the literature, mainly in favor of longitudinal gains (Bowman, 2009, 2011b; Gonyea & Miller, 2011; Herzog, 2011). Unfortunately, longitudinal gains (often limited to a pretest-posttest differences) are costly, time consuming, and have problems of their own, although the problems appear to diminish with larger sample sizes (Baird, 1988; Zimmerman, 2009). If the issues found in the literature concerning self-reported gains are valid, then it may be helpful to know if these concerns affect all self-assessments, including self-reported mastery. In the following literature review, I will review the concerns with self-reported gains, and then the literature about self-reported mastery.

Concerns with Self-Reported Gains

Two main issues concern self-reported gains, validity and moderators. The first concern, validity, is whether self-reported gains measure what they are expected to measure. Many researchers have found that self-reported gains correlate with extraneous factors and do not appear to correlate with cognitive gains. The second concern, moderators, is that many factors appear to influence the strength of the correlation with cognitive measures. Controlling for these factors or strengthening the presence of these factors may be a challenge. These two concerns will be discussed in detail below.

Validity of Self-Reported Gains

Validity is one of the main concerns with self-reported learning gains. Issues dealing with the validity of self-reported gains can be broken down into two areas: convergent validity and bias.

In convergent validity, the researcher is concerned with how as assessment relates to other similar assessments. If the assessment does not relate to other similar assessments, then it is possible that the assessment is measuring a different construct than the other assessments. With self-reported gains in learning, evaluating convergent validity consists of comparing self-reported gains to other measures of cognitive gains, like the gain scores obtained from pretests and posttests. Convergent validity can also be challenged by showing that self-reported gains in learning relate to other measures that appear not to measure cognitive gains, like affective measures, especially if the relationship with other measures is stronger than the relationship with cognitive gains. The relationships between assessments are often gauged through simple correlations or linear regression.

Another concern with validity is bias. A bias is any factor that systematically distorts scores obtained from a measure away from the target construct intended to be measured. Bias is often revealed by showing that a measure is related to an extraneous construct. For example, if an item intended to measure math ability, but has a strong correlation with reading ability, the correlation may indicate that students are being judged on both their math ability and their reading ability; students with poor reading ability may be perceived to have poor math ability because of the prerequisite. Biases can lower the validity of a measure and may result in erroneous conclusions about the participants.

Below I will discuss self-reported gains and longitudinal gains, whether self-reported gains are an affective measure or cognitive measure, the biases of social desirability and halo effect, the possibility of correcting for bias, and theories about the problems with convergent validity and bias. The findings in the first four sections generally point to low convergent validity and the presence of bias. Some researchers have attempted to correct for the biases found in self-reported gains with some success. Researchers have attempted to theorize why self-reported gains have low convergent validity and many biases. The theories help point future research in directions that may improve self-assessment.

Self-reported gains and longitudinal gains. Evidence of longitudinal learning gains typically consist of multiple data points across periods of time. Most of the longitudinal gains in the following studies consisted of two data points taken at the beginning and end of the time period (a pretest and posttest format).

When comparing longitudinal gains with self-reported gains measuring the same construct, a high correlation would be expected. Unfortunately, self-reported gains typically have a low correlation with longitudinal gains (Bowman, 2009, 2011b; Gosen & Washburn, 1999; Herzog, 2011; Pohlmann & Beggs, 1974). Bowman (2009) in a first study reported correlations between self-reported gains and longitudinal gains to be between -.01 and .22 and the adjusted correlations in the second study (2011a) to be between -.06 and .25. Herzog (2011) reported the correlations between self-reported gains and longitudinal gains to be between -.11 and .19. Pohlmann and Beggs found correlations of .21 for simple cognitive measures and .10 for complex cognitive measures between self-reported gains in learning and posttests while controlling for the pretest. Gosen and Washburn (1999) also found similar findings with correlations between objective longitudinal gains and self-reported gains, between -.375 and

.243. Consistently correlations between longitudinal gains and self-reported gains result in small correlations, and at times even negative correlations. The highest correlations in these studies explained less than 7% of the variance according to the coefficient of determination. Even though longitudinal gains may have some challenges, these low correlations do not appear to bolster the convergent validity of self-reported gains.

Researchers have hypothesized that self-reported gains would correlate stronger with longitudinal gains if the evidence of longitudinal gains was subjective in nature, because self-reported gains are also subjective. Bowman (2009) found that even though subjective longitudinal gains correlated stronger with self-reported gains (*r*s = .18 to .24) than objective longitudinal gains (*r*s = -.01 to .03), the correlation was still smaller than the correlations between the self-reported gains themselves (*r*s = .32 to .55). These findings also point toward low convergent validity of self-reported gains with longitudinal gains.

Bowman (2009, 2011a) found that predictors for longitudinal gains and self-reported gains are frequently significantly different and, at times, opposite. In the 2011 study, Bowman found that 15 of the 48 predictors were significantly different (chance would have limited the difference to two predictors). In the 2009 study, Bowman found many differences between predictors including opposite signs on 3 out of the 43 predictors.

Both low correlations and differences in predictors point toward low convergent validity between longitudinal gains and self-reported gains. In this case, subjectivity appears to be a slight moderator, improving correlations, but correlations continue to be small.

Affective measure or cognitive measure? Self-reported gains in learning should correlate with other cognitive measures of learning. Sitzmann, Ely, Brown, and Bauer (2010) claimed that self-reported gains correlate stronger with affective measures than with cognitive

measures. They conducted a meta-analysis reviewing 166 self-reported assessments. Findings showed that the "mean correlation corrected for measurement error based on predictor and criterion reliabilities" (p. 176) of self-assessments of knowledge with cognitive gains was .34, while the mean corrected correlation with motivation was .59 and mean corrected correlation with reactions (satisfaction) was .51. Both reactions (satisfaction) and motivation are considered affective outcomes and both had considerably stronger correlations than cognitive gains.

Multiple other studies also found that self-reported gains correlated with affective measures (Bowman & Hill, 2011; Gonyea & Miller, 2011; Pike, 1993; Pohlmann & Beggs, 1974). Bowman and Hill (2011) divided their study into first-year students and second-year students and beyond. First-year students correlated at .34 with college satisfaction, .30 with narcissism and at .53 with personal growth. Second-year students and beyond correlated .38 with college satisfaction, .31 with personal growth, and .46 with self-esteem. All of the above correlations represented relationships with affective measures. Other researchers also found that personal growth was a strong predictor for self-reported gains, stronger than the relationship with satisfaction (Pike, 1993; Pohlmann & Beggs, 1974).

Gonyea and Miller (2011) also found that self-reported gains correlated with affective measures, including deep learning (*r*s = .48 to .51) and overall satisfaction (*r*s = .37 to .50). Interestingly enough, they found other measures equally as strong, or more so, that dealt with environmental measures, the highest of which was a supportive campus environment (*r*s = .53 to .58). There is the possibility that self-reported gains correlate more with environmental variables than affective measures, although the environmental variables like supportive campus environment and level of academic challenge may also be highly connected with affective measures.

These findings concerning affective measures point to low convergent validity. Self-reported gains correlate with various other factors higher than with cognitive measures.

Social desirability. Social desirability is a bias caused by a respondent desiring to please others while answering an item. Researchers have been concerned whether students can answer subjective items without the effects of social desirability. Results concerning social desirability have been mixed.

Bowman and Hill (2011) found that first-year students are influenced by social desirability, but not other students. The correlation between self-reported gains and a social desirability scale was .32 for first year students, but only .06 for other students (and not significant). Because of the lack of bias found in more experienced students, the researchers stated that self-reported gains may be somewhat useful "among more advanced undergraduates" (p. 83).

Gonyea and Miller (2011) found little influence of social desirability in the NSSE, an assessment consisting of self-reported gains for either first-year students or seniors. The correlations for self-reported gains and the social desirability scale were between .06 and .14. Seniors had a higher correlation in two of the three types of self-reported gains, but by no more than a .06 difference.

Halo effect. Halo error is an alternative explanation to the relationship between self-reported gains and affective measures. Bowman (2009) defined halo error as "the tendency to respond to specific items based on general perceptions of a subject" (p. 4). Pike (1993), from a sample of 989 seniors, created two statistical models, with satisfaction in a direct relationship with perceived learning, or as "an artifact of a halo effect" (p. 24). The model emphasizing halo error proved to have a better fit than the model based on satisfaction as a direct relationship with

perceived learning. Even though the findings pointed toward halo error being the stronger model, Pike stated that he was not convinced halo error was a better fit because of a lack of variance in latent variables.

Pike (1999) again explored the possibility of halo error with two more studies. Confirmatory factor analysis pointed to stronger evidence that halo error was a factor in self-reported gains, although more of a factor for freshmen than for seniors. The halo error for freshmen explained more than half of the variance, while for seniors halo error explained somewhere between a fourth and half of the variance. Pike noted that his study was at a single university and might not be generalizable.

Correcting for bias. Some researchers have attempted to correct for bias present in self-reported gains. The main method used has been to collect self-reported high school gains, retrospectively, and use the data to correct for other biases (Bowman & Hill, 2011; Pascarella, 2001; Seifert & Asel, 2011). Pascarella (2001) suggested using students' self-reports of their high school experience to control for other bias. Researchers have found a high correlation with self-reported gains for first-year students and a weaker correlation for other students (Bowman & Hill, 2011; Seifert & Asel, 2011). Seifert and Asel (2011) found that controlling for high school self-reported gains explained between 1.7% and 8.9% more variance depending on the scale and whether the students were first year or seniors. Correcting for bias may be one method to improving the validity of self-reported gains in learning.

Theories about the problems with convergent validity and bias. A variety of researchers have proposed theories about the reasons for the problems with convergent validity and bias. Three of these theories involve the measurement of different constructs, misjudgment of time, and survey error.

