


Spring 2017

Retrospective versus prospective measurement of examinee motivation in low-stakes testing contexts: A moderated mediation model

Aaron J. Myers
James Madison University

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

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

Recommended Citation

Myers, Aaron J., "Retrospective versus prospective measurement of examinee motivation in low-stakes testing contexts: A moderated mediation model" (2017). *Masters Theses*. 511.
<https://commons.lib.jmu.edu/master201019/511>

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

Retrospective versus Prospective Measurement of Examinee Motivation in Low-Stakes
Testing Contexts: A Moderated Mediation Model

Aaron James Myers

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2017

FACULTY COMMITTEE:

Committee Chair: Sara J. Finney

Committee Members/Readers:

John D. Hathcoat

Kenneth E. Barron

Acknowledgements

I would first like to thank my academic advisor and thesis chair, Dr. Sara Finney. I can't thank you enough for your dedication and support. You have truly inspired me to be a better student, researcher, and professional. I could not ask for a better advisor. I would also like to thank the rest of my thesis committee, Dr. John Hathcoat and Dr. Kenn Barron. Thank you for all of your thoughtful feedback and suggestions throughout the thesis process. Third, I would like to thank the rest of the faculty and students at the Center for Assessment and Research Studies. I have learned so much from each and every one of you. Fourth, I would like to thank Courtney Sanders, Scott Strickman, and Thai Ong for all of your support and encouragement. You have been great friends that have always been there for me. Last, but certainly not least, thank you to my amazing wife, Jamie. Thank you for putting up with my late nights, encouraging me to persist, and helping me get through it all. I can't thank you enough.

Table of Contents

Acknowledgments.....	ii
List of Tables	vi
List of Figures	viii
Abstract.....	ix
Chapter I: Introduction.....	1
Measurement of Examinee Effort: Background of Problem Addressed by Study ..	2
Statement of the Problem Addressed by Current Study	8
Significance of the Current Study.....	10
Purpose of the Current Study	12
Research Questions	13
Comparing retrospective motivation scores across conditions.	13
Comparing prospective and retrospective motivation scores across conditions.....	15
Changes in importance and effort scores.....	18
Test performance differences: Research question 7.	19
Chapter II: Literature Review	21
Mediation	21
Importance of Mediation.....	23
Complete and partial mediation.....	25
Third Variables	30
Suppressor.....	30
Covariate	33
Confounder	34
Moderator.....	35
Differences between third variable hypotheses and correlation patterns...37	
Summary	39
Statistical Estimation to Support Mediation	40
Causal-steps approach.....	40
Structural equation modeling model-data fit approach.....	42
Tests of the indirect effect.....	45
Research Design Characteristics to Differentiate Third Variables	53
Experimental design.....	53
Classic experimental design.....	54
Causal-chain design	54
Temporal precedence	55
Summary	57
Assumptions of Mediation.....	57
Measurement error	57
Model misspecification.....	58

Current Measurement of Motivation	62
Retrospective and Prospective Measurement	63
Retrospective measurement	63
Prospective measurement.....	64
Type of Measurement Influences Item Response Patterns	64
Attributional bias	64
Behavioral commitment.....	66
Experience limitation	68
Implications of retrospective versus prospective measurement.....	70
Chapter III: Methods.....	72
Participants.....	72
Measures	74
Quantitative and scientific reasoning.....	74
Test importance and examinee effort.....	74
Procedures	76
Retrospective condition.	76
Combined condition.....	76
Prospective condition.....	77
Analytic Approach	77
Model-data fit.....	77
Nested model comparisons.	79
Research questions 1 and 3.	79
Research questions 2 and 4.	83
Research question 5.	85
Research question 6.	87
Research question 7.	88
Chapter IV: Results.....	90
Retrospective Importance and Effort Analyses	90
Research question 1.	90
Research question 2.	97
Summary of Retrospective Importance and Effort Analyses.	101
Prospective and Retrospective Importance and Effort Analyses	102
Research question 3.	102
Research question 4.	108
Summary of Prospective and Retrospective Importance and Effort Analyses.....	113
Changes in Test Importance and Effort Scores.....	113
Research question 5.	113
Research question 6.	117
Test Performance Differences: Research question 7.	118

Chapter V: Discussion	120
No Impact of Behavioral Commitment on Retrospective Importance and Effort	120
Measurement invariance and no average differences across conditions.	120
No difference in indirect effect across conditions.	122
Differences in Prospectively and Retrospectively Measured Effort.....	123
Measurement invariance and differences in effort across conditions.	123
Smaller standardized indirect effect modeling prospective importance and effort.....	125
Changes in Effort Scores Due to Experiencing the Test.....	127
Lower retrospective effort than prospective effort.	127
No effect of test performance on change in importance or effort.	128
Small Difference in Test Performance across Conditions	130
Summary of Findings.....	131
Limitations and Future Studies	132
Conclusion	137
Appendix A.....	173
Appendix B.....	174
Appendix C.....	175
Appendix D.....	176
Appendix E.....	177
Appendix F.....	179
Appendix G.....	180
References.....	182

List of Tables

Table 1: Third Variable Functions, Example Hypotheses, Path Diagrams, and Examples of Data Characteristics 138

Table 2: Research Questions, Hypotheses, Data Analyses, and Measures..... 140

Table 3: Demographics by Measurement Condition 144

Table 4: Correlations and Descriptive Statistics for Quantitative and Scientific Reasoning and Retrospective and Prospective Perceived Test Importance and Examinee Effort Scores by Measurement Condition..... 145

Table 5: Correlations and Descriptive Statistics for Retrospective SOS by Measurement Condition..... 146

Table 6: Model Fit Indices for One-Factor, Two-Factor, and Multiple-Condition Invariance Testing of the Retrospective SOS Scores across Measurement Conditions147

Table 7: Unstandardized and (Standardized) Factor Pattern Coefficients for the Retrospective SOS Scores across Measurement Conditions.....149

Table 8: Fit Indices for Completely Mediated Model using Retrospective SOS Scores by Measurement Condition and Invariance Testing.....150

Table 9: Correlations and Descriptive Statistics for Combined SOS by Measurement Condition.....151

Table 10: Model Fit Indices for One-Factor, Two-Factor, and Multiple-Condition Invariance Testing of the Prospective and Retrospective SOS Scores across Measurement Conditions.....152

Table 11: Unstandardized and Standardized Factor Pattern Coefficients for Prospective and Retrospective SOS Scores across Measurement Conditions.....154

Table 12: Fit Indices for Completely Mediated Model using Prospective and Retrospective SOS Scores by Measurement Condition and Invariance Testing	155
Table 13: Correlations and Descriptive Statistics for Prospective and Retrospective Importance and Effort Scores from the Prospective Condition	156
Table 14: Longitudinal Invariance Testing of Retrospective and Prospective Importance and Effort Scores from the Prospective Condition.....	157
Table 15: Unstandardized and (Standardized) Factor Pattern Coefficients from Longitudinal Analysis of Prospective and Retrospective SOS Scores from the Prospective Measurement Condition	158

List of Figures

Figure 1: Illustration of EV theory, EV theory applied to low-stakes testing, and EV theory as conceptualized in the current study	159
Figure 2: Illustration of measurement conditions with example items.....	160
Figure 3: Illustration of the relationship between political media and political engagement	161
Figure 4: Illustration of parameter symbols and regression formulas for partial and complete mediation	162
Figure 5: Illustration of complete and partial mediation	163
Figure 6: Illustration of ordinal and disordinal moderation effects	164
Figure 7: Illustration of classical and negative suppression models.....	165
Figure 8: Comparison of complete mediation and complete confounding models	166
Figure 9: Completed mediated models modeling retrospective SOS scores and combined prospective/retrospective SOS scores	167
Figure 10: Conditional LGM: Predicting change in importance and effort from test scores	169
Figure 11: Plot of importance change scores as a function of test performance	171
Figure 12: Plot of effort change scores as a function of test performance	172

Abstract

Expectancy-value theory applied to examinee motivation suggests examinees' perceived value of a test indirectly affects test performance via examinee effort. This empirically supported indirect effect, however, is often modeled using importance and effort scores measured *after* test completion, which does not align with their theoretically specified temporal order. Retrospectively measured importance and effort scores may be influenced by examinees' test performance, impacting the estimate of the indirect effect. To investigate the effect of timing of measurement, first-year college students were randomly assigned to one of three conditions where (1) importance and effort were measured retrospectively; (2) importance was measured prospectively; and (3) importance and effort were measured prospectively. Additionally, importance and effort were measured retrospectively in conditions two and three to determine if the rank-order and average importance and effort scores change from before to after the test. The indirect effect was invariant across conditions, indicating no effect of behavioral commitment on retrospective importance and effort scores. Although the unstandardized indirect effect using prospective and retrospective importance and effort scores was invariant across conditions, the standardized indirect effect was smaller in the prospective condition. Thus, testing practitioners should be cautious when interpreting and making decisions based on the standardized indirect effect. Average latent retrospective effort scores were lower than prospective effort scores via cross-sectional analyses. Moreover, examinees completing both importance and effort measures reported lower retrospective effort than prospective effort. This decrease in effort was not related to test performance, indicating change in effort is not influenced by test performance. However, given

examinees tend to decrease ratings of effort after experiencing the test, results from techniques such as motivation filtering will differ depending on when motivation is measured. Importantly, examinees primed by engaging in prospective ratings of motivation had higher average levels of test performance than examinees that did not provide prospective ratings. Although the increase in test performance was small, priming may provide a cheap intervention aimed at increasing test performance.

Chapter I

Introduction

Examinees assessed in low-stakes testing contexts have nothing to gain or lose as a result of their test performance. Accordingly, the effort put forth by examinees on a test with no personal consequences may be low or may be high. In contrast, examinees in high-stakes testing contexts such as certification and admission have something to gain or lose as a result of their performance; thus, they tend to put forth enough effort to demonstrate their ability. Discrepancy in expended effort in low- versus high-stakes testing contexts suggests inferences made from test scores across these testing contexts may not be of equal validity (Cole & Osterlind, 2008; Sundre & Kitsantas, 2004; Sungur, 2007; Wise & Smith, 2016; Wolf & Smith, 1995).

The prevalence of low-stakes testing internationally (e.g., institutional accountability assessment, National Assessment on Educational Progress, Program for International Student Assessment, Trends in International Mathematics and Science Study) has prompted research investigating examinee behavior and affect during low-stakes testing. Researchers have empirically established that test scores are a function of both examinee ability and expended effort (O'Neil, Sugrue, & Baker, 1995/1996; Wise, 2006; Wise & DeMars, 2005; Wise & Kong, 2005; Wolf & Smith, 1995). More specifically, examinee effort impacts the validity of inferences from test scores (e.g., Wise & Smith, 2016), value-added scores (e.g., Finney, Sundre, Swain, & Williams, 2016), and international comparisons of ability (e.g., Eklöf, Pavešič, & Grønmo, 2014). Consequently, professional organizations that focus on testing and measurement issues stipulate that practitioners collect examinee effort data and interpret test scores in light of

examinee effort. The *Standards for Educational and Psychological Testing* assert, “In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation” (AERA, APA, & NCME, 2014, p. 213). Given the relevance and attention to examinee effort in low-stakes testing contexts, it is not only important to measure effort, but to investigate *how* and *when* to measure effort.

Measurement of Examinee Effort: Background of Problem Addressed by Study

A common method of measuring examinee motivation in low-stakes testing contexts is via examinee self-report (Wise & Smith, 2016).¹ A widely used self-report measure of examinee motivation is the Student Opinion Scale (SOS; Finney, Mathers, & Myers, 2016; Thelk, Sundre, Horst, & Finney, 2009). Based on expectancy-value (EV) theory (Eccles et al., 1983), the SOS consists of two subscales: perceived test importance (e.g., “Doing well on this test was important to me”) and expended effort (e.g., “I gave my best effort on this test”). The importance and effort subscales are operationalizations of the task value and motivation components of EV theory, respectively.