The measurement of different constructs. One of the theories that attempts to explain the problem concerning the validity of self-reported gains is that self-reported gains measure a different construct than longitudinal gains (Bowman, 2009, 2011b; Herzog, 2011). As noted above, affective measures, environmental measures, and personal growth appear to have a stronger correlation with self-reported gains than with cognitive measures, which may mean that self-reported gains measure a construct other than cognitive gains. No overarching construct has been found to explain all the different correlations with self-reported gains. Pike (1996), for example, found that halo error was stronger than a model connected to satisfaction, but was not confident in his findings. Considering the complexity of human motivations, there may not be one specific construct, but many, that influence self-reported gains.

Another theory about the construct connected to self-reported gains is that the relationships involving satisfaction, halo error, personal growth, and environmental variables may be authentic relationships. Pike (1993) pointed out that previous researchers have found a relationship between aspects of learning and satisfaction, and the nature of that relationship is not understood. Satisfaction, personal growth, and environmental variables may be expected to correlate with learning on a regular basis, and a halo effect may be a normal relationship between items.

A theory not stated in the literature that would result in a different construct is that students may have a different definition of learning than the definition outlined by the research community. Evidence of this different construct may be found in the high correlations of self-reported gains with deep learning (Gonyea & Miller, 2011) and personal growth (Bowen & Hill, 2011; Pike, 1993; Pohlmann & Beggs, 1974). If students define learning as deep learning or personal growth, then an objective longitudinal assessment could demonstrate gains in learning,

while students fail to report a gain because of a belief that the knowledge did not contribute to their learning. Students may express that they memorized knowledge to perform well on an exam and then quickly proceeded to forget the knowledge. Results of students' self-reported gains would coincide with objective measures only when both definitions of learning by students and the research community correspond. This would result in constructs that overlap and diverge depending on when the definitions of learning correspond.

Different definitions of learning may cause administrators and professors to make poor decisions based on their perceptions of the data. For example, at times students may not understand the purpose behind their learning, but come to understand that purpose later as they progress in a discipline. The resulting low scores in student ratings may cause professors to change their teaching methods, or administrators to not fund the area of study. Another scenario that would result in lower student ratings scores would be a class where the content applies to some of the students' future careers, but not all of the students, resulting in lower scores for some of the students and higher scores for those who feel they are achieving personal growth. No matter the circumstances, the definition of learning has to be shared by the administrators and students in order to avoid misunderstanding and misusing the results.

In contrast to the theory that self-reported gains and longitudinal measures measure different constructs is the idea that they measure the same construct, but on opposite ends of a spectrum. Astin (1993) stated that objective tests are accurate at measuring specific content, but do not cover a broad range of content. Bowman (2009) hypothesized that self-reported gains may be opposite in nature, covering more content, but with less exactness. Although not completely aligned with this theory, Pike (1996) believed that self-reported gains and cognitive measures may be congeneric—related to the same construct, but not equivalent. Even if the two

measures are measuring the same construct, the differences between the two measures cause the dilemma of which measure we should chose when reaching our decisions.

Misjudgment of time. Another theory concerning the problems with self-reported gains is that students misjudge the concept of time when measuring learning. Bowman (2009) pointed out that to effectively answer a self-reported gains item, theoretically students have to determine their ability at both the beginning and end of the requested time period, and then compare the two to report the difference. Bowman believed that students may be shortcutting the process. He notes research on satisficing, or students giving satisfactory answers on more cognitively challenging items. Bowman believed that self-reported gains are cognitively challenging items, which would result in students satisficing. Another possibility is that students may not recognize the existence of multiple steps in a self-reported gains item, and instead lean toward an intuitive answer, which may correspond better with deep learning and satisfaction, and not changes over time.

Self-reported gains in learning correlate stronger with measurements of a single point in time instead of multiple points as found in longitudinal gains. This fact may support the theory that students are misjudging the concept of time, that students are reporting their results at a single point in time at the end of the learning period instead of their changes across a period of time. Cohen (1981), in his meta-analysis of student ratings, found that self-reported gains in learning correlated with cognitive measures at a single point in time at .47. Cohen's results were much stronger than any of the correlations self-reported gains had with longitudinal gains. Not all of the studies Cohen included in his meta-analysis had such strong correlations though. He reported that two had negative correlations, two had no correlation, and ten had positive correlations. The broad range of findings may not completely support that self-reported gains

measure a single point in time. Pike (1995, 1996) more recently found a strong relationship when comparing self-reported gains to a single point in time. Pike used a cognitive measure called College BASE, which measured a single point in time, and found a strong relationship with self-reported gains using structural equation modeling. The strong relationships between measurements of a single point in time and self-reported gains bring up doubt that self-reported gains are actually measuring gains. If self-reported gains do not measure gains, it may be more effective to focus on measuring ability at one point in time to specify the item and increase the correlation with cognitive gains.

Measurement error. The apparent reason for biases like halo error and social desirability is an error in human judgment, but there may be other reasons for these biases; another possibility may be measurement error. Bowman (2009) pointed out that halo error may be influenced by the consecutive nature of survey items. Halo error occurs when overall perceptions of a subject influence a student's response, but if a student was influenced to respond in a similar way by a chain of similar items, it may statistically appear like halo error. Gonyea and Miller (2011) also noted the importance of examining the influence of the order of survey items. These influences could be examined through statistical methods common to developing surveys. Measurement error could explain some of the inconsistent findings, like Bowman and Hill (2011) finding that self-reported gains correlated with a social desirability scale, while Gonyea and Miller (2011) did not. The NSSE (reported by Gonyea and Miller in 2011) is a well refined tool that may have less influence between items than other studies. Although plausible, future research may be needed to support Gonyea and Miller's claim that measurement error found in the survey items may be influencing results.

Moderators

Sitzmann et al. (2010) found that self-assessments had various moderators. Moderators, in this case, consist of different variables that change the strength of the relationship between self-reported gains and cognitive measures. Other researchers have also found various variables strengthen and weaken the correlation between self-reported gains and cognitive measures (Bowman, 2011b; Bowman & Hill, 2011; Gonyea & Miller, 2011; Pike, 1995; Pike, 1996).

Moderators result in contexts that shift the strength of the relationship between self-reported gains and cognitive measures. These differences result in instable statistics. Self-assessments cannot be used as proxies for other measures unless they are consistent across contexts (Ewell, Lovell, Dressier, & Jones, 1993 as cited in Pike, 1995). Some of these inconsistencies could be controlled. If the inconsistencies remain, the resulting error would produce faulty data that could cause incorrect decisions.

Moderators are different from bias in that moderators are not the result of error. Biases systematically pull the statistic away from the population parameter. Moderator variables change the relationship between two other variables. The changes in the relationship produced by moderators may cause misunderstandings of relationships if the moderator is not accounted for or left concealed. Moderators in self-assessments can be found across universities, student bodies, subject matter, measured content, and whether feedback is received.

Differences across universities. Inconsistencies in the correlation between self-reported gains and other measures have been reported by Bowman (2011b) and Pike (1996). Bowman (2011b) found two types of differences across universities. While correlating self-reported gains with longitudinal gains, selective schools correlated .07 higher (r=.18 compared to r=.11) than less selective schools on two of the four constructs, while liberal arts colleges compared to

universities correlated .12 higher on one construct (r=.08 compared to *r*=-.04) and .07 higher on another (*r*=.18 compared to .11). Although not large differences, even small differences make comparisons across universities challenging.

Pike (1996) found a difference between two- and four-year colleges. He reported that there were small differences in the method factors. The differences in method factors meant that the learning outcomes, specifically in this study: mathematics, science, English, and social studies, may be different between two-year and four-year colleges. Comparisons between universities will be slanted if learning outcomes are not parallel across universities.

Differences between students. Differences have been found between different groups of students who have responded to self-reported gains. These groups of students consist of first-year students compared to other students and first-generation college students compared to other students (Bowman, 2011b; Bowman & Hill, 2011; Gonyea & Miller, 2011). Bowman (2011b) found that the correlation between self-reported gains and longitudinal gains were stronger for first generation students, although he did not report by how much. Bowman and Hill (2011) found that biases were different between students in their first year of school and other students. The social desirability index, narcissism, personal growth, and self-esteem when correlated with self-reported gains for first year students and other students had a difference of at least .20. All the factors but self-esteem were higher for first-year students; Bowman claimed that self-reported gains may be useful for more advanced students because the biases were less correlated. Gonyea and Miller (2011), on the other hand, did not find large differences between university first-year students and seniors when looking at social desirability, the largest difference being .06 for gains in personal and social development down to no difference for gains in practical

competence. Universities may have to individually verify the amount of variance across the student population because of the lack of uniformity in the findings.

Differences in content. Correlations between self-assessments and cognitive measures change depending on the subject matter being measured. Sitzmann et al. (2010) found that self-reported gains correlated with cognitive measures at .41 when content focused on interpersonal content, cognitive content correlated at .25, and psychomotor content correlated at .15. Pike (1995) also found different relationships using structural equation modeling of content measured with self-reported gains and an objective assessment. Pike concluded that both self-reported gains and the objective assessment measured the same construct in math, but did not measure the same construct in social studies, while English and science was between math and social studies. Within each of the content areas, there were subsections that had a stronger or weaker relationship between self-reported gains and the objective assessment. English had a strong relationship with the subscale writing, but not with reading and literature. Social studies had a strong relationship with the subscale history, but not with social science. It appears that certain types of content are better measured by self-reported gains than others; these differences must be understood before self-reported gains can be used effectively.

Differences in measured content. Correlations between self-reported gains and other types of measures are stronger when the items measure similar content. Sitzmann et al. (2010) found in their meta-analysis that when self-assessments were similar to objective measures in content, the correlation was .36, but when they were dissimilar the correlation was .19 (an increase of .17). Pike (1995) called this principle "content correspondence." He found that content correspondence was highest for English and math, then science, and lastly for social

studies. Similarly, the strongest relationships between the objective assessment and self-reported gains were found in math, followed by English, then science, and lastly with social studies.

Previous studies that compared self-reported gains to cognitive measures have focused on broad learning outcomes (Bowman, 2009, 2011b; Bowman & Hill, 2011b; Herzog, 2011; Pike 1995; Pike, 1996), which may cause a lessening in content correspondence between objective measures and self-reported gains. Future studies could focus on more specific learning outcomes, possibly even at the class level, to evaluate whether content correspondence improves the relationship between self-assessments and objective measures.

Feedback. Two types of feedback have been found to improve the correlation between self-reported gains and cognitive measures. Sitzmann et al. (2010) found that students who received feedback on their performance increase the correlation from .11 to .21. In the same meta-analysis, the researchers also found that students who received feedback on their self-assessments throughout a course also increased the correlation between assessments and cognitive measures from .23 to .40. A combination of these two types of feedback would be an increase of .27, although in reality the actual increase of applying these two moderators at the same time may result in a different outcome. The challenge is that feedback is typically in the professors' hands and will most likely be included only if it increases the ratings for the learning outcomes. As students are often overconfident unless knowledgeable (Kennedy, Lawton, & Plumlee, 2002), feedback may be easier to implement at higher level classes and more difficult at lower level classes.

Self-Reported Mastery

The majority of studies in the educational literature about self-assessments has focused on self-reported gains. The absence of studies about self-reported mastery may be because of the

desire to capture learning, which is conceptualized as a change in knowledge, understanding, or some other affective or behavioral trait. In other words, learning is perceived to be evidenced by a gain. Self-reported mastery measures an attribute at a single point in time. In order to measure a gain, self-reported mastery would have to be administered at two points of time, as a subjective longitudinal gain, or as part of a retrospective pretest-posttest design. Most have assumed that self-reported gains and self-reported mastery fall under the same umbrella of self-assessment and should have the same results. The two studies outlined below provide a different viewpoint.

Sitzmann et al. (2010) conducted a meta-analysis reviewing self-assessments, not only in the area of education, but also in communication, psychology, business, and other disciplines. The researchers compared the difference between self-reported gains and self-reported mastery, along with other factors discussed above. Even though self-reported mastery is not common in education, the researchers found 108 studies exploring self-reported mastery, and 25 studies exploring self-reported gains. Self-reported mastery had a sample weighted mean correlation of .34 (adjusted sample weighted mean correlation of .44), while self-reported gains had a sample weighted mean correlation of .00 (adjusted sample weighted mean correlation of .00). The difference in the correlations of .34 (or .44 when adjusted) was the largest difference found in the moderators of the researchers' study. Even though the findings appear to be solid evidence of the superior performance of self-reported mastery over self-reported gains, it would be rash to jump to the conclusion. This meta-analysis did not report the methods used by each study to measure cognitive learning, and as noted earlier, self-reported gains in education have been found to correlate positively with cognitive measures (Cohen, 1981). These findings do point to the need to further evaluate the differences between self-reported gains and self-reported mastery.

Berdie (1971) wrote the only article in the educational literature which the author found that focused on self-reported mastery. Berdie asked a sample of 216 students to report their level of knowledge on a 3-point scale about 12 public figures, 13 authors, and 14 painters. (The three options for public figures were "know who he is, have heard of him but cannot identify him, have never heard of him;" for authors were "read a book by him, heard of him but have not read a book by him, have never heard of him;" and for painters were "seen a picture by him, heard of him but have not seen a picture by him, have never heard of him," p. 631). The students were then given an objective multiple choice test of 39 items. Each item asked students to identify one of the person's achievements from five options. Berdie obtained the students from two samples and reported the samples divided between male and female. Overall correlations for the four groups were .47, .65, .67, and .74. Subsections of the test showed that public figures had the highest correlations, ranging from .40 to .76; authors had the middle range of correlations, from .30 to .69, and artists had the lowest correlations, from -.07 to .07. Berdie points out that "we have here an excellent method for determining the extent of a person's ignorance, perhaps a less satisfactory method for determining the extent of his knowledge" (p. 635-636).

Berdie's study displayed possible evidence that self-reported mastery may have moderators similar to self-reported gains, specifically differences across subject matter. If the results for self-reported mastery are not consistent, then self-reported mastery cannot be a proxy for cognitive measures of ability.

The wide range of results in Berdie's study could also have been because some of the items did not measure the full range of expertise. The item about painters, for example, asked if students had seen a single painting of an artists, but were then given 39 multiple choice questions about that artist covering a wide range of their works. This would be like asking if we had seen

any paintings by Edward Munch. Many people would be able to remember *The Scream*, but if asked to identify his paintings *The Kiss* or *Death in the Sickroom* from a list of other paintings, we would be hard pressed to succeed. To better determine expertise, the three point scale could have been expanded to ask if they had seen multiple paintings by the artist, or could name multiple paintings by the artists. Future use of self-reported mastery may need to evaluate the effectiveness of each item to verify that it measures the full range of the expected construct.

Self-reported mastery also avoids the need for students to judge a change in ability over time. Bowman (2009), as noted above, pointed out that students responding to self-reported gains may not complete the steps needed to judge a change in an attribute over time. Self-reported mastery focuses on one point in time, which removes the error caused by students' misjudgment of time. The amount of error removed is unknown, but could explain the stronger correlations between self-reported mastery and cognitive measures. Simplicity often helps to reduce error, and self-reported mastery may require fewer steps for student to measure their ability than when responding to self-reported gains. Combining simplicity with the lack of measuring a change in ability over time may improve convergent validity and reduce bias.

Bias has been found when self-reported mastery is used as part of other methods. Douglass, Thomson, and Zhao (2012) used the retrospective pretest-posttest method to measure learning outcomes as an alternative to standardized tests and self-reported gains. The retrospective pretest-posttest method uses a self-reported mastery item as the posttest method, and asks them to retrospectively measure their ability at the beginning of the period. The differences between the self-reported mastery item and the retrospective item results in a gains score. Reviewing the data from the retrospective pretest-posttest items, Douglass, Thomson, and Zhao (2012) found that Asians reported lower scores than other ethnic groups. They also noted

that previous studies have found that retrospective pretest-posttest methods have an upward bias. Self-reported mastery, being part of the retrospective pretest-posttest method, may have some of these same biases. Other biases affecting self-reported gains should also be researched to assure that self-reported mastery is an effective method.

Single-Item Indicators

Measurement experts typically suggest using multiple items to measure a construct and frown on any attempts to use one item to measure a construct (Diamantopoulos, Sarstedt, Fuchs, Wilezynski, & Kaiser, 2012; Fuchs and Diamantopoulos, 2009). Even though there is a negative view toward using single items to measure a construct, there may be a few instances where a single item may be useful and may still maintain a measure of validity and reliability (Diamantopoulos et al., 2012; Fuchs & Diamantopoulos, 2009; Postmes, Haslam, & Jans, 2013; Spörrle & Bekk, 2014; Wanous, Reichers, & Hudy, 1997).

Reasons to Use Single-Item Indicators

Single-item indicators may have the benefit of "[taking] up less space and time (for responding as well as data coding)" (Spörrle & Bekk, 2014, p. 272). These benefits are extremely important when respondent have little time to complete an evaluation (Credé, Harms, Niehorster, & Gaye-Valentine, 2012). Single-item indicators have been used to measure many constructs (the concepts or ideas that the researchers are attempting to measure), including job satisfaction (Wanous et al., 1997), organizational justice (Jordan & Turner, 2008), power bases in organizations (Schriesheim, Hinkin, & Podsakoff, 1991), mathematical anxiety (Núñez-Peña, Guilera, & Suárez-Pellicioni, 2014), quality of instructors and subjects in education (Ginns & Barrie, 2004), social identification (Postmes et al., 2013), personality (Credé, Harms, Niehorster, & Gaye-Valentine, 2012; Spörrle & Bekk, 2014), pleasure and arousal (Russell, Weiss, &

Mendelsohn, 1989), group cohesiveness in psychotherapy (Hornsey, Olsen, Barlow, & Oei, 2012), and many other areas (Postmes, Haslam, & Jans, 2013).

Even though single-item indicators are prevalently administered, their use may be a mistake if the results are not reliable and valid. Some authors, for example, have claimed that single-item indicators lead to false conclusions about constructs (Credé et al., 2012; Schriesheim, Hinkin, & Podsakoff, 1991). Are there scenarios in which single-item indicators are valid and reliable?

Reliability

Reports of reliability of single-item indicators has not been consistent both within and across constructs. Reliability of single-item indicators has been recorded across a range of low to high levels (Postmes et al., 2013; Spörrle & Bekk, 2014; Wanous et al., 1997).

Postmes, Haslam, and Jans (2013) conducted a meta-analysis of 16 articles focused on single-item indicators. The authors used the correlation with the larger scale and the scale alpha to estimate the reliability of the single-item indicators. The average reliability of the single-item indicators was .51 with a range between .14 and .68.

Other researchers have conducted meta-analyses of specific single-item indicators. Wanous, Reichers, and Hudy (1997) collected 16 articles measuring overall job satisfaction through single-item indicators. Minimum estimates of reliability for the single-item indicators ranged from .45 to .69. Spörrle and Bekk (2014) reviewed articles on frequently measured personality traits through single-item indicators. The 33 indicators whose reliability was stability based had a mid 50% range between .68 and .91. The 207 indicators whose reliability was consistency based had a mid 50% range between .35 and .77. Even within specific types of single-item indicators, reliability covers a wide range.

Reliability of single-item indicators appears to cover a wide range across differing constructs, and even within the same construct. Reliability of single-item indicators may need to be measured in each scenario to ensure high reliability, although some of the guidelines provided under validity may also help to improve reliability.