¹ Response time effort (RTE) is another method employed to measure effort (Wise & Kong, 2005). RTE is operationalized as the time interval between visualization of an item and selection of a response to the item. Conceptually, examinees either expend the effort necessary to select the most appropriate response, labeled solution behavior, or do not expend the effort necessary and randomly select a response, labeled rapid-guessing behavior. For example, if an examinee responds to an item in less time than possible, given the length of time necessary to read and respond to the item, then the examinee’s response would be classified as rapid-guessing behavior. Examinee RTE is then calculated by the ratio of items where the examinee exhibited solution behavior to the total number of items. A disadvantage to RTE is it presumably only identifies examinees expending the lowest levels of effort. Because item responses are scored one for solution behavior and zero for rapid-guessing behavior, it is likely that little effort is expended on some item responses scored as solution behavior. For example, if the time an examinee takes to respond to an item qualifies as solution behavior, but the examinee truly exerts little effort, the behavior will be incorrectly recorded as solution behavior. In short, variability in effort is artificially reduced as a result of the dichotomously scored response-level behavior and, in turn, incorrect classifications of behavior can result (DeMars, Bashkov, & Socha, 2013).

Expectancy-value theory. Motivation, as it pertains to testing, refers to the persistence, or expended effort, an examinee puts forth toward the task of completing a test. As originally conceptualized, EV theory posits that motivation is primarily a function of the interaction between expectancy of success on a task and the value of that task (Atkinson, 1957). An important characteristic of Atkinson's conceptualization of EV theory is it applies to tasks that are prescribed, where the individual has no choice whether to engage in the task or not, but only has a choice as to the level of engagement in the task. For this reason, EV theory seems especially applicable to low-stakes testing contexts where examinees only choose the level of effort to expend given their expectancy and value associated with the test.

A revision of Atkinson's EV theory extended and further developed the expectancy and value components and their respective interrelationships (Eccles et al., 1983; Wigfield & Eccles, 1992, 2000, 2002). Expectancy of success on a task refers to an individual's judgment of the likelihood of personal success on the task (Eccles et al., 1983). Implicit within the expectancy component are perceptions of personal competence, locus of control, and task difficulty. That is, individuals with higher levels of expectancy of success likely perceive themselves to be relatively more competent and more in control, and perceive the task to be less difficult (i.e., more attainable). Eccles and colleagues conceptualized EV theory's value component as four distinct elements: attainment, intrinsic, utility, and cost. Attainment value is defined by the importance of doing well on a task. Intrinsic value is defined as the enjoyment one gets from doing the task. Utility value refers to how accomplishing the task may be useful to the individual. Cost refers to the negative aspects of engaging in a task such as time devoted to the task

and cognitive and affective demand. Whereas higher levels of attainment, intrinsic, and utility value are associated with higher levels of motivation, cost is inversely related to motivation (Eccles et al., 1983; Wigfield & Eccles, 1992). An important departure from Atkinson's EV theory was that Eccles and colleagues made no mention of the interaction between the value and expectancy components. Further, they suggested a positive, instead of negative, relationship between expectancy and value.

The theoretically suggested pattern of interrelationships between expectancy, value, motivation, and performance is consistent with mediation. Expectancy of success on a task *and* task value are theorized to affect task performance via an individual's motivation (see Figure 1). Empirically, research in the domain of academic achievement has demonstrated the indirect effects of expectancy of success and task value on performance via motivation (e.g., Plante, O'Keefe, & Théorêt, 2013).

EV theory applied to examinee motivation. Adapting EV theory to low-stakes testing contexts, researchers have operationalized task value as perceived test importance and motivation as examinee effort. Accordingly, researchers posit that perceived test importance affects examinee effort which affects test performance (Wise & DeMars, 2005; Wolf & Smith, 1995). When a test is of little consequence to examinees, they may perceive the test to be of little importance (i.e., attainment value), perceive little enjoyment from the test (i.e., intrinsic value), perceive the test to be useless (i.e., utility value), and realize the cognitive and affective demand necessary to complete the test (i.e., cost; Cole, Bergin, & Whittaker, 2008; Wise & DeMars, 2005). An examinee attributing little value to the test will not be inclined to expend the effort necessary to demonstrate true ability, resulting in an attenuated ability estimate (see Figure 1).

The aforementioned example did not mention expectancy of success. Expectancy of success is not often modeled in empirical studies of motivation in low-stakes testing contexts. One argument for focusing on perceived value and not expectancy during low-stakes testing centers on examinees' conceptualization of success. Expectancy refers to examinees' perceived likelihood of success on a test; however, the definition of success on low-stakes tests is unclear. Examinees do not typically receive test scores, let alone interpretive feedback about their performance; thus, it is difficult to comprehend what success means to examinees in low-stakes testing contexts (Cole et al., 2008). Moreover, given perceptions of competence and task difficulty are subsumed within expectancy of success, it may be difficult for examinees to consider these characteristics when they have no experience with the test. Examinees completing a high-stakes test are aware of the content and difficulty of the test and have prepared appropriately. Examinees completing a low-stakes test are often unaware of the content and difficulty and have not explicitly prepared for the test. Accordingly, it may be difficult for examinees to form accurate perceptions of expectancy of success on low-stakes tests.

Another reason expectancy is not often measured or modeled may be due to expectancy not being easily manipulated by testing practitioners, whereas task value could potentially be manipulated. For example, to increase the validity of low-stakes test scores, researchers and test practitioners have investigated methods of increasing examinee effort by trying to manipulate perceived test importance via financial incentives (Baumert & Demmrich, 2001; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005) and via test instructions (Baumert & Demmrich, 2001; Bensley et al., 2016; Finney, Sundre et al., 2016; Hawthorne, Bol, Pribesh, & Suh, 2015; Kornhauser, Minahan, Siedlecki, &

Steedle, 2014; Liu, Bridgeman, & Adler, 2012; Liu, Rios, & Borden, 2015; Mathers, Finney, & Myers, 2016; Myers, Finney, & Mathers, 2016; Wise, 2004). In contrast, it would be difficult for researchers or test practitioners to manipulate examinees' expectancy of success on a test within the confines of a testing session. Expectancy will not be evaluated in the current study. The indirect effect of perceived importance on test performance via examinee effort will be the focus of the current study.

Researchers using cross-sectional research designs have empirically demonstrated the indirect effect of perceived importance on test performance via expended effort (e.g., Cole et al., 2008). Moreover, this indirect effect has generalized across age groups and countries: ninth-grade students in Germany (e.g., Penk & Schipolowski, 2015), first-year college students in the U.S. (e.g., Mathers et al., 2016; Myers et al., 2016; Zilberberg et al., 2014), and upper-class college students in the U.S. (e.g., Cole et al., 2008; Mathers et al., 2016; Myers et al., 2016). In addition, the indirect effect has been demonstrated when examinees complete tests of different knowledge domains, such as quantitative and scientific reasoning (Mathers et al., 2016; Myers et al., 2016; Penk & Schipolowski, 2015; Zilberberg et al., 2014), mathematics, science, language, and social studies (Cole et al., 2008). The indirect effect has been demonstrated across test instruction conditions (i.e., examinees told: their test scores would be used for accountability mandates, they would receive performance feedback, or faculty would see their personal test scores) intended to increase the personal relevance of the test to the examinee (Mathers et al., 2016; Myers et al., 2016). Finally, the indirect effect has been demonstrated when controlling for gender and prior knowledge (Cole et al., 2008; Mathers et al., 2016;

Zilberberg et al., 2014), conscientiousness (Myers et al., 2016), test anxiety (Mathers et al., 2016; Penk & Schipolowski, 2015), and expectancy (Penk & Schipolowski, 2015).

Given the number of studies demonstrating the indirect effect of perceived importance on test performance via examinee effort, it is important to consider the practical significance of the indirect effect. The standardized indirect effect is the product of the standardized direct effect of importance on effort and the standardized direct effect of effort on performance. Crude guidelines are .01, .09, and .25 standard deviations for small, medium, and large effect sizes, respectively (Kenny, 2016; Kenny & Judd, 2014). In general, the magnitude of the indirect effect has been stable and practically meaningful across different tests. Specifically, the indirect effect was .20, .26, .27, and .19 for English, mathematics, science, and social studies tests, respectively (Cole et al., 2008). The indirect effect has been medium to large in magnitude across test instruction conditions. More specifically, when controlling for gender, prior knowledge, and test anxiety, the indirect effect was .13, .10, and .10 for first-year students who were told their test scores would be used for accountability mandates, who were told they would receive feedback, and who were told faculty would see their personal test scores conditions, respectively (Mathers et al, 2016). When controlling for conscientiousness, the indirect effect was .20, .29, and .32 for first-year students and .34, .29, and .29 for upper-class students who were told their test scores would be used for accountability mandates, who were told they would receive feedback, and who were told faculty would see their personal test scores, respectively (Myers et al., 2016). Moreover, in studies that did not manipulate test instructions, but simply told examinees their scores would be used for accountability, the indirect effect was medium to large in magnitude after controlling for

gender and prior knowledge (.09, Zilberberg et al., 2014) and test anxiety and expectancy (.24, Penk & Schipolowski, 2015).

Statement of the Problem Addressed by Current Study

The studies estimating the indirect effect of importance on performance via effort have focused little to no attention on the assumption of temporal precedence of value, effort, and performance. Recall, EV theory applied to low-stakes testing posits perceived importance precedes and influences expended effort, which in turn precedes and influences test performance (Wise & DeMars, 2005; Wolf & Smith, 1995). With the exception of Penk and Schipolowski (2015), previous studies modeling the indirect effect have assessed perceived importance and expended effort retrospectively, *after test completion*, despite the theoretical specification of value and motivation occurring before task performance (see the retrospective model in Figure 2 for an illustration). Given this theorized temporal precedence of variables, researchers are making an assumption that retrospectively measured perceived importance and examinee effort scores are the same as prospectively measured importance and effort scores (Cole & Maxwell, 2003).

With respect to importance, researchers and testing practitioners could easily employ the same measure of importance before or after completion of the test (e.g., “I am not concerned about the score I receive on this test”). However, importance scores collected retrospectively may actually reflect self-protective attributions of performance (Dong, Stupinsky, & Berry, 2013; Perry, Stupinsky, Daniels, & Haynes, 2008; Weiner, 1985, 2000). For example, after a rigorous test, examinees may incorrectly attribute the cause of their poor performance to low levels of perceived importance (e.g., “I did not care about the test”). Thus, retrospective self-reported levels of importance may be

invalid because these scores may reflect post-test attributions of test performance rather than true perceptions of importance.

With respect to examinee effort, it is not as easy to employ the same measure of effort before and after test completion (“I will give my best effort on this test” vs. “I gave my best effort on this test”). Examinees’ retrospective responses reflect *expended* effort, whereas examinees prospective responses reflect *intended* effort. Expended effort scores are, presumably, examinees’ reflections of their exhibited behavior while completing the test. However, expended effort scores may also be reflections of self-protective post-test attributions instead of actual expended effort (Dong et al., 2013; Perry et al., 2008; Weiner, 1985, 2000). In fact, researchers have found that students tend to attribute effort as the primary cause of test (Dong et al., 2013) and academic course (Perry et al., 2008) performance, over test difficulty, ability, and learning/test strategy. Prospectively measuring importance and effort may eliminate contamination of these scores.

Moreover, measuring importance and effort prospectively may affect test performance via priming mechanisms. Researchers investigating behavioral commitment have found that priming students by asking them to report their intended level of performance prior to a task can affect task performance (Carver & Scheier, 2000; Higgins, 1997). Behavioral commitment is conceptualized as individuals’ tendency to carry out a behavior to avoid the dissonance between their stated behavioral intentions and exhibited behavior. Research has shown a large effect of stated behavioral intentions on academic-related behavior and achievement (Manstead & van Eekelen, 1998; Sheeran,

Orbell, & Trafimow, 1999). If examinees' report their intended effort before the test, they may, in turn, exhibit effort reflective of their intended effort throughout the test.²

In sum, EV theory and previous research on motivation in low-stakes testing contexts specify perceived importance influences examinee effort, which influences test performance. However, researchers tend to violate the temporal precedence assumption of mediation by modeling this indirect effect using data collected out of temporal order. The timing of measurement of perceived importance and examinee effort relative to test completion may influence the validity of the inferences made from importance and effort scores. No previous research has examined the consequences of measuring perceived importance and examinee effort after test completion rather than prior to test completion, which follows the temporal order specified by EV theory.