Validity

Even if there are reasons to use single-item indicators, the benefits are worthless if the indicator is invalid. Two areas of validity concerning single-item indicators are discussed in the literature, convergent validity and content validity.

Convergent validity. Many studies have reviewed whether single-item indicators correlate with a large scale measuring the same construct. Reports of convergent validity of single-item indicators have a broad range much like reliability.

Two of the meta-analyses used to report reliability also provided statistics on convergent validity. The meta-analysis by Postmes, Haslam, and Jans (2013) found that the average correlation between the single-item indicators and their corresponding larger scale was .64 with a range between .34 to .76. The meta-analysis by Wanous, Reichers, and Hudy (1997) provide a mean correlation of .63 between single-item indicators and their corresponding larger scale and a mean of .67 when adjusted for unreliability, but did not provide the range. The means of the two studies were only .01 apart, but it is hard to say if this was chance or points to a theme across single-item indicators.

Content validity. High content validity insures that all the areas of a construct are represented by the items purported to measure the targeted construct. Because of this need, Fuchs and Diamantopoulos (2009) advise that single-item indicators should only be used for constructs that are concrete and not complex/multidimensional. Concreteness, as opposed to

abstractness, assures that different raters would view the construct in the same way. A complex or multidimensional construct would need multiple items to cover the breadth of a complex construct. Many of the single-item indicators that are currently in use, for example, job satisfaction (Wanous et al., 1997) or mathematical anxiety (Núñez-Peña et al., 2014), appear to be measures of simple constructs. Some more complex constructs use multiple single-item indicators, each representing a different dimension of the construct. Personality, for example, is a complex construct, and researchers have used single-item indicators to measure each of the personality traits (Credé et al., 2012).

A more recent study (Diamantopoulos et al., 2012) used a Monte Carlo simulation experiment to compare single-item indicators to multi-item scales. They found that the multi-item scale performed better 59.90% of the time, the single-item indicator performed better 14.10% of the time, and 26.00% of the time there was no significant difference. The authors presented guidelines on when to use single-item indicators: (a) sample size is limited to less than 50 participants, (b) effect size is expected to be below .30, and (c) inter-item correlations are above .80 or the alpha is above .90 (suggesting one dimension). The authors proposed that concreteness may not be enough to justify using a single-item indicator, and they recommended using a multi-item scale instead of single-item indicators in most scenarios.

CHAPTER 3: METHOD

The current study collected data from over 2700 students participating in two undergraduate university courses, Calculus and Statistics. Data were collected from final exams, the old student ratings form, and the pilot student ratings form. The data were analyzed through the use of correlations, confirmatory factor analysis, and the measurement of various psychometric properties.

Participants

Students enrolled in two courses, Math 112: Calculus 1 (Calculus) and Stat 121: Principles of Statistics (Statistics), completed the final examination, and were invited to respond to both a pilot version of the new teacher evaluation and the old course and teacher evaluation form at Brigham Young University during Fall semester 2012. A total of 1,781 students completed the Statistics final exam, while 963 students completed the Calculus final exam (Table 1). Of those who completed the final exam in Calculus, 329 students completed both the pilot and the old student ratings form. Out of those who completed the final exam in Statistics, 519 students completed the pilot and old student ratings form. The Calculus final exam also had two forms. Calculus Form A was completed by 476 students and Calculus Form B was completed by 487 students. Of those who completed Calculus Form A, 167 student also completed the pilot and old student ratings form, while 162 students who completed Calculus Form B also completed the pilot and old student ratings form.

Table 1

Participants

	Number of Students	
Examination	Final Exam	Final Exam and Both Students Ratings Forms
Statistics	1781	519
Calculus Form A	476	167
Calculus Form B	487	162

Instruments

Three instruments were used in this study, including the pilot student ratings form, the final exams for Calculus and Statistics, and the old student ratings form. The old student ratings form provided one item, which was a self-reported gains item.

Pilot Student Ratings Form

The pilot version of the student ratings form contained 11 survey items and was administered online. The first three items were self-reported mastery items related to specific course level learning outcomes. An example of the form is displayed in Appendix A. Departments chose multiple expected learning outcomes before the courses began, which were then narrowed down to three expected learning outcomes to be used in this evaluation. The other items on the form were general items that related to the university goals and teaching quality.

The self-reported mastery items for the Calculus test are listed below:

1. (Limits) Evaluate limits of functions described graphically and algebraically, including recognizing when and how a limit does not exist. Write the definition of a derivative or an integral as a limit, and use the limit to compute the derivative or integral.

2. (Differentiation and Integration) Find derivatives and integrals of common functions. Know and apply differentiation rules to compute derivatives. Use geometry, the fundamental theorem of calculus, and u-substitution to compute integrals.
3. (Applications) Use derivatives and integrals to solve common problems, such as optimization, related-rates, approximation, indeterminate limits, and curve sketching for derivatives, and net change and area problems for integrals.

The self-reported mastery items for Statistics were:

1. Understand the importance of data collection and how it dictates the appropriate statistical method and acceptable inference.
2. Understand and communicate using technical language about probability and variation.
3. Interpret and communicate the outcomes of estimation and hypothesis tests in the context of a problem.

Final Exams for Statistics and Calculus

Final exams for both Statistics and Calculus were used as objective measures of students' achievement of the learning outcomes. The Statistics exam consisted of 90 multiple choice items. The Calculus test consisted of 20 multiple-choice items and 8 constructed-response items. The Calculus exam was also divided into two forms; both form A and form B consisted of the same items, but the multiple-choice items (the first 20 items) and the response options for those items were ordered differently, while the constructed response items (the last 8 items) were in the same order.

Self-Reported Gains Item

One self-reported gain item was used from the old student ratings form. A specimen copy of the old student ratings form used for both Calculus and Statistics appears in Appendix B.

The old student ratings form consisted of 23 items: two comprehensive items, 11 items regarding the course, nine items regarding the instructor, one item about the aims of BYU, and one item allowing them to provide a comment. The first item regarding the course, the 14th item, was a self-reported gains item: "I learned a great deal in this course." Eight response options were presented for each item:

1. *Very Strongly Disagree*
2. *Strongly Disagree*
3. *Disagree*
4. *Somewhat Disagree*
5. *Somewhat Agree*
6. *Agree*
7. *Strongly Agree*
8. *Very Strongly Agree*

Procedures

Students were invited to participate through an initial email and two follow-up emails. Professors were also asked to invite their students to fill out both the pilot form and the old student rating form, and were sent one reminder email. Students were allowed to fill out the forms at their own leisure, but completed the forms before taking their final exams. Both student ratings forms were administered online. Final exams were administered in a supervised location on the campus of Brigham Young University.

Data Analysis

To examine the extent to which self-reported mastery can serve as a proxy for objective measures of course achievement when measuring learning outcomes, the three research questions were answered.

Correlations

The Pearson product-moment correlation coefficient was used to describe the relationship of self-reported mastery with both self-reported gains and objective measures of learning outcomes.

Confirmatory Factor Analysis

To further understand the relationships between self-reported mastery, self-reported gains, and course achievement and the influences of measurement error, a confirmatory factor analysis was used to evaluate the relationships between the three variables. The analysis for the two courses was done separately, and the two analysis are represented by the diagram found in Figure 1. Six items were removed from the Statistics exam in order to run the confirmatory analysis. The three self-reported mastery items (LO_K) consisting of the three learning outcomes from each course loaded onto self-reported mastery. A single-item indicator (GM_1), the self-reported gain item from the old student ratings form, loaded onto self-reported gain. The items from the final (X_K), either Statistics or Calculus, loaded onto course achievement.

Because the self-reported gain variable was operationally measured by a single question, the error could not be computed. The average error from the self-reported mastery items, another self-reported item with the same sample of students, was used as an estimate of the error for the single item indicator. In addition, a range of error values between 0.1 and 0.5 was also analyzed to show the possible results for other error values (see Appendix C).

Psychometric Properties

Three analyses were computed to provide data about the psychometric properties of the instruments. Cronbach's alpha was used to estimate the overall reliability of the final exams and self-reported mastery items. A point-biserial correlation was used to evaluate the items. A frequency distribution for the self-reported mastery items was obtained. A test of factorial invariance was used to verify that Calculus form A and B have the same factor structure (including factor loadings, intercepts, and variance). Then a confirmatory factor analysis was used to measure the relationships between self-reported mastery, self-reported gains, and course achievement for the combined data from both forms.

Three software packages were used to conduct the analyses. The Pearson correlations and the Cronbach's alpha coefficients for the final exams were computed using SPSS. The confirmatory factor analysis and test of invariance were conducted using M-Plus. The point-biserial correlations, the counts of response options, and the Cronbach's alpha for the self-reported mastery items were analyzed in Bond&FoxSteps, a version of Winsteps.

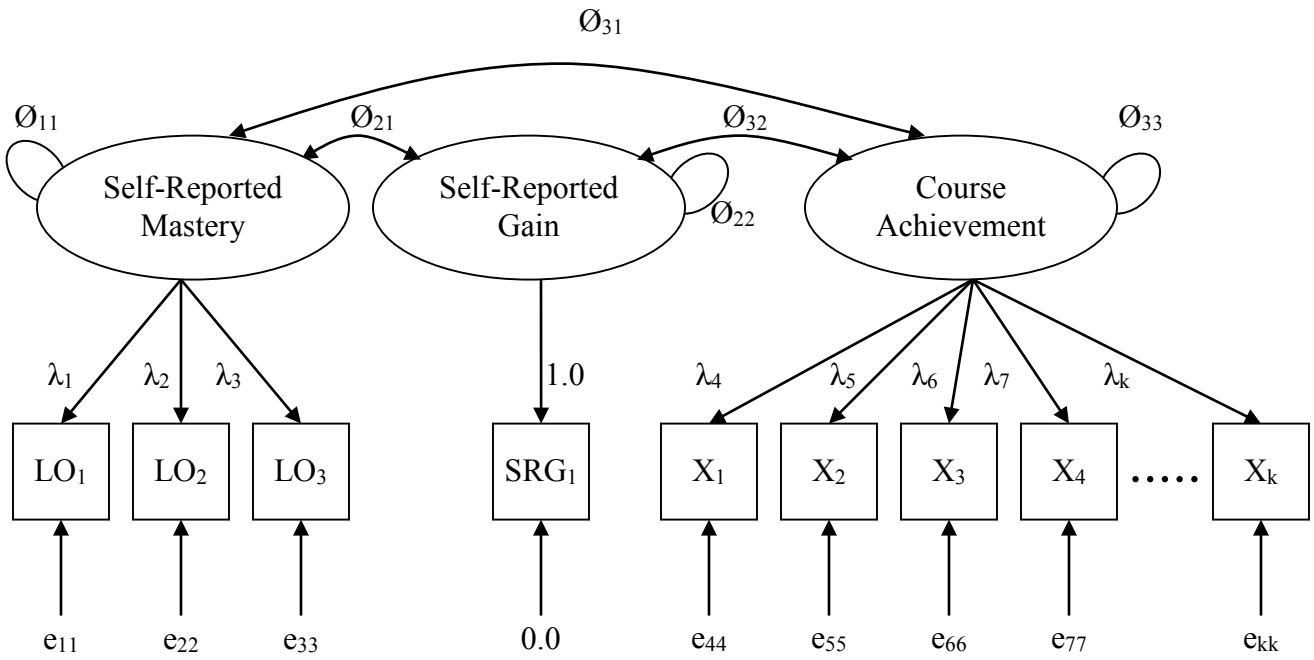

Figure 1. Path Diagram of the CFA Model for Self-Reported Mastery, Self-Reported Gain, and Course Achievement for Statistics and Calculus. (Analysis for the two courses was done separately).

CHAPTER 4: RESULTS

Results from this study were gathered from Pearson product-moment correlations, correlations between the latent variable in a confirmatory factor analysis, and various measures of psychometric properties. Comparisons between the Pearson product-moment correlations and the correlations between the latent variable in the confirmatory factor analysis were used to estimate the influences of measurement error

Correlations

Pearson product-moment correlations were calculate to explore the relationships between self-reported mastery, self-reported gains, and the final exam scores. Correlations were calculated for both the Statistics and Calculus courses.

Statistics Course

The Pearson product-moment correlations between students' showed that the relationships between the composite score of the self-reported mastery items and the objective measure of the learning outcomes (final exam scores) was .416 for Statistics. The Pearson correlations between the self-reported gain items and the objective measures of the learning outcomes was .306 for Statistics. The Pearson correlations between the composite scores of the self-reported mastery items and the self-reported gains items was .628 for Statistics (Table 2).

Table 2

Correlations between the Various Self-Reported Measures and Objective Measures of Learning (Final Exam) for the Statistics Course

Self-Reported Measures	Self-Reported Gain	Final Exam Score
Mastery Item Composite Score	.628	.416
Self-Reported Gain	1.000	.306

Calculus Course

The Pearson product-moment correlations between the composite score of the self-reported mastery items and the objective measure of the learning outcomes (final exam scores) were .436 for Calculus Form A, and .361 for Calculus Form B. The Pearson correlations between the self-reported gain items and the objective measures of the learning outcomes were .160 for Calculus Form A, and .223 for Calculus Form B. The Pearson correlations between the composite scores of the self-reported mastery items and the self-reported gains items were .473 for Calculus Form A, and .500 for Calculus Form B (Table 3).

Table 3

Correlations between the Various Self-Reported Measures and Objective Measures of Learning (Final Exam) for the Calculus Course

Calculus Form A		
Self-Reported Measures	Self-Reported Gain	Final Exam Score
Mastery Item Composite Total	.473	.436
Self-Reported Gain	1.000	.160
Calculus Form B		
Mastery Item Composite Score	.500	.316
Self-Reported Gain	1.000	.223

Confirmatory Factor Analysis

Before conducting the confirmatory factor analysis, a test of invariance for the Calculus test forms was performed. Goodness-of-fit statistics and the correlations between the latent variables were reported from the confirmatory factor analysis.

Test of Invariance for the Calculus Test Forms

Before conducting the confirmatory factor analysis, the decision had to be made on whether to combine Calculus Forms A and B or to conduct the analysis separately. The test for

measurement invariance between Calculus Form A and Calculus Form B showed that weak factorial invariance (factor loadings for Calculus Form A and Form B were constrained to be equal) could be assumed. The chi-square test for difference testing produced a value of 28.798 with 14 degrees of freedom and a p-value of .0112, which with a sample size of close to a thousand participants was enough to assume weak factorial invariance. In testing for strong factorial invariance, the chi-square test for difference testing produced a value of 923.917 with 25 degrees of freedom and a p-value of .0000, which showed that strong factorial invariance could not be assumed. The goodness-of-fit indices also were much stronger for the model testing for weak invariance than the model testing for strong invariance (see Table 4).

Table 4

Goodness-of-Fit Results for the Test of Invariance for Calculus Forms A and B

Goodness-of-Fit Index	Form A and B Weak Invariance Test	Form A and B Strong Invariance Test
CFI	.950	.842
TLI	.954	.861
RMSEA	.036	.064
Chi-square	1462.592	1462.592

Goodness-of-Fit

The goodness-of-fit for the confirmatory factor analysis of the models for Statistics produced a CFI of .914, a TLI of .952, and a RMSEA of .031. The confirmatory factor analysis of Calculus with weak invariance produced a CFI of .951, a TLI of .956, and a RMSEA of .038. (see Table 5).

Table 5

Goodness-of-Fit Results for the Confirmatory Factor Analysis of Statistics and Calculus Form A and B Assuming Weak Invariance

Goodness-of-Fit Index	Statistics	Calculus Form A and B assuming Weak Invariance
CFI	.914	.951
TLI	.952	.956
RMSEA	.031	.038
Chi-Square	1399.621	460.575

Correlations between Latent Variables

Correlations between the latent variables were provided to help understand the relationships between self-reported mastery, self-reported gain, and course achievement. Correlations between the latent variables were found for both the Statistics course and the Calculus course.

Statistics course. The results of the confirmatory factor analysis showed that the correlation between the latent variables for self-reported mastery and course achievement was .449 for Statistics. The correlation between the latent variables for self-reported gain and course achievement was .346 for Statistics. The correlation between the latent variables between self-reported mastery and self-reported gain was .726 for Statistics (see Table 6).

Calculus course. The results of the confirmatory factor analysis, assuming weak invariance for Calculus Form A and B, showed that the correlations between the latent variables for self-reported mastery and course achievement were .538 for Calculus Form A, and .365 for Calculus Form B. The correlations between the latent variables for self-reported gain and course achievement were .204 for Calculus Form A, and .276 for Calculus Form B. The correlations

between the latent variables between self-reported mastery and self-reported gain were .609 for Calculus Form A, and .609 for Calculus Form B, (see Table 6).

Table 6

Correlations between Latent Variables in the Confirmatory Factor Analysis of Calculus Form A and B, Assuming Weak Invariance, and Statistics

Examination	Self-reported mastery with self-reported gain	Self-reported mastery with course achievement	Self-reported gain with course achievement
Calculus Form A	.609	.538	.204
Calculus Form B	.609	.365	.276
Statistics	.726	.449	.346

Influences of Measurement Error

For comparison purposes, the results from the attenuated correlations and the confirmatory factor analysis have been juxtaposed in Tables 7 and 8. The juxtaposition allows for easier viewing of the differences between the two types of correlations. It must be noted that the error for the single-item indicator of self-reported gains was fixed at the same error that was found in the self-reported mastery items. Other possibilities for the results of the confirmatory factor analysis, based on differing levels of error, are found in Appendix C.

Table 7

Attenuated Correlations and Correlations between the Latent Variables for Statistics

	Self-Reported Mastery	Self-Reported Gains	Course Achievement
Self-Reported Mastery	–	.726	.449
Self-Reported Gains	.628	–	.346
Course Achievement	.416	.306	–

Note: Attenuated correlations are on the bottom of the correlation matrix and correlations between the latent variables are on the top.

Table 8

Attenuated Correlations and Correlations between the Latent Variables for Calculus Form A and Calculus Form B

Calculus Form A			
	Self-Reported Mastery	Self-Reported Gains	Course Achievement
Self-Reported Mastery	–	.609	.538
Self-Reported Gains	.473	–	.208
Course Achievement	.436	.160	–
Calculus Form B			
Self-Reported Mastery	–	.609	.365
Self-Reported Gains	.500	–	.276
Course Achievement	.316	.223	–

Note: Attenuated correlations are displayed below the diagonal. Correlations between the latent variables are shown above the diagonal.

Psychometric Properties

The psychometric properties of the various items in terms of traditional item analysis statistics including frequency distributions, point-biserial correlation coefficients, and Cronbach's alpha reliability coefficient were computed. The Cronbach's alpha coefficients for the three final examinations .933 for Statistics, .802 for Calculus Form A, and .781 for Calculus Form B. Point-biserial correlations below .30 were found on 20 Statistic's items, 7 items from Calculus Form A, and 5 items from Calculus Form B. One item out of the Statistics final exam had a point-biserial correlation below .15; item 77 was at .06. In the dichotomous portion of the Calculus Final Exam, item 4 on Form A had a point-biserial correlation of .12, and none of the items on Form B had a point-biserial correlation below .15.