Significance of the Current Study

This study will not be able to determine whether prospectively or retrospectively measured scores are more valid. However, it will help determine if inferences made from perceived importance scores, examinee effort scores, and test scores differ depending on when importance and effort are measured relative to test completion.

With respect to importance and effort scores, if both importance and effort scores are similar before and after test completion, then the timing of measurement may have no effect on the interpretation of importance and effort scores. However, if importance and

² Behavioral commitment is just one priming mechanism. Stereotype threat is another well-studied priming mechanism, yet less relevant to this study. Researchers investigating stereotype threat have shown that examinees identifying as Black tend to underperform on verbal tests after being asked to indicate their ethnicity immediately prior to beginning the test (Steele & Aronson, 1995). Examinees identifying as female tend to underperform on mathematics tests after being asked to indicate their gender immediately prior to beginning the test (Gresky, Eyck, Lord, & McIntyre, 2005).

effort scores are different from before to after test completion—likely due to self-protective attributions (e.g., Weiner, 2000) or behavioral commitment (Carver & Scheier, 2000; Higgins, 1997)—then the timing of measurement may have implications on the interpretation of importance and effort scores. Moreover, procedures such as motivation filtering used to account for attenuated test scores due to low motivation (Swerdzewski, Harmes, & Finney, 2011) may result in different findings depending on when importance and effort are measured.

With respect to test performance scores, asking examinees to report their perceived importance and intended effort before the test may influence test performance due to behavioral commitment (e.g., Perry et al., 2008). Consequently, test scores may be more reflective of examinee ability and may not be contaminated by construct irrelevant variance due to low effort (Haladyna & Downing, 2004). Assuming the low-stakes nature of the test is biasing test scores downward, the priming may result in, on average, higher test scores than without priming. If so, the simple act of engaging in prospective ratings of importance and effort would be an effective (and “cheap”) motivation intervention.

With respect to relationships between perceived importance, examinee effort, and test performance, if the relationships differ depending on when importance and effort are measured relative to test completion, then interpretation and implications of the indirect effect may differ. For example, motivation interventions aimed at manipulating effort via importance may be differentially effective (e.g., Finney, Sundre et al., 2016; Liu et al., 2015). More specifically, if the indirect effect is nil or small when importance and effort are measured before the test, then interventions created to increase importance would be useless with respect to increasing test performance.

Purpose of the Current Study

The purpose of the current study is to empirically examine the potential differential effects of measuring perceived importance and examinee effort before and after test completion. To help parse out these potential effects, I will examine if the indirect effect of importance on performance via effort differs when variables are measured in alignment with the theoretically specified temporal order of occurrence versus when variables are measured according to typical practice of measuring importance and effort after test completion in low-stakes testing. Uncovering differences in the relationships between importance, effort, and test scores does not indicate if the timing of measurement influences the magnitude of importance, effort, and test scores; hence, I will examine if average levels of importance, effort, and performance differ when importance and effort are measured prospectively versus retrospectively.

To accomplish this, first-year students were randomly assigned to one of three conditions during a university-wide, institutional accountability testing session (see Figure 2 for illustrations of conditions, theorized models, and example items). In the retrospective condition, data were collected consistent with most previous studies of the indirect effect of perceived importance on test performance: perceived importance and expended effort were measured after test completion. In the combined condition, which combines prospective importance and retrospective effort, perceived importance data were collected before the test and were collected again with expended effort after test completion. In the prospective condition, both perceived importance and effort data were collected before the test and again after test completion.

Research Questions

Comparing retrospective motivation scores across conditions. Research questions one and two focus on retrospective perceived importance and examinee expended effort scores across conditions that manipulated the presence and absence of behavioral priming (see Table 2). Does asking examinees to engage in rating their perceived importance and intended effort before test completion influence retrospective perceived importance and expended effort scores? To answer these questions, I will model cross-sectional retrospective importance and effort scores across conditions to examine possible effects of behavioral commitment on average levels of importance, effort, and test performance and their interrelationships.

Research question 1. On average, do examinees report different levels of retrospectively measured perceived test importance and examinee expended effort across measurement conditions? Examinees may attribute the cause of their test performance to low levels of retrospectively measured importance and effort (e.g., Perry et al., 2008). However, priming examinees by asking them to engage in rating their importance and effort before test completion may increase subsequent average test scores (e.g., Carver & Scheier, 2000). Accordingly, behavioral commitment may mitigate the effects of attributional bias on retrospective importance and effort scores in the prospective condition. Consequently, I expect average levels of retrospectively measured perceived importance and examinee effort scores will be higher in the prospective condition than the retrospective and combined conditions.

Research question 2. Given the typical practice of estimating the indirect effect of perceived test importance on test performance via examinee effort using retrospective

importance and effort scores (e.g., Zilberberg et al., 2014), this question focuses solely on retrospective importance and effort scores. Is the indirect effect of retrospective perceived importance on test performance via retrospective examinee effort affected by asking examinees to rate their perceived importance and intended effort before test completion? The indirect effect of retrospective importance on performance via retrospective effort estimated in the retrospective condition of Figure 2 will be compared to the indirect effects estimated using retrospective importance and effort scores from the combined and prospective conditions (see Table 2). The equivalence, or lack thereof, of the indirect effect across the three conditions will be assessed using a multiple-condition path model (i.e., moderated mediation model). Given retrospective importance and effort are modeled in all three conditions, any difference in the magnitude of the indirect effect will be due to the presence or absence of examinees simply engaging in prospective ratings.

I expect the completely mediated model of importance on performance via effort will fit the data across the three conditions. Mediation would be suggested if correlations between retrospective importance and retrospective effort and between retrospective effort and performance are larger than the correlation between retrospective importance and performance. Given adequate model-data fit across conditions, I will estimate the magnitude of the indirect effect for each condition.

Examinees may attribute the cause of their test performance to low levels of importance and effort (e.g., Perry et al., 2008). Thus, the relationship between test performance and retrospective importance and effort scores may be strong due to retrospective scores being post-test attributions of test performance. However, asking examinees in the prospective condition to rate their perceived importance and intended

effort before the test may result in behavioral commitment. That is, if students engage in effort according to their ratings of intended effort, their test scores may better reflect their true ability and thus behavioral commitment may mitigate the effect of attributional bias on retrospective importance and effort scores. In sum, removing the effect of attributional bias on retrospective importance and effort scores in the prospective condition may attenuate the relationship between test performance and retrospective importance and effort scores. As a result, the relationship between test performance and retrospective effort may be weaker in the prospective condition than in the retrospective condition. Accordingly, I expect the magnitude of the indirect effect of retrospective importance on performance via retrospective effort will be smaller in the prospective condition than in the retrospective condition. Although these results are not able to confirm the influence of attributional bias on retrospective importance and effort scores, they would indicate if behavioral commitment via priming buffers the possibly inflated indirect effect that may be a result of attributional bias.

Comparing prospective and retrospective motivation scores across conditions.

Research question 3. Do examinees reporting retrospective perceived importance and expended effort have lower scores, on average, than examinees reporting prospective importance and intended effort for the same test? In contrast to comparing average levels of retrospective perceived importance and expended effort (research question 1), here I will compare average retrospective importance and effort scores from the retrospective condition, average prospective importance and retrospective effort scores from the combined condition, and average prospective importance and effort scores from the

prospective condition (see Figure 2). Given examinees, on average, tend to perform poorly on the test administered in the current study (Finney, Sundre et al., 2016), I expect average prospective importance and effort scores from the prospective condition will be higher than retrospective importance and effort scores from the retrospective condition due to retrospective importance and effort scores being self-protective attributions of test performance. Research question one will indicate if behavioral commitment mitigates the effect of attributional bias on importance and effort scores. In other words, does simply engaging in prospective ratings of importance and effort influence retrospective ratings of importance and effort. This comparison will indicate if examinees report different levels of prospective importance and effort than retrospective importance and effort.

Research question 4. Does the theoretically and empirically suggested indirect effect of perceived test importance on test performance via examinee effort differ depending on if examinees report importance and effort prospectively versus retrospectively? In the retrospective condition, the indirect effect of retrospective importance on performance via retrospective effort will be estimated, aligning with the typical practice. In the combined condition, the indirect effect of prospective importance on performance via retrospective effort will be estimated. In the prospective condition, the indirect effect of prospective importance on performance via prospective effort will be estimated, aligning with the hypothesized temporal order of occurrence (see Figure 2).

Similar to research question two, the equivalence of the indirect effects of importance on performance via effort across the three conditions will be assessed via moderated mediation analysis. I expect the completely mediated model of importance on performance via effort will fit the data across the three conditions. Research question two

will indicate if the indirect effect estimated from retrospective importance and effort scores is influenced by merely asking examinees to engage in prospective ratings of importance and effort, whereas this comparison will indicate if the indirect effect depends on when importance and effort are measured (prospective vs. retrospective).

Whereas modeling retrospective importance and effort scores (research question 2) may help parse out potential effects of behavioral commitment due to priming, this question may help determine the effect of attributional bias and behavioral commitment on the indirect effect. The indirect effect when modeling retrospective importance and effort scores in the retrospective condition will be free from any influence of priming; thus, no possibility of behavioral commitment. However, these scores may reflect post-test attributions of test performance rather than true perceptions of importance and effort. The indirect effect when modeling prospective importance and effort scores in the prospective condition will be free from any contamination due to attributional bias, but may be influenced by behavioral commitment.

Attributional bias may serve as a self-protective mechanism to a greater degree for examinees who perform poorly on the test than for examinees who perform well (Pekrun et al., 2004). Imagine an examinee performs poorly on a test and attributes the cause of their poor performance to low effort and thus decreases their effort rating from before to after the test. Another examinee performs well on the test, has no reason to self-protect, and rates their effort equivalently before and after the test. Consequently, one would expect a larger range of scores and more variability in retrospective effort scores than prospective scores. However, the unstandardized relationships from importance to effort and effort to performance may change negligibly. As a result, the unstandardized

indirect effect may be essentially equivalent across conditions, but the standardized indirect effect in the prospective condition would be smaller in magnitude—due to less variability in effort scores—than the indirect effect in the retrospective condition. Similarly, one would expect higher correlations between importance, effort, and performance in the retrospective condition than in the prospective condition.

An important implication of the expected results suggests the mediated model ($I \rightarrow E \rightarrow P$) is misspecified. That is, if retrospective importance and effort scores are contaminated by attributional bias, then test performance is the common cause of both importance and effort ($P \rightarrow I, P \rightarrow E$). Research question six may be able to further explicate the effect of attributional bias on retrospective importance and effort scores.

Changes in importance and effort scores. Unlike the above research questions which focus on differences in average importance and effort scores or indirect effects of importance on performance via effort across testing conditions, research questions five and six focus on *changes* in importance and effort across time (see Table 2). Using longitudinal data from the prospective condition, I will compare examinees' prospectively measured perceived importance and intended effort scores to their retrospectively measured importance and expended effort scores (research question 5). I will also assess if this change is related to test performance (research question 6).

Research question 5. Do examinees completing both prospective and retrospective measures report different levels of prospective perceived importance and examinee effort, on average, than retrospective importance and effort? That is, do self-reported importance and effort scores *change* as a result of examinees' experience with the test? Examinees may exhibit behavior aligning with their behavioral intentions (e.g.,

Carver & Scheier). If this were true, I would expect prospective and retrospective importance and effort scores to be equivalent. However, examinees may make post-test attributions (e.g., low expended effort, disinterest in test) to explain their test performance (e.g., Perry et al., 2008). Whereas this effect may be nil for examinees who perform relatively well, examinees who perform poorly may be more susceptible to the effect of attributional bias (Perry et al., 2008). Given the rigor of the quantitative and scientific reasoning test, many examinees may perform poorly and may attribute the cause of their poor performance to low expended effort. Thus, I expect average levels of retrospective importance and effort scores will be lower than average levels of prospective scores (i.e., there will be a decrease in reported importance and effort from before to after the test).