The self-reported mastery items taken by the Calculus students had a Cronbach's alpha of .87 and self-reported mastery items taken by the Statistics students had a Cronbach's alpha of .89. The students' responses to the self-reported mastery items for both the Statistics and Calculus courses were skewed toward the successful side of the scale, but the highest response option ("extremely successful") was consistently lower than the two options right below ("successful" and "extremely successful) (Tables 9 and 10).

Table 9

Distribution of Responses to the Self-Reported Mastery Items by Course and Expected Learning Outcome for Statistics

	Learning Outcome 1		Learning Outcome 2		Learning Outcome 3	
Response	Count	%	Count	%	Count	%
Not at all successful	4	1	4	1	5	1
Not very successful	8	1	22	3	18	2
Moderately successful	81	10	107	14	88	11
Successful	225	29	280	36	259	33
Very successful	306	39	280	36	300	38
Extremely successful	162	21	93	12	116	15

Table 10

Distribution of Responses to the Self-Reported Mastery Items by Course and Expected Learning Outcome for Calculus

	Learning Outcome 1		Learning Outcome 2		Learning Outcome 3	
Response	Count	%	Count	%	Count	%
Not at all successful	2	1	0	0	3	1
Not very successful	10	3	7	2	28	7
Moderately successful	73	18	52	13	73	18
Successful	138	35	113	28	118	30
Very successful	126	32	155	39	123	31
Extremely successful	50	13	72	18	54	14

CHAPTER 5: DISCUSSION

This study supports the conclusion that self-reported mastery is not a perfect proxy for objective measures of course achievement, but serves the purpose of measuring learning outcomes better than self-reported gains. The relationship between self-reported mastery and the objective measures of the learning outcomes were of moderate strength and were stronger than the relationships between self-reported gains and the objective measures of the learning outcomes. Self-reported mastery was also strongly related to self-reported gains, which could indicate that self-reported mastery may have some of the same biases and moderators as self-reported gains.

Correlations

The results from the first research question concerning the correlations between students' (a) self-reported mastery of the expected learning outcomes in a course, (b) self-reported learning gain, and (c) objective measure of their course achievement as measured by score on the final examination helped provide information about the relationships between the self-reported measures and the objective measures.

The self-reported mastery items in this study had attenuated correlation of medium strength with objective measures of learning outcomes (Calculus Form A: r = .436; Calculus Form B: r = .361; Statistics: r = .416). The coefficient of determination showed that these relationships accounted for between 13% and 19% of the variance (Calculus Form A: r^2 = .190; Calculus Form B: r^2 = .130; Statistics: r^2 = .173). The confirmatory factor analysis also supported the same conclusion with slightly stronger relationships (Statistics: r = .449; Calculus Form A: r = .538; Calculus Form B: r = .365).

The results were consistent with most of the previous research. The Sitzmann, Ely, Brown, and Bauer (2010) study reported that the sample weighted mean correlation between self-reported mastery and cognitive learning was .34 (although the adjusted sample weighted mean correlation was higher at .44), very similar to the results from their study. Berdie (1971) found correlations ranging from -.07 to .76. The results from the current study fell in the middle of Berdie's results showing consistency, but also that there may be situations that create both a higher and lower relationship.

Self-reported gains scores were correlated lower with objective measures of learning outcomes than self-reported mastery. The attenuated correlations were consistently weak correlations (Calculus Form A: r = .160, Calculus Form B: r = .223, Statistics: r = .306) which accounted for between 2.6% of the variance and 9.4% of the variance (Calculus Form A: r^2 = .026, Calculus Form B: r^2 = .050, Statistics: r^2 = .094). The confirmatory factor analysis resulted in a slightly higher relationship (Statistics: r = .346, Calculus Form A: r = .208, Calculus Form B: r =.276). The consistently stronger relationship of self-reported mastery with objective measures of learning outcomes may point to self-reported mastery being a better tool for measuring learning outcomes than self-reported gains.

Sitzmann et al. (2010) also compared self-reported mastery items to self-reported gains items and found an even greater difference in the relationship with cognitive measures. Their study found that self-reported gains had a sample weighted mean correlation of .00, while self-reported mastery was at .34 (or .44 when adjusted). The differences in the current study were not as large, but self-reported mastery consistently correlated stronger with objective measures of learning outcomes than the same relationship for self-reported gains. The correlations between self-reported gains and objective measures of the learning outcomes were above zero, which has

also been found in other studies (see Cohen, 1981). The consistency of the stronger relationship between self-reported mastery and objective measures of learning suggests that self-reported mastery will be the more effective choice than the more commonly used tool of self-reported gains.

Even though self-reported mastery appears to be a more effective tool for measuring learning outcomes than self-reported gains, self-reported mastery is not without fault. This may be evident in the fact that self-reported mastery has a stronger relationship with self-reported gains than with objective measures of learning outcomes (Calculus Form A: r = .473, Calculus Form B: r = .500, Statistics: r = .628), which account for between 22.4% and 39.4% of the variance (Calculus Form A: r^2 = .224, Calculus Form B: r^2 = .250, Statistics: r^2 = .394),. The confirmatory factor analysis showed an even stronger relationship (Statistics: r = .726, Calculus Form A: r = .609, Calculus Form B: r =.609). A few factors could be influencing the strength of the relationship between self-reported mastery and self-reported gains: (a) the nature of both types of items was subjective allowing for similar error, (b) self-reported gains may not measure a gain, but actually measure mastery resulting in the same type of item, and/or (c) the similar subject matter may have strengthened the relationship. The strong relationship between self-reported mastery and self-reported gains may suggest that self-reported mastery may have many of the same biases that self-reported gains do. Some influences on self-reported mastery (that have been found to influence self-reported gains) may be social desirability, halo effect, affective influence, differences in context, differences in content, and cultural differences (see literature review above). Self-reported mastery may be a better proxy for measuring learning outcomes than self-reported gains, but users must be aware of the error that may be involved.

Measurement Error Concerns

Researchers have recognized for some time that the presence of measurement error results in an attenuated (or underestimated) estimate of the actual association between variables (Spearman, 1904). In this study we used confirmatory factor analysis to obtain estimates of what the correlations would be without the attenuating influence of measurement error.

The results from this study (see Tables 6 and 7) showed that the differences between the attenuated correlations and the confirmatory factor analysis were the largest for the relationship between self-reported mastery and self-reported gains (differences for Statistics: .098, Calculus Form A: .136, and Calculus Form B: .109). The increase in the strength of the relationship shows that the two self-reported measures may have a closer relationship than expected. The closer relationship may point even more to the fact that self-reported mastery and self-reported gains may share similar biases and moderators.

The other difference between the attenuated correlations and the confirmatory factor analysis that increased by more than .10 was the relationship between self-reported mastery and course achievement for Calculus Form A. (The other differences were about by half that magnitude.) This difference was noteworthy because it increased the difference between the relationships of self-reported mastery and course achievement for Calculus Form A and Calculus Form B (a difference of .173 for the confirmatory factor analysis and a difference of .120 for the attenuated correlations). These two test contained the same items presented in a difference sequence and differences in judgment for the eight constructed-response items, but the results show a continued increase in the difference in their relationship.

Much of the previous research on self-reported measures has been done with attenuated correlations occasionally supplemented with regression without accounting for the measurement

error (Bowman, 2009, 2011b; Herzog, 2011; Gonyea & Miller, 2011; Gosen & Washburn, 1999; Pohlmann & Beggs, 1974; Seifert & Asel, 2011). In very few cases have the authors attempted to account for measurement error (Bowman & Hill, 2011; Pike, 1993). This study found differing amounts of measurement error, specifically depending on the error of each instrument. Future researchers would be wise to recognize the existence of error and compensate for the attenuation in correlations, as lack of correction can distort our perceptions of the relationships between variables.

Psychometric Properties

The results from the third research question concerning the psychometric properties of the various items in terms of traditional item analysis statistics revealed that the objective measures for both Calculus and Statistics had flaws, but were well designed. Reliability, using Cronbach's alpha, was higher for Statistics than Calculus (Statistics: .933, Calculus Form A: .802, Calculus Form B: .781). Point-biserial correlations below .30 were found on 20 Statistic's items, 7 items on Calculus Form A, and 5 items on Calculus Form B. Two items had point-biserial correlation below .15 (one item from Statistics and one item from Calculus Form A). The point-biserial correlations pointed out that the majority of students who were expected to answer the items correctly actually provided a correct response.

Even through not directly explored by the second research question, one of the most interesting findings came from the differences in the correlation coefficients relating to Calculus Form A and Calculus Form B. The confirmatory factor analysis comparing the two groups found that the correlation between self-reported mastery and course achievement was .538 for Form A and .365 for Form B. The difference appears to be large considering the fact that the two forms consisted of the same items. The only differences were that the first twenty items and

the response options were in different orders and judgment on the constructed response options was different. Possibly some of these differences were reflected in the fact that the scores from Form B were less reliable with a Cronbach's alpha of .781, while the reliability of the scores from Form A was at .802. It is interesting that a context effect and differences in judgment could create such differences in the relationships. Future researchers may want to verify if this was an isolated event, or if measurement error can produce such varied results in relationships between objective and subjective forms. The Statistics exam, for example, was the most reliable of the three exams, and had a weaker correlation than Calculus Form A, but stronger than Calculus Form B. If two similar forms resulted in such different relationships, then differences across exams and measurement error could be hugely influencing results.

Error in the self-reported mastery items could also have lowered the correlation. The frequency of the response items for the self-reported mastery items (see Table 8) showed that few students place their level of mastery at the bottom of the scale. It may be that there are few students that actually have low levels of mastery, but students may also be overestimating their ability. Even if students are not overestimating their ability, the lack of range in the response options creates a lower correlation. Resolving some of the error found in the self-reported mastery items may improve the relationships.

Measurement error may have affected the results of previous research. Berdie (1971) reported a wide range of correlations between self-reported mastery items and objective measures of ability. Some of the wide range in the results may have been because his self-reported mastery items did not measure the full range of expertise; some items measured a low level of expertise and some measured a high level of expertise. Covering the whole range of

expertise in one item may have produced stronger relationships between subjective and objective measures.

Recommendations

Administrators or researchers who wish to measure course learning outcomes on a large scale may benefit by using self-reported mastery items. Self-reported mastery items are easier to administer to many students than objective measures and are more comparable across contexts (unless the same objective measurement is administered). Self-reported mastery items may also have a stronger relationship with objective measures of learning outcomes that self-reported gains do. Theoretically, self-reported mastery items make more sense than self-reported gain items in measuring a student's achievement at one point in time. On a smaller scale, objective measures may be more effective that self-reported mastery items because of the many biases that appear to affect the results.

Departments can be delegated the responsibility of creating learning outcomes for the course. Delegating the responsibility to departments could lower the quality of the learning outcomes if departments do not have the level of expertise needed to develop quality learning outcomes. Quality controls and education about how to write effective learning outcomes may be needed to ensure that professors write learning outcomes that are relevant to the course, understandable to students, and can be measured appropriately.

The question of how many learning outcomes to provide may also cause concern. This study allowed professors to write three learning outcomes for their course, which resulted in broad learning outcomes. Administrators of student ratings forms may want to consider that allowing fewer questions will cause learning outcomes to be broad, but too many learning outcomes may be difficult for some departments to complete and place undue stress on

completion of the form for students. This study was unable to provide information about how self-reported mastery items with more specific learning outcomes may relate to the specific objective measures of content relating to that learning outcome. Further study could clarify that the items distinctly measure each outcome and whether the specificity strengthens the relationship between the objective and subjective measures.

Researchers and administrators may be able to use results from self-reported mastery items to compare different teachers teaching the same course, or the effectiveness of the same teacher over time. It may be tempting to compare results of self-reported mastery items across a full department or university, but each learning outcome would have completely different language resulting in different functioning. Comparisons across different learning outcomes would be unadvisable.

Administrators and researchers may also consider correcting for bias and moderators to reduce error and strengthen the relationship between self-reported mastery and objective measures of learning outcomes. Controlling for high school self-reported gains may help to reduce some bias (Seifert & Asel, 2011). It may also be possible to control for moderators like differences across universities (Bowman, 2011b; Pike, 1996), types of content (Pike, 1995; Sitzmann et al. 2010), year in school (Bowman, 2011b; Bowman & Hill, 2011; Gonyea & Miller, 2011), cultural background (Douglass, Thomson, and Zhao, 2012), content correspondence (Pike, 1995; Sitzmann et al., 2010), and amount of feedback (Sitzmann et al., 2010). Instead of controlling for the amount of feedback in a class, teachers can also be encouraged to provide more feedback to students. The increased amount of feedback can increase the strength of the correlations between self-reported mastery and objective measures of learning outcomes.

Administrators and researchers may want to consider when to administer the self-reported mastery items. Traditionally, student ratings are administered before students take their final exams (as in this study), but self-reported mastery items may have stronger relationships with objective measures of learning if given later in the semester. Support for this recommendation came from the findings of a self-assessment study, and findings from the literature review.

Grimes (2002) found that students predicting their exam scores (in essence a self-assessment of their predicted score) lowered the amount they overestimated their grades depending on the time of prediction; two days before the exam, the students predicted they would score 85 percent, which dropped to 83 percent right before the test, and then to 77 percent immediately after the test. Inviting student to take the self-reported mastery items after they completed their final examination could improve the strength of the relationship between self-reported mastery items and course achievement. Considering that prediction of exam scores is not equal to self-reported mastery, the concept may or may not hold when measuring learning outcomes. Further study may be needed to verify this prediction.

Two findings reviewed in the literature may also support the recommendation to administer student ratings after the final exam. Sitzmann et al. (2010) found that feedback was a moderator for the relationship between self-reported mastery and objective measures. Giving students more time to study (providing self-feedback) and receiving feedback from the final examination may allow students to more accurately gauge their mastery of the learning outcomes. In relation to the concept of feedback influencing learning, the authors suggested that further research needed to look at how learning over time changes the relationship between self-reported measures and objective measures. More time to provide feedback for students may

increase the strength of the self-reported mastery items. Even though asking students to fill out student ratings after the exams may strengthen the relationship between self-reported mastery and objective measures, administrators and researchers may need to evaluate the effect timing has on response rate.

Second, Berdie (1971) suggested that students' knowledge of the subject influenced the strength of the relationship—or that ignorance caused a low correlation. Increased studying for the final exam and the recognition and recall the final exam would provide may create an increase in knowledge and lower ignorance. This would support offering student ratings at a later date as long as knowledge and ignorance actually have an effect on self-reported mastery of learning outcomes.

One might argue against administering student ratings after the exam because the exam score might influence students' self-reports. Students may be influenced by exam scores, but this scenario assumes that the score would bias students (creating satisficing) instead of influencing them as a form of feedback. Further research may be needed to understand the influence of examinations on the beliefs of students in relation to this scenario.

Administrators and researchers may also be interested in learning and not just the mastery of learning outcomes. In this case, they may consider pairing a retrospective question about the students' initial level of mastery of the learning outcome with the self-reported mastery item. If administrators or researchers decide to use this approach, they must also realize that a retrospective pretest-posttest will also have many of the same flaws that other self-reported assessments have. Douglass, Thomson, and Zhao (2012), for example, found issues with cultural bias and noted that previous studies have found that participants overestimate their answers.

Strengths

This study furthers the understanding of self-reported mastery, which has been documented little in the educational literature. This fills a growing need as learning outcomes become more popular and the demand to measure them becomes greater.

This study may also help to point out the ineffectiveness of self-reported gains, which are commonly used in student ratings of instruction. If self-assessment methods measuring gains are needed, future researchers may want to explore other methods like the retrospective pretest-posttest method, which utilizes self-reported mastery, to evaluate if these methods are more effective than self-reported gains. Caution would also be needed with these methods because of the many biases that have been found connected to self-assessment methods.

This study took another step toward evaluating how to measure specific learning outcomes. Future research will benefit by continuing to look at specific learning outcomes and researching how to measure learning outcomes more effectively.

Weaknesses

This study was not without weaknesses, four of them being the lack of specificity of the learning outcomes, the lack of research into the biases that may plague self-reported mastery items, the use of a single-item indicator, and the possibility that the results may not generalize to other subject areas.

The statements of expected learning outcomes created by the professors (the self-reported mastery items) were very broad. This may have been because the professors were asked submit a maximum of three learning outcomes. Because the learning outcomes were broad, the objective measures for the course were also broad (the complete final exam). Future research may benefit by evaluating more specific learning outcomes and more specific objective

measures. More specificity may change the strength of the relationship, and may also allow for more flexibility in examining context issues and other biases.

This study also did little to look at the biases that may plague self-reported mastery. Some of the possible biases that may affect self-reported mastery were reported above in the literature review. The only error explored in this study was measurement error, and even the topic of measurement error could be explored more fully in the future.

This study used a single-item indicator to represent self-reported gains. Measurement experts have warned against using single-item indicators (Diamantopoulos, Sarstedt, Fuchs, Wilezynski, & Kaiser, 2012; Fuchs & Diamantopoulos, 2009) and suggest using multiple items in most cases. It may be noted that the single-item indicator in this study did not explore a complex construct. Learning is a complex construct, but studies have reported that self-reported gains are an affective measure (Bowman & Hill, 2011; Gonyea & Miller, 2011; Pike, 1993; Pohlmann & Beggs, 1974; Sitzmann et al., 2010), which results in a much less complex construct. Even though the single-item indicator may be somewhat justified, future research would be wise to use multiple items to measure the construct of self-reported gains.

The results of this study may not generalize to other subject areas. Both Pike (1995) and Sitzmann et al. (2010) found differences across content in the relationship between subjective and objective measures. Pike (1995) compared objective and subjective measures in four different undergraduate subject areas and found that math had the strongest relationship, followed by English, science, and then social studies. The current study included mathematically related classes (Calculus and Statistics) and did not consider courses in other subject areas. Pike also pointed out that content correspondence (whether the subjective and objective items measure the same concepts) may explain some of the across content differences. Aligning

subjective and objective items may be more difficult in less concrete subjects, but this may not mean that subjects in the humanities will have lower relationships between subjective and objective measures. Pike's study, for example, found that English had a stronger relationship between objective and subjective measures than science. Future studies may want to assure that both subjective and objective measures assess equivalent content to verify whether there are consistent differences in relationships across content.

References

Astin, A. W. (1993). What matters in college? *Liberal Education*, *79*(4), 4–15.

Baird, L. L. (1988). Value-added: Using student gains as yardsticks of learning. In C. Adelman (Ed.), *Performance and Judgment: Essays on Principles and Practice in the Assessment of College Student Learning* (p. 205-216). Washington, D.C.: U.S. Government Printing Office.

Berdie, R. F. (1971). Self-claimed and tested knowledge. *Educational and Psychological Measurement*, *31*, 629–636.

Bowman, N. A. (2009). Can 1st-year college students accurately report their learning and development? *American Educational Research Journal*, *47*, 466–496.

Bowman, N. A. (2011a). Examining systematic errors in predictors. *New Directions for Institutional Research*, *150*, 7–20.

Bowman, N. A. (2011b). Validity of college self-reported gains at diverse institutions. *Educational Researcher*, *40*, 22–24. doi:10.3102/0013189X10397630

Bowman, N. A., & Hill, P. L. (2011). Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains, *New Directions for Institutional Research, 150*, 73–86.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Educational Research*, *51*, 281–309.