Research question 6. Are perceived importance and examinee effort change scores (difference between prospective and retrospective importance and effort) related to performance on the quantitative and scientific reasoning test? That is, is the main effect estimated in research question five actually moderated by test performance? Attributional bias may serve as a self-protective mechanism to a greater degree for examinees who perform poorly on the test than for examinees who perform well (Pekrun et al., 2004). I expect larger decrease in importance and effort scores for examinees who perform relatively poorly than for the examinees who perform relatively well. These results would suggest examinees make post-test attributions to explain their (poor) test performance (e.g., Perry et al., 2008).

Test performance differences: Research question 7. Does test performance differ, on average, as a function of measurement condition? Using cross-sectional data to answer this question, I will compare average levels of test scores across three conditions.

Given students tend to exhibit behavior in accordance with their behavioral intentions (Carver & Scheier, 2000; Higgins, 1997), I expect average test scores will be higher in the prospective condition than test scores from the other conditions. Accordingly, test scores may be more reflective of examinees' ability after engaging in rating perceived test importance and intended effort prior to test completion. In turn, engaging in prospective ratings would then serve as a powerful yet resource-light test-taking motivation intervention.

Chapter II

Review of the Literature

Scientists aspire to gain a deeper understanding of their research domain by determining *why* or *how* one variable affects another variable (Frazier, Tix, & Barron, 2004; Judd & Kenny, 1981b). That is, regardless if observational, quasi-experimental, or experimental research designs are used, cause and effect inferences are the fundamental objective of many scientists. Nonetheless, inferences from scores gathered using particular research designs or measurement approaches may not be equally valid.

This chapter reviews two concepts related to the validity of inferences made from scores and statistical analyses. First, I will discuss the concept of mediation and how mediation hypotheses can be tested statistically. Second, I will discuss retrospective and prospective measurement, how type of measurement can influence response patterns, and the implications of retrospective versus prospective measurement.

Mediation

Scientists began explaining relationships between variables in terms of a simple two-variable stimulus-response or cause and effect model, Woodworth (1928) then expanded this simple model by describing psychological phenomena as a process, or sequence of events. Instead of the stimulus-response model, he suggested that a cause may be the effect of a previous cause. For example, a stimulus such as a loud noise may *stimulate* an organism, which *processes* the stimulus, subsequently *causing* a response.

This conceptualization of some organism, or intermediary, being responsible for the link between a stimulus and a response, now known as mediation, has since become a popular model for explaining behavior. Mediation refers to a process in which an

intervening variable, or mediator, helps explain how a predictor transmits its effect to a criterion (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). Said another way, a predictor (X) causes an effect in a mediator (Z) which causes an effect in a criterion (Y). The mediator functions as both the criterion of the predictor (X) and the predictor of the criterion (Y). For example, the theory of planned behavior suggests attitudes and norms (predictors) affect behavioral intentions (mediator) which in turn affect behavior (criterion; Azjen, 1991). Mass communications researchers suggest political mass media (predictor) affects voter likelihood (mediator) which affects voting behavior (criterion; McLeod, Kosicki, & McLeod, 2002). Cognitive psychologists suggest age (predictor) affects executive functioning (mediator) which affects cognitive functioning (criterion; Salthouse, Atkinson, & Berish, 2003). Prevention researchers posit that smoking interventions (predictor) affect attitudes toward smoking (mediator) which affect smoking behavior (criterion; Worden & Flynn, 2002). Educational researchers suggest the relationship between identification as a racial minority (predictor) and test performance (criterion) is mediated by anxiety (Osborne, 2001). Educational psychologists suggest the effect of achievement emotions (predictor) on academic performance (criterion) is mediated by students' motivation (Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011). Accordingly, mediators (e.g., voter likelihood, attitudes toward smoking, anxiety) help explain *why* or *how* predictors (e.g., political media, smoking intervention, racial differences) affect criterions (e.g., voting behavior, smoking behavior, test performance).

Considering the language used in the previous examples, a *theoretical* mediation model is inherently causal in nature, as reflected by the directional pattern of

relationships specified by the *theorized* model (Rose, Holmbeck, Coakley, & Franks, 2004), not unlike most research hypotheses. Regardless of the experimental design or data analytic technique, causal inferences necessitate a program of research. We do not know the underlying true mechanism that causes variables to be related; we can only hypothesize the mechanism based on theory and provide empirical evidence that supports or refutes the theory. Fortunately, theories about the mechanism imply certain patterns of relationships in data. If we observe those patterns of relationships in data, then there is empirical evidence that the theory is plausible. However, empirical evidence does not confirm theories; it simply fails to disconfirm theories. In sum, despite the theory and language used to communicate mediation hypotheses, which is causal nature (e.g., smoking interventions affect smoking behavior via attitudes toward smoking), causal inferences are only justified from empirical research under certain methodological conditions and necessitate replication.

Importance of Mediation

Elucidation of the interrelations among variables by testing, integrating, and refuting theories suggestive of mediation processes allows for more informed causal inferences (James & Brett, 1984; see Table 1 for a quantitative, visual, and narrative comparison of *third variable* functions). Consequently, this clearer understanding accommodates applied research in the areas of prevention, intervention, and treatment (MacKinnon, Fairchild, & Fritz, 2007; Rose et al., 2004). Instead of developing interventions intended to solely target and manipulate the criterion, interventions can be designed to target single or multiple mediators in a causal chain, thereby making the intervention more effective (Kraemer, Stice, Kazdin, Offord, & Kupfer, 2001).

For example, communications researchers investigating mass media have established the relationship between political mass media (X) and voting behavior (Y) is mediated by voter attitudes (Z ; McLeod et al., 2002). Subsequent research in this domain found political news (X) affects political engagement (Y) two ways: (1) indirectly through online political discussion (Z_1) and (2) indirectly through interpersonal political discussion (Z_2) and political knowledge (Z_3 ; see Figure 3 for an illustration of this multiple mediator example; Jung, Kim, & Zúñiga, 2011). These results imply that future interventions to encourage voting behavior should be aimed at manipulating political media, which will influence both online and interpersonal political discussion. Thus, development of effective interventions can be facilitated by determining what variables (Z) transmit the effect of X on Y . By strengthening the effect of the intervention on the mediating variables, one can strengthen the effect of X on Y .

An example of mediation in a drug prevention program helps highlight the utility in examining indirect effects. Researchers evaluated the effect of participating in a drug prevention program (dichotomous X) on drug use (Y) via patients' perceived friend's reactions to drug use (Z ; MacKinnon et al., 1991). The drug prevention program (X) influenced how patients perceived their friends would react to their drug use (Z). Subsequently, patients' perceptions about how their friends would react discouraged patients' drug use (Y). Importantly, the researchers provided evidence of *why* a patient decreased their drug use, and by designing interventions to manipulate perceived friend's reactions to drug use (Z), researchers were able to decrease drug use.

This concept of mediation has implications for the current study. Testing practitioners may not be able to *directly* manipulate examinee effort. However,

practitioners may be able to design interventions that influence a predictor of effort such as perceived test importance which will, in turn, influence effort and test performance.

Complete and partial mediation. There is a clear distinction both theoretically and statistically between complete and partial mediation. With respect to both complete and partial mediation, the *total* effect (c) of a predictor (X) on a criterion (Y) can be decomposed into two effects: *indirect* effect via the mediator (Z) and *direct* effect. Potentially, the predictor can affect the criterion in two ways: an indirect effect (ab) from the predictor (X) to the criterion (Y) through the mediator (Z) and a direct effect (c') from the predictor to the criterion (see Figure 4 for an illustration).

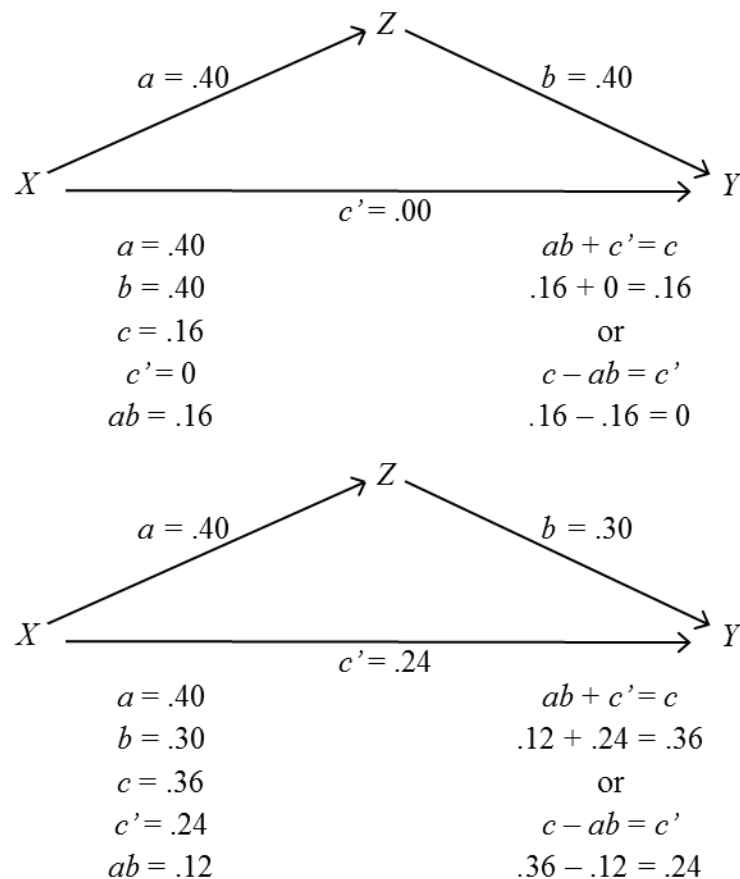


Diagram 1. Illustration of Complete Mediation (top) and Partial Mediation (bottom)

Complete mediation refers to the phenomenon where the indirect effect of the predictor on the criterion through the mediator (ab) *fully* accounts for the total effect of the predictor on the criterion (c ; see row 1 of Table 1). Consequently, because the full effect of X on Y is through Z , the direct effect of X on Y (c') is nil. In contrast, *partial mediation* refers to the phenomenon where the indirect effect of the predictor on the criterion through the mediator (ab) *partially* accounts for the total effect of the predictor on the criterion (c ; see row 2 of Table 1). Consequently, because only part of the effect of X on Y is through Z , the direct effect of X on Y (c') must be nonzero. That is, X directly affects Y and indirectly affects Y through Z (see Figure 5 for an illustration of complete and partial mediation). To help distinguish between complete mediation and partial mediation, an example of complete mediation follows. The effect of parents' medical care responsibilities (X) on parents' quality of life (Y) is completely mediated by family environment (Z ; Crespo, Carona, Silva, Canavarro, & Dattilio, 2011). The total effect of medical responsibilities (X) on quality of life (Y) is due to the influence of medical responsibilities (X) on family environment (Z ; as responsibilities increase, environment gets worse) which directly influences quality of life (Y ; as environment gets worse, quality of life gets worse). There is no direct effect of medical responsibilities (X) on quality of life (Y) given the direct effect of family environment (Z) on quality of life (Y). Thus, attention could be directed to bolster the family environment to lessen the impact of medical care responsibilities on parents' quality of life. Another example of complete mediation involves the effect of a drug intervention (X) on adult substance abuse (Y) via delayed substance use as an adolescent (Z ; Spoth, Trudeau, Gyll, Shin, & Redmond, 2009). Participation in a drug intervention program tends to delay participants' substance

use as an adolescent, which tends to mitigate substance abuse as an adult. The two previous examples illustrate how knowledge of mediated relationships can facilitate development of effective interventions. The first example illustrates how one can focus on trying to manipulate the mediator (family environment) via an intervention because it is not feasible to manipulate parents' medical care responsibilities. The second example illustrates manipulation of the predictor (drug intervention program) that will influence the mediator (delayed substance abuse) and, in turn, the criterion (adult substance abuse).

In contrast, partial mediation exists when the predictor (X) has both an effect on the criterion (Y) directly and indirectly through the mediator (Z). For example, the effect of examinees' ethnicity (X) on test performance (Y) is partially mediated by anxiety (Z ; Osborne, 2001). The total effect of examinees' ethnicity (X ; self-identification as a minority group member) on test performance (Y) is due to the influence of examinees' ethnicity (X) on anxiety (Z) which influences test performance (Y ; i.e., indirect effect) *and* the influence of examinee ethnicity directly on test performance (Y). Notice anxiety *partially* explains the relationship between examinees' ethnicity and test performance. As another example of partial mediation, the effect of state anxiety (X) on quantitative reasoning performance (Y) is partially mediated by verbal working memory (Z). Thus, state anxiety (X) has a direct effect on quantitative reasoning (Y) not accounted for by working memory (Z ; Owens, Stevenson, Norgate, & Hadwin, 2008).

The conceptual explanation of the direct effect of X on Y (c') in partial mediation models can prove difficult to generate. The direct effect of examinee ethnicity (X) on test performance (Y) is likely due to unmeasured mediators (variables affected by ethnicity, which in turn affect test performance). Perhaps examinee effort (Z_2) should have been

measured and modeled; thus, it may be that the effect of examinee ethnicity (X) on test performance (Y) is mediated by examinee anxiety (Z_1) and mediated by examinee effort (Z_2). By not assessing examinee effort (Z_2), the direct effect of examinee ethnicity (X) on test performance (Y) is needed to account for the relationship between ethnicity (X) and test performance (Y) not explained by examinee anxiety (Z_1). Partial mediation may also be a result of a poorly measured mediator. If the mediator is not measured well, the X - Z and Z - Y path coefficients may be attenuated due to measurement error. Accordingly, the indirect effect will be attenuated. As a result, a direct path between X and Y is needed to account for the observed correlation between X and Y . In short, a poorly measured mediator is essentially not represented in the model (Baron & Kenny, 1986).

To further illustrate mediation, I will refer to the three variables from this research study: perceived test importance (X), examinee effort (Z), and test performance (Y). The literature on test taking suggests perceived importance of a test is positively related to test performance (e.g., Cole et al., 2008; Finney, Sundre et al., 2016; Zilberberg et al., 2014). Moreover, examinee effort may help explain *how* or *why* perceived importance and test performance are related. That is, examinees high in perceived importance tend to be high in examinee effort, relative to examinees low in perceived importance. Examinees high in examinee effort tend to have high test performance, relative to examinees low in examinee effort. These empirical relationships may be explained theoretically (Eccles et al., 1983; Wigfield & Eccles, 1992, 2000, 2002): examinees' perceived importance positively affects test performance indirectly via its positive effect on examinee effort (i.e., perceived importance \rightarrow examinee effort \rightarrow test performance). If the effect of perceived importance on test performance were completely transmitted through examinee

effort, this effect would be considered complete mediation. Conceptually, complete mediation is interpreted as perceived importance causes test performance *only* through effort. Thus, importance has no relationship with performance that cannot be accounted for by examinee effort. Said another way, examinee effort explains how and why perceived importance is related to test performance.

In contrast, perhaps examinee effort only accounted for part of the effect of perceived importance on test performance. Although perceived importance may directly *cause* test performance, more proximal variables such as examinee effort or test anxiety are more likely to directly cause test performance. A more likely possibility of why importance would have a direct effect on performance, in addition to the indirect effect via effort, is because of model misspecification. Perhaps an omitted or unmeasured variable such as test anxiety can partially explain why importance is related to performance. That is, it is possible the effect of perceived importance on test performance is completely mediated by *both* examinee effort ($I \rightarrow E \rightarrow P$) and test anxiety ($I \rightarrow A \rightarrow P$). Although this theory would specify *complete* mediation, failure to model test anxiety as a mediator between importance and performance necessitates a nonzero direct effect from importance to performance to account for the total relationship between perceived importance and test performance (i.e., a partial mediation). One would theorize a completely mediated model (indirect effect of X on Y via Z and *no* direct effect of X on Y) if they believed that Z *completely* explains how and why X causes Z . In contrast, it is hard to conceptualize when one would theorize a partially mediated process (indirect effect of X on Y via Z *and* direct effect of X and Y). It is more likely that partial mediation is

empirically supported post hoc, after the theorized completely mediated model is rejected due to model misspecification (e.g., omitted mediator) or measurement error.

Third Variables

Clearly, going beyond two-variable experiments enables scientists to better understand *how* and *why* a predictor and a criterion are related. Mediation is just one way to explicate the relationship between two variables. *Third variables* can be specified to represent theories reflecting other mechanisms as well such as suppression, moderation, and confounding (see Table 1 for examples of third variables and example hypotheses). Although third variables can present certain challenges in interpretation and estimation, well-established theory and proper specification of third variables is essential when designing experiments and testing competing models to support or refute existing theories. The following section will present the different functions a third variable can take on with respect to the relationship between predictors and criteria. I will discuss these different functions the third variable, examinee effort, can take on with respect to perceived importance and test performance.

Suppressor. A suppressor variable is special case of intervening variable where, when compared to a bivariate X - Y relationship, estimation of a model with a suppressor results in the regression weight of the predictor (X) on the criterion (Y) increasing in magnitude and/or a change in the sign (Conger, 1974; Kline, 2010; MacKinnon, Krull, & Lockwood, 2000; Shrout & Bolger, 2002). In the classical case of suppression, the bivariate X - Y relationship is nil (i.e., $r_{yx} = .00$ or $c = .00$); however, when the suppressor (Z) is added into the model predicting the criterion (Y), the effect of X on Y increases in magnitude after controlling for Z ($c' = -.16$). In classical suppression, a suppressor is not

identified by its regression weight; a variable is a suppressor if the regression weights for other variables increase or change signs.

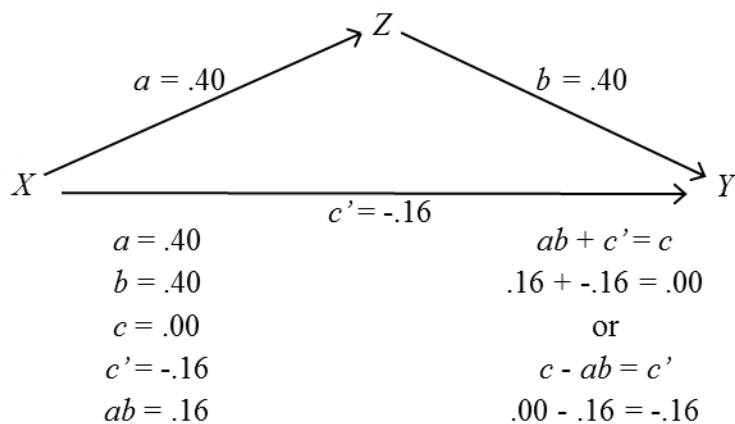


Diagram 2. Illustration of Classical Suppression

Consider a hypothetical example (see Diagram 2 above) provided by McFatter (1979) where assembly line workers' intelligence (X) and number of errors made on an assembly line (Y) are measured. Intuitively, one would hypothesize intelligence is negatively related to number of errors. However, if only intelligence and number of errors are measured, the *bivariate* relationship is nil (e.g., $c = .00$). Suppose researchers decided to measure workers' boredom (Z ; suppressor) as well. More intelligent workers tend to be more bored when working on an assembly line ($a = .40$), increased boredom tends to be associated with more errors ($b = .40$). Thus the indirect effect of intelligence on number of errors becomes nonzero ($ab = .16$) and positive, as originally hypothesized. Notice decomposition of the total effect ($c = .00$) results in a negative direct effect ($c' = -.16$) and an equal in magnitude, but opposite sign indirect effect ($ab = .16$). Consequently, because the total effect is equal to the sum of the indirect and direct effects, the total effect is nil. Thus, misspecification of the model by not including boredom would result in a failure to accurately portray the components resulting in the nil relationship between intelligence and number of errors on an assembly line.

Whereas classical suppression presumes a nil bivariate X - Y relationship (r_{xy}), negative suppression presumes all bivariate relationships (r_{xy} , r_{xz} , and r_{zy}) are positive. However, when the suppressor is added, one of the regression coefficients is in the opposite direction of its bivariate correlation (Conger, 1974; Kline, 2010; Pandey & Elliot, 2010). As with classical suppression, when the suppressor is added into the model, the total effect is decomposed into nonzero direct (c') and indirect (ab) effects. An

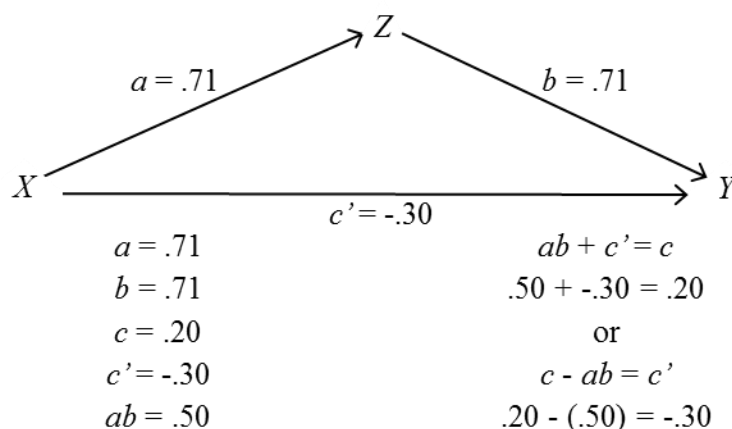


Diagram 3. Illustration of Negative Suppression

example given by Kline (2010) of negative suppression follows (see Diagram 3 above). Psychotherapy (X) and prior suicide attempts (Y) are, *surprisingly*, positively related ($r_{xy} = .20$); thus, it might be interpreted that psychotherapy is harmful. The bivariate correlations between psychotherapy (X) and depression (Z), and depression (Z) and suicide attempts (Y) are both positive. When estimating path coefficients, the psychotherapy (X) and depression (Z) path coefficient ($a = .71$), and the depression (Z) and suicide attempts (Y) path coefficient ($b = .71$) are positive. However, by modeling depression (suppressor), the path coefficient between psychotherapy (X) and suicide attempts (Y ; controlling for depression) is negative ($c' = -.30$), as one might expect.

Notice these suppression effects can be conceptualized as mediation effects: the effect of X on Y is mediated by Z (see rows 1, 2, and 3 of Table 1 for comparisons of suppression and mediation). When a suppressor functions as a mediator, the mediation effect is often termed as *inconsistent mediation* in contrast to consistent mediation where the signs of the direct and indirect effects are both the same (MacKinnon et al, 2007, 2000; Shrout & Bolger, 2002). Importantly, when classical or negative suppression is the underlying mechanism, the direct effect of X on Y will be underestimated when the suppressor is not modeled, further illustrating the importance of understanding theory and specifying models accordingly. In sum, addition of a suppressor (Z) variable into a model *increases* or *changes* the direct effect between the predictor (X) and the criterion (Y).

Covariate. Another function of a third variable is when the third variable (Z) is related to the criterion (Y), has a negligible relationship with the predictor (X), and does not interact with the predictor with respect to the effect of the predictor on the criterion. In this situation, the third variable is referred to as a covariate (MacKinnon, 2008). Including a covariate adds predictive utility to the model over and above the predictor(s). Because the predictor and the covariate are ideally negligibly related, controlling for the effect of the covariate on the criterion does not substantially change the relationship between the predictor(s) and the criterion. Said another way, adding a third variable—covariate—to the model will have little to no effect on the predictor to criterion relationship, but will add to the prediction of the criterion.