Créde, M., Harms, P, Niehorster, S, & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the big five personality traits. *Journal of Personality and Social Psychology, 102*, 874-888.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilezynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science, 40*, 434-449.

Douglass, J. A., Thomson, G., & Zhao, C.-M. (2012). The learning outcomes race: The value of self-reported gains in large research universities. *Higher Education*. *64*, 317-335.

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft, 69*, 195-210.

Ginns, P., & Barrie, S. (2004). Reliability of single-item ratings of quality in higher education: A replication. *Psychological Reports, 95*, 1023-1030.

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. In P. D. Umbach (Ed.), *Survey research: Emerging issues* (New Directions for Institutional Research, No. 127, pp. 73–89). San Francisco, CA: Jossey-Bass.

Gonyea, R. M., & Miller, A. (2011). Clearing the AIR about the use of self-reported gains in institutional research. *New Directions for Institutional Research*, *150*, 99–112.

Gosen, J., & Washbush, J. (1999). Perceptions of learning in TE simulations. *Developments in Business Simulation & Experiential Learning*, *26*, 170–175.

Grimes, P. W. (2002). The overconfident principles of economics student: An examination of a metacognitive skill. *Journal of Economic Education*, *33*, 15–30.

Herzog, S. (2011). Gauging academic growth of bachelor degree recipients: Longitudinal vs. self-reported gains in general education. *New Directions for Institutional Research*, *150*, 21–40.

Hornsey, M. J., Olsen, S., Barlow, F. K., & Oei, T. P. (2012). Testing a single-item visual analogue scale as a proxy for cohesiveness in group psychotherapy. *Group Dynamics: Theory, Research, and Practice, 16*, 80-90.

Jordan, J. S., & Turner, B. A. (2008). The feasibility of single-item measures for organizational justice. *Measurement in Physical Education and Exercise Science, 12*, 237-257.

Kennedy, E. J., Lawton, L., & Plumlee, E. L. (2002). Blissful ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education*, *24*, 243–252.

Núñez-Peña, M. I., Guilera, G., & Suárez-Pellicioni, M. (2014). The single-item math anxiety scale: An alternative way of measuring mathematical anxiety. *Journal of Psychoeducational Assessment, 32*, 306-317.

Pascarella, E. T. (2001). Using student self-reported gains to estimate college impact. *Journal of College Student Development*, *42*, 488–492.

Pike, G. R. (1993). The relationship between perceived learning and satisfaction with college: An alternative view. *Research in Higher Education*, *34*, 23–40.

Pike, G. R. (1995). The relationship reports of college experiences and achievement test scores. *Research in Higher Education*, *36*, 1–21.

Pike, G. R. (1996). Limitations of using students' self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education*, *37*, 89–114.

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education*, *40*, 61–86.

Pohlmann, J. T., & Beggs, D. L. (1974). A study of the validity of self-reported measures of academic growth. *Journal of Educational Measurement*, *11*, 115–119.

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology, 52*, 597-617.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*, 493-502.

Schriesheim, C. A., Hinkin, T. R., & Podsakoff, P. M. (1991). Can ipsative and single-item measures produce erroneous results in field studies of French and Raven's (1959) five bases of power? An empirical investigation. *Journal of Applied Psychology, 76*, 106-114.

Seifert, T. A., & Asel, A. M. (2011). The tie that binds: The role of self-reported high school gains in self-reported college gains. *New Directions for Institutional Research*, *150*, 59–72.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, *9*, 169–191.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72-101.

Spellings' Commission on the Future of Higher Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. US Department of Education, September 26, 2006.

Spörrle, M. & Bekk, M. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment, 21*, 272-285.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82*, 247-252.

Zimmerman, D. W. (2009). The reliability of difference scores in populations and samples. *Journal of Educational Measurement*, *46*, 19–42.

Appendix A

Proposed New Student Rating Form—Fall 2012 Pilot

Student Questionnaire

Student feedback on faculty and courses is very important at BYU. This feedback is used by faculty to improve their teaching. Department chairs review your ratings as one of several pieces of information to assess teaching effectiveness. And University committees consider student feedback carefully in determining who is retained and who is promoted. Your responsible input is essential to assessing and improving teaching performance and student learning at BYU. **Be honest, fair, and constructive as you complete this questionnaire.**

Expected Learning Outcomes

Three of the expected learning outcomes for your course are listed below. In your judgment, how successfully have you achieved these outcomes:

1. Understand the importance of data collection and how it dictates the appropriate statistical method and acceptable inference.

Not at all successful	Not very successful	Moderately successful	Successful	Very successful	Extremely successful
●	●	●	●	●	●

Comments:

2. Understand and communicate using technical language about probability and variation.

Not at all successful	Not very successful	Moderately successful	Successful	Very successful	Extremely successful
●	●	●		●	●

Comments:

3. Interpret and communicate the outcomes of estimation and hypothesis tests in the context of a problem.

Not at all successful	Not very successful	Moderately successful	Successful	Very successful	Extremely successful
●	●	●		●	●

Comments:

Answer the following questions. Although it is easy to give the same response to all of the questions, consider your response to each question separately.

4. To what extent was the instructor (not the TA) willing to help students when they needed it?

Not at all willing	Not very willing	Moderately willing	Willing	Very willing	Extremely willing
●	●	●	●	●	●

Comments:

5. How effective was the instructor (not the TA) in providing meaningful opportunities and encouragement for you to actively participate in the learning process?

Not at all effective	Not very effective	Moderately effective	Effective	Very effective	Extremely effective
●	●	●	●	●	●

Comments:

6. How effective was the instructor (not the TA) in teaching challenging concepts or skills?

Not at all effective	Not very effective	Moderately effective	Effective	Moderately effective	Very effective
●	●	●	●	●	●

Comments:

7. For this course, about how many hours per week did you spend **out of class** (doing assignments, readings, etc.)?
(e.g. 4, 4.5)

What effect did this course and instructor have in helping you achieve the Aims of a BYU Education?

8. Spiritually Strengthening:

Detracted	No effect	Slightly enhanced	Moderately enhanced	Strongly enhanced	Very strongly enhanced
●	●	●	●	●	●

Comments:

9. Intellectually Enlarging:

Detracted	No effect	Slightly enhanced	Moderately enhanced	Strongly enhanced	Very strongly enhanced
●	●	●	●	●	●

Comments:

10. Character Building:

Detracted	No effect	Slightly enhanced	Moderately enhanced	Strongly enhanced	Very strongly enhanced
●	●	●	●	●	●

Comments:

11. Leading to Lifelong Learning and Service:

Detracted	No effect	Slightly enhanced	Moderately enhanced	Strongly enhanced	Very strongly enhanced
●	●	●	●	●	●

Comments:

Appendix B

Brigham Young University Student Ratings Form (Old Form)

Comparing this course with other university courses you have taken, please indicate an OVERALL rating for the following:

Course: DIET 123

Exceptionally Poor	Very Poor	Poor	Somewhat Poor	Somewhat Good	Good	Very Good	Exceptionally Good
●	●	●	●	●	●	●	●

Instructor: Smith, John Q

Exceptionally Poor	Very Poor	Poor	Somewhat Poor	Somewhat Good	Good	Very Good	Exceptionally Good
●	●	●	●	●	●	●	●

Please respond to each of the following items regarding this course: DIET 123

I learned a great deal in this course.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Course materials and learning activities were effective in helping students learn.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

This course was well organized.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Evaluations of students' work (e.g., exams, graded assignments and activities) were good measures of what students learned in the course.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Course grading procedures were fair.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree

●	●	●	●	●	●	●	●

This course helped me develop intellectual skills (such as critical thinking, analytical reasoning, integration of knowledge).

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

This course provided knowledge and experiences that helped strengthen my testimony of the Gospel of Jesus Christ.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

For this course, about how many hours per week did you spend **in class**?
(e.g. 2, 2.5)

What percentage of the time you spent in class was valuable to your learning?

%

For this course, about how many hours per week did you spend **out of class** (doing assignments, readings, etc.)? (e.g. 4, 4.5)

What percentage of the time you spent out of class was valuable to your learning (as opposed to just busy work)?

%

Please respond to each of the following statements regarding this instructor: Smith, John Q

The instructor:

Showed genuine interest in students and their learning.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Provided adequate opportunities for students to get help when they needed it.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Provided opportunities for students to become actively involved in the learning process.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Gave students prompt feedback on their work.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Provided students useful feedback on their work.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Responded respectfully to students' questions and viewpoints.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree

●	●	●	●	●	●	●	●

Was effective in explaining difficult concepts and ideas.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Appropriately brought Gospel insights and values into secular subjects.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Was spiritually inspiring insofar as the subject matter permitted.

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

This instructor and course contributed to the Mission and Aims of a BYU Education (i.e., Spiritually Strengthening, Intellectually Enlarging, Character Building, Leading to Lifelong Learning and Service).

Very Strongly Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Very Strongly Agree
●	●	●	●	●	●	●	●

Please add any comments or suggestions you have about your learning experience in this course with this instructor.

Appendix C

Correlations between the Latent Variables in the Confirmatory Factor Analysis of the Statistics and Calculus Tests by Varying Amounts of Error by Variance

Statistics

Self-Reported Gains Error by Variance	Self-Reported Mastery with Self-Reported Gains	Self-Reported Mastery with Course Achievement	Self-Reported Gains with Course Achievement
.1503	.687	.449	.327
.3006	.728	.449	.347
.4509	.779	.449	.371
.6012	.841	.449	.401
.7515	.922	.449	.440

Calculus

Self-Reported Gains Error by Variance	Self-Reported Mastery with Self-Reported Gains	Self-Reported Mastery with Course Achievement	Self-Reported Gains with Course Achievement
.1932	.544	.454	.216
.3864	.577	.454	.230
.5796	.617	.454	.245
.7728	.667	.454	.265
.7515	.716	.454	.285