For example, although difficult to imagine, that perceived test importance and examinee effort are negligibly related, yet both have strong, positive relationships with test performance. Accordingly, including the covariate, examinee effort, to a model that

includes the predictor, perceived importance, adds to the accuracy of predicting test performance, but does not substantially change the regression coefficient associated with predicting test performance from test importance (see row 4 of Table 1 for an illustration of a path model that includes a covariate). Thus, assuming X and Z are negligibly related, failure to model covariates does not result in substantially biased estimates of the X - Y relationship, yet decreases the prediction of Y .

Confounder. Unlike a covariate that adds to the prediction of a criterion in a model and has a negligible effect on the X - Y relationship, or how a suppressor can result in increased magnitude of the direct effect between X and Y , confounders account for part (or all) of the relationship between the predictor (X) and the criterion (Y); thereby reducing the effect of X on Y . In the extreme case, when partialling the effects of a complete confounding variable from the predictor and criterion relationship, the (partial) relationship between the predictor and the criterion (c') is nil. In this case, the predictor and criterion are said to have a spurious relationship.

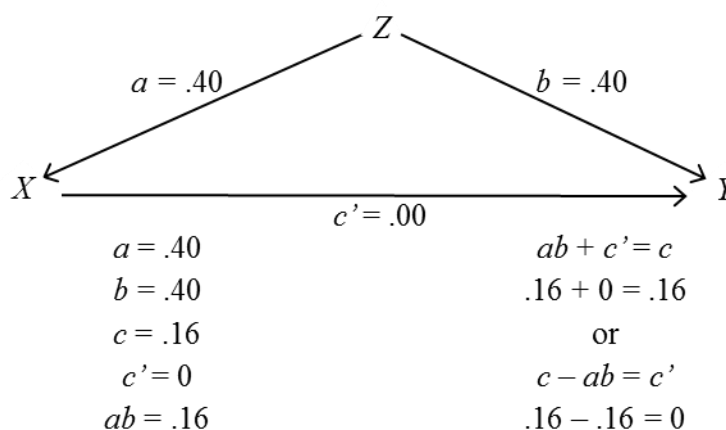


Diagram 4. Illustration of Confounder

Returning to the psychotherapy (X), depression (Z), and suicide attempts (Y) example above. It may be that engaging in psychotherapy and suicide attempts is a

function of level of depression. After controlling for depression, the positive relationship between psychotherapy and suicide attempts is nil.

With respect to the current study, perhaps examinees perceive tests to be of more importance (X) because they expend relatively higher levels of effort (Z). Moreover, examinees perform better on tests (Y) because they expend relatively higher levels of effort on the test (Z). Thus, the bivariate relationship between importance and test performance can be explained by effort (i.e., importance and performance are spuriously related due to effort; once this common cause is controlled, their partial correlation is nil). Confounders and completely mediated effects are equivalent in accounting for the X - Y relationship (see rows 1 and 5 of Table 1 for a comparison of confounding and complete mediation). Thus, although conceptually they represent different theories for the variable interrelations, confounders and mediators cannot be distinguished statistically—both will have the same pattern of interrelationships.

Moderator. If the effect of a predictor (X) on a criterion (Y) depends, or is conditional, on a third variable (Z), that third variable is considered a moderator (Baron & Kenny, 1986). Similar to mediators, moderators may be trait-like attributes (e.g., personality), experimentally manipulated variables (e.g., treatment vs. control), or background variables (e.g., gender, socioeconomic status, school attended; Wu & Zumbo, 2008). When discussing moderation, the term *interaction effect* is often used. The predictor (X) and the third variable (Z) *interact* in the way they affect the criterion (Y). If the strength of the relationship between the predictor and the criterion depends on the level of third variable, then the effect of the predictor on the criterion is moderated by the third variable (see Figure 6 for examples of ordinal and disordinal moderation).

If the effect of perceived importance on test performance is conditional on the students' level of examinee effort, the relationship between perceived importance and performance is moderated by examinee effort. Given this example, one would likely hypothesize an ordinal interaction where, at high levels of examinee effort, there is no relationship between perceived importance and test performance, but at lower levels of effort, the relationship between importance and test performance is positive. Not only can the strength of the relationship between X and Y differ across levels of Z , the sign of the relationship may be affected by the moderator as well (Baron & Kenny, 1986; James & Brett, 1984). If the sign of the relationship changes, one would likely hypothesize a disordinal interaction. For example, although difficult to imagine, that at high levels of effort, the relationship between importance and test performance is negative, but at low levels of effort the relationship between importance and performance is positive. In the current study, one may hypothesize an interaction between importance and effort if previous studies suggested inconsistent relationships between importance or effort and performance. However, studies have demonstrated the effect of importance on performance does not depend on effort (Mathers et al., 2016; Myers et al., 2016).

It is critical to understand the distinction between moderators and other third variables such as mediators, especially in the development of intervention and prevention programs. Consider the ordinal interaction between importance and effort described above. Now consider an intervention designed to increase test importance for the examinees. Conceptually, test performance is not affected by importance for those examinees with high levels of examinee effort (moderator). In contrast, test performance is positively affected by importance for those examinees with lower levels of examinee

effort. Thus, the intervention would be useless for examinees expending higher levels of effort, but may be effective for examinees expending lower levels of effort.

Differences between third variable hypotheses and correlation patterns. To further clarify the different functions of a third variable, I will provide a brief overview of third variable hypotheses and expected pattern of relationships between X , Z , and Y . Does X cause Y indirectly via Z ? Empirically, if complete mediation were plausible, one would expect the correlation between X and Y to be smaller than the correlations between X and Z or Z and Y , because the X - Y relationship is entirely a function of the indirect effect of X on Y via Z . For example, imagine the effect of X on Z is .40, the effect of Z on Y is .40, and the bivariate X - Y relationship is .16. The product of the X - Z and Z - Y relationships equals the indirect effect (X on Y via Z) of .16. Thus, the indirect effect (.16) fully accounts for the observed X - Y relationship of .16 (i.e., complete mediation). In contrast, imagine the effect of X on Z is .40, the effect of Z on Y is .30, and the bivariate X - Y relationship is .36. The product of the X - Z and Z - Y relationships equals the indirect effect (X on Y via Z) of .12. Partial mediation would be suggested because indirect effect (.12) does not fully account for the bivariate X - Y relationship (.36). A direct path equal to .24 (i.e., $.36 - .12 = .24$) is necessary to fully account for the X - Y relationship (see Figure 5).

Is X more predictive of Y when Z is modeled? Empirically, if suppression were plausible, one would expect a particular pattern of intercorrelations between X , Z , and Y . Similar to mediation, the correlation between X and Y would be smaller than the correlations between X and Z or Z and Y , because the X - Y relationship is entirely a function of the indirect effect of X on Y via Z . Imagine positive correlations between X and Z and Z and Y and a correlation between X and Y of zero. However, when estimating

the model, one of the path coefficients between a predictor and the criterion becomes larger in absolute value than its respective bivariate correlation. Thus, X or Z may become more predictive of Y when in the presence of the suppressor (see Figure 7).

After controlling for Z , is X related to Y ? Empirically, if a covariate were plausible, one would expect a particular pattern of intercorrelations between X , Z , and Y . For example, imagine the correlation between X and Y is .50, the correlation between Z and Y is .30, and the correlation between X and Z is .02. The semi-partial correlation between X and Y controlling for Z (.49) would be negligibly smaller than the bivariate X - Y correlation (.50). Similarly, the semi-partial correlation between Z and Y controlling for X (.29) would be negligibly smaller than the bivariate Z - Y correlation (.40). Further, the X - Y relationship should be similar at all levels of Z . The magnitude of the X - Y partial correlation would decrease as the magnitude of the X - Z correlation increased. An important distinction between a mediator and a covariate is temporal precedence. That is, a mediator must occur in time and space between X and Y , whereas a covariate can occur before or at the same time as X .

Is the observed relationship between X and Y spurious due to both variables being caused by Z ? Empirically, if confounding were plausible, one would expect a particular pattern of intercorrelations between X , Z , and Y . Despite confounders and mediators being conceptually different, both confounders and mediators can partially or completely account for the X - Y relationship equivalently. Thus, one would expect the same pattern of correlations between X , Z , and Y when Z is a mediator and when Z is a confounder. A distinction between mediators and moderators is a mediator occurs *within* a causal sequence of variables ($X \rightarrow Z \rightarrow Y$), whereas no assumption of temporal precedence of a

confounder is made (MacKinnon et al., 2000). In sum, when Z is a confounder, instead of the arrow (i.e., direction of causation) pointing from X to Z , when Z is a confounder, the arrow points from Z to X . Thus, specification of confounders and mediators is largely dependent on the theoretical rationale describing the relationships (see Figure 8).

Does the effect of X on Y depend on Z ? Empirically, if moderation were plausible, one would expect a particular pattern of intercorrelations between X , Z , and Y . For example, imagine at one level of Z the path coefficient between X and Y is .40, the path coefficient between Z and Y is .20, and the correlation between X and Z is .20. Now imagine at another level of Z the path coefficient between X and Y is .25, the path coefficient between Z and Y is .20, and the correlation between X and Z is .20. Thus, the X - Y relationship changes across levels of Z . A further distinction between mediators and moderators is a moderator occurs before or at the same time as X .

Summary. Although third variables are conceptually very different (Table 1 provides a concise summary of these third variables), it is sometimes difficult or impossible to differentiate between third variables empirically. Accordingly, the importance of a strong theoretical framework cannot be understated. As you will see in the following section, in addition to theory, characteristics of the research design can also help differentiate mediators from other third variables. Further, a fundamental differentiation to point out is covariates, moderators, and confounders are all positioned as predictors (Z) occurring at the same time, or sometimes before, the predictor of interest (X). In contrast, because mediators are effects of the predictor and causes of the criterion, they function as intermediaries occurring between the predictor and criterion.

Statistical Estimation to Support Mediation

Statistical estimation cannot confirm mediation hypotheses, but it can provide evidence to support or reject mediation hypotheses. Mediation is best inferred when using a combination of theory suggesting a mediation process, causal-chain research design, and statistical evidence replicated over a several studies (Little, Card, Bovaird, Preacher, & Crandall, 2009; Shrout & Bolger, 2002; Spencer, Zanna, & Fong, 2005). In the following section, I present some of the more common methods of estimating mediation. Although the concept of causal chains of variables (MacCorquodale & Meehl, 1948; Woodsworth, 1928) and path analysis (Wright, 1923) has been around for several decades, not until recently has statistically modeling mediation become popular (e.g., MacKinnon, 2008). A multitude of techniques to statistically estimate mediation have been proposed across several disciplines (for a review, see Preacher, 2015), albeit the most popular in the literature is the causal-steps approach (Baron & Kenny, 1986; Judd & Kenny, 1981a, 1981b; Kenny Kashy, & Bolger, 1998).

Causal-steps approach. The following four steps represent the Baron and Kenny approach or “causal-steps approach” (Baron & Kenny, 1986; Judd & Kenny, 1981a; 1981b; Kenny, Kashy, & Bolger, 1998). As thoroughly covered in the section above, inferences of mediation heavily rely on the theoretical arguments supporting mediation and not entirely on statistical estimation and testing. The following section will present the steps of statistically estimating and testing mediation via the multi-step, ordinary least squares (OLS) regression, causal-steps approach with presentation of formulas to assist the reader in understanding the concepts. See Figure 4 for an illustration of the path coefficients being estimated in the following steps.

Step 1: The criterion (Y) is regressed on the predictor (X) to determine if a statistically and practically significant X - Y relationship exists to be mediated. Note that e represents the residual of Y .

$$Y = \beta_0 + \beta_{yx}X + e \quad (1)$$

Step 2: The mediator (Z) is regressed on the predictor (X) to determine if a potential mediator exists. Note that e represents the residual of Z . Statistical significance of the β_{zx} path coefficient suggests Z is a plausible mediator between X and Y .

$$Z = \beta_0 + \beta_{zx}X + e \quad (2)$$

Step 3: The criterion (Y) is regressed on mediator (Z) controlling for the predictor (X) to determine if a relationship ($\beta_{yz.x}$) exists between the mediator and the criterion above and beyond what the predictor (X) accounts for. Statistical significance of the $\beta_{yz.x}$ path coefficient suggests the mediation hypothesis is plausible. Whereas nonsignificance of the $\beta_{yz.x}$ path coefficient suggest that the relationships between the mediator (Z) and the criterion (Y) is spurious due to the predictor (X).

$$Y = \beta_0 + \beta_{yx.z}X + \beta_{yz.x}Z + e \quad (3)$$

Step 4: The criterion (Y) is regressed on the predictor (X), controlling for the mediator (Z ; estimated in step three). If the β_{zx} , $\beta_{yz.x}$, and $\beta_{yx.z}$ path coefficients from the first three steps are statistically and practically significant, partial mediation is plausible. That is, if the $\beta_{yx.z}$ path coefficient is significantly different from zero, part of the effect of the predictor on the criterion is indirect and part of the effect is direct. Whereas nonsignificance of the $\beta_{yx.z}$ path coefficient suggests complete mediation. If the $\beta_{yx.z}$ path coefficient is not significantly different from zero, the total effect of X on Y is indirectly through Z . As such, given a statistically significant $\beta_{yx.z}$ path coefficient controlling for Z

in step four, we reject the null hypothesis of complete mediation (i.e., path coefficient c' equals zero) and conclude partial mediation. In contrast, given a nonsignificant $\beta_{yx.z}$ path coefficient controlling for Z , we fail to reject the null hypothesis of complete mediation and, by default, conclude complete mediation.

At this point, it is important to point out one caveat to consider when hypothesizing and estimating mediation. A well-established, empirically supported, and statistically and practically significant bivariate relationship between the predictor on the criterion has been considered necessary by many researchers (Kraemer, Kiernan, Essex, & Kupfer, 2008; Little et al., 2009; Preacher & Hayes, 2004; Rose et al., 2004; Shrout & Bolger, 2002). However, consideration must be given with respect to suppression (MacKinnon et al., 2000), attenuation of bivariate relationship (X - Y) when multiple mediated paths are present (Hayes, 2009), and the distance in time and space between X and Y (Kenny et al., 1998; Shrout & Bolger, 2002). That is, if X and Y are close in time and space (i.e., proximal mediation), one would expect a significant bivariate relationship between X and Y . In contrast, as X gets further away from Y in time and space (i.e., distal mediation), the relationship between X and Y may be attenuated by other factors. Consequently, many researchers no longer require a significant bivariate relationship between X and Y (Kenny et al., 1998; MacKinnon et al., 2004; Shrout & Bolger, 2002).

Structural equation modeling model-data fit approach. Although there may be differences in estimation, specification, and testing of mediation via the causal-steps and structural equation modeling (SEM) approaches, both result in similar findings. In fact, if OLS estimation is used and we assume variables are measured without error, the path coefficient estimates from both approaches, assuming partial mediation, will be

essentially identical (Bollen & Stine, 1990). Major advantages of estimating mediation via SEM over OLS regression are the ability to formally test the completely mediated model and model error-free latent variables, which results in less biased estimates of direct and indirect effects (James, Mulaik, & Brett, 2006). In addition, the SEM approach is able to model multiple mediators, moderators, and covariates.

Contrary to the causal-steps approach, the SEM approach assumes the more parsimonious complete mediation as the baseline model (James et al., 2006). The causal-steps approach is only able to estimate partially mediated models. SEM allows for specification of either complete or partial mediation models. As a result, the Z - Y parameter will be similar for both the causal-steps and SEM approaches when partial mediation models are estimated; however, the Z - Y parameter will be different in SEM when complete mediation models are estimated. Note if a simple three-variable partial mediation model is hypothesized, the model has no degrees of freedom; consequently, it is not falsifiable in an SEM or regression framework (James et al., 2006).

Complete Mediation. If a completely mediated model is hypothesized, the SEM approach requires estimation of two path coefficients. The direct effect of the predictor (X) on the mediator (formula 2) and the direct effect of the mediator (Z) on the criterion (Y) shown below.

$$Y = \beta_0 + \beta_{yz}Z + e \quad (4)$$

Notice, when complete mediation is hypothesized, the SEM framework has a distinct advantage over OLS regression in that SEMs are able to constrain the X - Y path to zero therefore the Z - Y path coefficient controlling for X is not estimated (formula 3). Modeling mediation hypotheses via SEM allows for statistical tests of complete mediation. Because

the X - Y path is constrained to zero, the complete mediation model has one degree of freedom and a global goodness-of-fit test can be used to determine how well the model fits the data. The goodness-of-fit of the complete mediation model is essentially assessed by comparing the SEM model-reproduced X - Y correlation (formula 5 shown below) to the observed X - Y correlation. If the two correlations are not significantly different, as determined by goodness-of-fit indices, the completely mediated model is said to fit the data. That is, the entire effect of the predictor on the criterion is via the mediator.

Partial Mediation. If SEM is used to estimate a partial mediation model, the parameter estimates are similar to those estimated via the causal-steps approach. The X - Z (β_{yx} ; formula 2), the Z - Y controlling for X ($\beta_{yz.x}$), and the X - Y controlling for Z ($\beta_{yx.z}$; formula 3) path coefficients are simultaneously estimated when estimating the model via the SEM approach. Partial mediation is supported if all three path coefficients are statistically significant. However, because the model has no degrees of freedom (i.e., just identified model), global goodness-of-fit tests are not possible and thus the partial mediation model is not falsifiable.

Estimation of indirect effect. When complete mediation is estimated, the indirect effect of the predictor on the criterion via the mediator is computed as the product of the X - Z path coefficient (β_{zx}) and the Z - Y path (β_{yz}) coefficient ($\beta_{zx}\beta_{yz}$ or equivalently, ab), where \hat{r} represents the SEM model-reproduced correlation:

$$\hat{r}_{xy} = \beta_{zx}\beta_{yz} \quad (5)$$

When partial mediation is estimated, the indirect effect is computed as the product of the X - Z path coefficient (β_{zx}) and the Z - Y path coefficient controlling for X ($\beta_{yz.x}$; $\beta_{zx}\beta_{yz.x}$).

Note, for the causal-steps approach, computation of the indirect effect is only possible from a partially mediated model. As a result, if the underlying mechanism is truly completely mediated, estimating the indirect effect via the SEM approach will produce more accurate estimates of the indirect effect. For example, if the SEM standardized path coefficient between perceived importance and examinee effort is .40 ($\beta_{\text{effort importance}}$) and the SEM standardized path coefficient between examinee effort and test performance is .40 ($\beta_{\text{performance effort}}$), the standardized indirect effect would equal .16. This indirect effect would be interpreted as for every standard deviation unit change in the predictor, the criterion would change by .16 standard deviation units. The value .16 also represents the model-implied correlation between perceived importance and test performance. If the observed correlation between perceived importance and test performance was not significantly different than .16, a direct path from perceived importance to test performance would not be necessary to account for the relationship between perceived importance and test performance. However, holding all else constant, if the observed correlation between perceived importance and test performance was .36, a direct path would need to be modeled to reproduce the relationship between perceived importance and test performance (i.e., partial mediation). See Figure 5 for an illustration of this example of complete versus partial mediation.

Tests of the indirect effect. Neither the causal-steps approach nor the SEM approach explicitly test the statistical and practical significance of the indirect effect (ab). The causal-steps approach tests the a , b , c , and c' path coefficient estimates for statistical significance (see Figure 5). The SEM approach tests the a and b paths for complete mediation models, and a , b , and c' path coefficients for partially mediated models. Thus,

other analyses are needed to test the statistical significance of the indirect effect (MacKinnon et al., 2002). Considering the sampling distributions of the parameter estimates a , b , c , and c' are normally distributed (i.e., based on the central limit theorem), the use of standard errors to compute confidence intervals and statistical significance tests are straightforward and do not pose complications in a regression or SEM framework (Bollen & Stine, 1990). However, as discussed below, the product term (i.e., ab) is not always normally distributed (Bollen & Stine, 1990).

Sobel test. The most popular method to construct confidence intervals and test the indirect effect for statistical significance via a z -test is referred to as the Sobel test (Sobel, 1982). Sobel computed a standard error of the ab parameter (s_{ab}) based on the multivariate-delta method, which assumes a multivariate asymptotic normal distribution of the ab parameter. In the following formula, s_a^2 and s_b^2 represent the squared standard errors of the a and b parameters, respectively.

$$s_{ab} = \sqrt{a^2 s_b^2 + b^2 s_a^2} \quad (6)$$

Although the standard error proposed by Sobel in formula six is the most commonly used, the formula for the exact standard error derived by the multivariate-delta method can be used. Notice the only difference between the formulas is the addition of the a and b parameter variances.

$$s_{ab} = \sqrt{a^2 s_b^2 + b^2 s_a^2 + s_a^2 s_b^2} \quad (7)$$

The standard errors estimated by formula six and seven have been shown to approximate the population parameter in large samples (MacKinnon & Dwyer, 1993; MacKinnon et al., 2002; Stone & Sobel, 1990). However, the exact standard error (formula 7) tends to perform slightly better (MacKinnon, Warsi, & Dwyer, 1995). The product of the path

coefficients (ab) is then divided by its estimated standard error (s_{ab}) and compared to a normal distribution (MacKinnon et al., 2002; Sobel, 1982). Significant results from the Sobel test (ab/s_{ab}) indicate a significant indirect effect.

However, accuracy of the test is contingent on how well the sampling distribution of ab approximates a normal distribution. Unfortunately, several studies have found the distribution of ab to be asymmetrical (Lockwood & MacKinnon, 1998; MacKinnon et al., 2004; Shrout & Bolger, 2002), which poses problems in statistical significance testing and when constructing confidence intervals (Bollen & Stine, 1990; MacKinnon et al., 1995). Specifically, if ab is positive, the distribution tends to be positively skewed and vice versa (Bollen & Stine, 1990; MacKinnon et al., 2002; Preacher & Hayes, 2004). Thus, the Sobel test results in symmetric confidence intervals being constructed around the ab point estimate, when, in fact, the ab distribution is often asymmetrically distributed. Consequently, because of the (typically) positive skew, the confidence interval hangs to the left near zero. Thus, detecting a statistically significant indirect effect is underpowered, especially in smaller samples (MacKinnon et al., 2002; Preacher, Rucker, & Hayes, 2007). This non-normal distribution of ab does not tend to result in biased point estimates of ab or its standard error.

Bootstrap Resampling. On account of the typically non-normal distribution of the population indirect effect (ab) and the resulting underpowered performance of the Sobel test, bootstrap resampling has emerged as a viable alternative to the Sobel test (Bollen & Stine, 1990; Little et al., 2009; MacKinnon et al., 2004; Preacher & Hayes, 2004). Advantages of bootstrap resampling methods are that no assumptions are made with respect to shape of the sampling distribution of ab . Furthermore, bootstrapping can

accommodate more complex mediation models, where most methods are more relevant to single-mediator designs. Bootstrapping is a way of empirically estimating parameters, standard errors, and confidence intervals with nonparametric distributions.

Generally speaking, bootstrapping treats the observed data as if it were a population distribution where k independent samples of size N are drawn with replacement from the observed data, the model is fit to the independent samples, and parameter estimates (e.g., ab) are calculated for each sample. Conceptually, the bootstrap sample is to the original sample what the original sample is to the population; thus, mimicking the traditional sampling process. The mean and standard deviation of the resultant cumulative distribution of the ab estimates become the point estimate of the indirect effect (ab) and its standard error (s_{ab}). The confidence interval is then constructed from the distribution. Because the bootstrapped distribution is an unknown distribution, and usually asymmetric (Lockwood & MacKinnon, 1998; Shrout & Bolger, 2002), the percentile confidence interval method is often used. The percentile method orders the bootstrapped ab estimates from low to high. The confidence interval's lower and upper bounds are then determined by the parameter estimates corresponding with the $\alpha/2$ and $1-\alpha/2$ percentiles (e.g., $.05/2 = .025$ and $1-.05/2 = .975$). Thus, assuming a 95% confidence interval, the first 2.5% of the ab parameter estimates make up the values in the lower tail left of the lower bound, the middle 95% of the estimates make up the 95% confidence interval, and the highest 2.5% of the ab parameter estimates make up the values in the upper tail to the right of the upper bound. A statistically significant indirect effect is said to be present when this confidence interval does not contain zero as a plausible value.

The percentile confidence intervals generated from the bootstrap results tend to be asymmetrical (matching the underlying ab distribution) with good power and Type I error control (Bollen & Stine, 1990). Subsequent studies have found the percentile confidence interval method has performed well with good power and Type I error in simulation studies (MacKinnon et al., 2004; Valente, Gonzalez, Miočević, & MacKinnon, 2015; Shrout & Bolger, 2002).

A slight variation of the percentile confidence interval method is the bias-corrected bootstrap method. However, recommendations on which of the two methods to use has been mixed. High power and excessive Type I error with the bias-corrected bootstrap method has been reported (Biesanz, Falk, & Savalei, 2010; Fritz & MacKinnon, 2007; Fritz, Taylor, & MacKinnon, 2012). Other researchers have reported the bias-corrected bootstrap method has reasonable Type I error control and slightly better power than the percentile bootstrap method (e.g., Cheung, 2007; Cheung & Lau, 2008; Hayes & Scharkow, 2013). Yet some researchers have reported higher Type I error control and reasonable power from the percentile bootstrap method (Hayes & Scharkow, 2013; MacKinnon et al., 2004; Valente et al., 2015).

Distribution of the product. In response to the underpowered Sobel test and computationally-intensive bootstrap resampling approach, the distribution of the product approach was introduced as an analytical method of testing the indirect effect (MacKinnon et al., 2002; MacKinnon, Fritz, Williams, & Lockwood, 2007; Tofighi & MacKinnon, 2011). This method assumes the parameters a and b are normally distributed, an assumption that usually holds (Falk & Biesanz, 2015). However, as pointed out earlier, the distribution of the product of ab is not usually normally

distributed. Specifically, if the mediation null hypothesis is true ($H_0: a = b = 0$), the ab distribution is symmetrically distributed with excess kurtosis, but when the null is false ($a \neq 0$ and/or $b \neq 0$), ab is asymmetrically distributed with excess kurtosis (MacKinnon et al., 2002). Consequently, the symmetry of the confidence interval will change as the values of a and b change.

To analytically construct asymmetric confidence intervals, the distribution of the product method uses the product of the standardized path coefficients ($z_a = a/\sigma_a$ and $z_b = b/\sigma_b$). Then critical values of the product of the standardized path coefficients ($z_a z_b$) are obtained from tables provided by Meeker, Cornwell, and Aroian (1981). These critical values (CV) were derived from the theorized distribution of the product of two normally distributed standardized variables (Meeker et al., 1981; see MacKinnon et al., 2002 for details). Using the exact multivariate-delta standard error of ab (formula 7), computation of the confidence interval then proceeds along the lines of computing a conventional confidence interval, with the exception of different critical values for upper and lower bounds. Formulas for the upper and lower bounds for the asymmetric confidence intervals are illustrated by the formulas below (MacKinnon et al., 2004).

$$CI_{\text{Lower Bound}} = ab + \text{lower CV} * s_{ab} \quad (8)$$

$$CI_{\text{Upper Bound}} = ab + \text{upper CV} * s_{ab} \quad (9)$$

A limitation to the tables provided by Meeker and colleagues (1981) is the CVs are provided in 0.40 increments and only for a Type I error rate of .05. To facilitate easier and more accurate construction of confidence intervals (i.e., no rounding) and to allow for other Type I error rates, the PRODCLIN program was developed (MacKinnon, Fritz,

et al., 2007). Subsequently, the PRODCLIN program has been implemented into the RMediation package for use in R (R Core Team, 2016; Tofighi & MacKinnon, 2011).

A statistically significant indirect effect is said to be present when distribution of the product generated confidence interval does not contain zero as a plausible ab parameter value. Researchers suggest the distribution of the product asymmetric confidence interval method generally performs better than conventional confidence intervals via the Sobel test and percentile and bias-corrected bootstrap methods (Falk & Biesanz, 2015; Fritz et al., 2012; MacKinnon et al., 2002; MacKinnon et al., 2004; Tofighi & MacKinnon, 2011; Valente et al., 2015).

Monte Carlo. Monte Carlo has some advantages over other methods. Like bootstrap resampling, Monte Carlo makes no assumptions of the distribution of the parameter (ab). However, unlike bootstrapping, it can be used with summary statistics (raw data is not needed), it is faster than bootstrapping, and can be used with multilevel models (Preacher & Selig, 2012; Tofighi & MacKinnon, 2016).

Generally speaking, instead of fitting the model to the data k times, as done in bootstrap resampling, Monte Carlo estimation generates specific sample statistics based on the observed data parameter estimates. Using the observed parameter estimates a , b , s_a , and s_b as means and standard deviations, respectively, k (typically 1000 to 5000) pairs of a and b sample statistics are generated, the product of the pairs (ab) are computed, and the (typically 1000 to 5000) products form a distribution of ab . Confidence intervals are then constructed similar to how the bootstrap percentile confidence intervals were formed. That is, the confidence interval's lower and upper bounds are determined by the parameter estimates corresponding with the $\alpha/2$ and $1-\alpha/2$ percentiles. Thus, assuming a

95% confidence interval, the parameter estimates corresponding to the 2.5 percentile and the 97.5 percentile represent the lower and upper bounds.

Simulation studies have shown the Monte Carlo method performs very well and essentially equivalently to the distribution of the product with respect to power and Type I error control (Hayes & Scharkow, 2013, MacKinnon et al., 2004; Preacher & Selig, 2012; Valente et al., 2015). Although the bias-corrected bootstrap resampling method had more power, the Monte Carlo method had a better balance of Type I error control and power (Hayes & Scharkow, 2013; Tofighi & MacKinnon, 2016).

Summary of tests of indirect effects. Numerous methods of estimating and testing the indirect effect for statistical significance have been developed since the seminal causal-steps approach was introduced by Baron and Kenny (1986). New methods such as permutation testing (e.g., Taylor & MacKinnon, 2012), Bayesian Markov chain Monte Carlo (e.g., Biesanz et al., 2010; Yaun & MacKinnon, 2009), and posterior p -value show great promise (e.g., Biesanz et al., 2010; Falk & Biesanz, 2016), but have not been widely accepted in applied research as of yet. Further, methods such as multilevel modeling (e.g., Pituch, & Stapleton, 2011; Preacher, Zyphur, Zhang, 2010), and longitudinal modeling (e.g., Cheong, MacKinnon, & Khoo, 2003; Maxwell & Cole, 2007) have been specifically designed for more complex mediation models and research designs.

Nonetheless, the estimation methods presented are among the most popular in applied and basic literature. Despite the similarity in performance of many of the methods discussed, the methods best able to balance Type I error control and power across nearly all studies were the percentile bootstrap resampling, distribution of the product, and Monte Carlo estimation techniques. However, out of the methods discussed, the

distribution of the product method seems to be slightly favored (Hayes & Scharkow, 2013; Preacher & Selig, 2012). Furthermore, Tofighi and MacKinnon (2011) noted empirical methods (e.g., bootstrapping) should not be used when analytic solutions of testing the indirect effect are readily available (e.g., PRODCLIN: MacKinnon, Fritz, et al., 2007; RMediation: Tofighi & MacKinnon, 2011).

Research Design Characteristics to Differentiate Third Variables

As previously discussed, mediation is a theoretical specification of causal relationships rather than simply a data-analytic technique or statistical model. To make appropriate causal conclusions, a scientist must consider the validity evidence supported by the research design. Although validity of inferences is often thought of as a dichotomous decision (valid vs. not valid), one must recognize the validity of causal inferences lies on a continuum depending on several characteristics such as research design, sampling, measurement, and statistical analyses (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). Characteristics of research design most commonly discussed in mediation literature that lend to increased validity evidence or valid inferences are randomization, manipulation, and temporal precedence (Iacobucci, Saldanha, & Deng, 2007; MacKinnon et al., 2000; Spencer et al., 2005). The following paragraphs will discuss experimental design (i.e., random assignment and manipulation of variables) and temporal precedence in the context of mediation.

Experimental design. Research studies often randomly assign participants to treatment and control conditions with the intent of manipulating the criterion variable via the influence of the experimentally manipulated predictor. Because mediation specifies two causal effects ($X \rightarrow Z$ and $Z \rightarrow Y$), experimental studies of mediation require both

experimental manipulation of the predictor, independent from the mediator, and experimental manipulation of the mediator, independent from the predictor (Shrout & Bolger, 2002). Several researchers have proposed experimental designs such as causal-chain strategy that more readily enable mediation inferences (Smith, 1982; Spencer et al., 2005; Stone-Romero & Rosopa, 2008). Causal-chain designs randomize both X and Z ; thus, two separate studies or groups must be used. The first study evaluates the effect of X on Z by experimentally manipulating X . Followed by a second study that evaluates the effect of Z on Y by experimentally manipulating Z (Smith, 1982). For obvious reasons, these designs are only applicable when the predictor and mediator variables are easily experimentally manipulable. Although predictors are often manipulable, mediators are not frequently able to be experimentally manipulated in isolation (Spencer et al., 2005).

Classic experimental design. The most common experiment is performed where examinees are randomly assigned to control and intervention groups (X) with the intent of manipulating the mediator (Z). Because the mediator is related to the criterion, the criterion will be manipulated as well. Accordingly, the predictor, mediator, and criterion are all measured and mediation is estimated. Given adequate research design, the researcher is able to make causal inferences with respect to the effect of the predictor on the mediator (e.g., no confounders, no omitted variables, temporal precedence). However, the relationship between the mediator and the criterion is only correlational; thus, no causal inferences are justified. That is, the effect of the mediator on the criterion is prone to the effects of confounding variables and temporal order of occurrence.

Causal-chain design. A more robust experimental design to make causal inferences requires two separate experiments (Spencer et al., 2005). In the first

experiment, examinees are randomly assigned to control and intervention groups (experimentally manipulated X) with the intent of influencing the mediator (Z). The predictor and mediator are then measured. In the second experiment, a different group of examinees is randomly assigned to control and intervention groups (experimentally manipulated Z) with the intent of influencing the criterion (Y). The mediator and the criterion are then measured. Combining the data from both experiments, mediation is estimated. Given adequate research design, the researcher is able to make causal inferences with respect to the effect of the predictor on the mediator and the mediator on the criterion (e.g., no confounders, no omitted variables, temporal precedence).

As you can see, experimental design to strengthen mediation inferences is an arduous task, but, when properly executed, evidence of causal order of variables and unbiased estimation of the indirect effect are possible. Mediation is often examined when participants are not able to be randomly assigned to levels of *either* the predictor or the mediator (Frazier et al., 2004; Kraemer, Wilson, Fairburn, & Agras, 2002; Rose et al., 2004); thus, researchers should employ other design characteristics such as temporal precedence to strengthen mediation inferences.

Temporal precedence. A critical assumption of mediation is temporal precedence. Temporal precedence is defined as a variable (X) that occurs in time before another variable (Y ; Shadish et al., 2002). Oftentimes, temporal precedence is referred to as the order of measurement (Iacobucci et al., 2007; Wu & Zumbo, 2008). That is, the predictor is measured first, followed by the mediator, and finally the criterion. A point of emphasis with respect to temporal precedence is simply measuring the predictor before the mediator and the mediator before the criterion is *not* sufficient when making an

inference of causal ordering of events. It is the temporal relationships of the underlying constructs that is most important (James, Mulaik, & Brett, 1982). That is, the predictor (X) must occur prior to the mediator (Z) in time and space and the mediator (Z) must occur prior to the criterion (Y) in time and space. This temporal precedence may be better illustrated by an example. With respect to this study, an examinee must have some inherent level of perceived test importance (predictor) before the individual exerts effort (mediator) on the test. Further, the examinee must exert effort (mediator) before answering each test item (criterion). Thus, perceived importance (X) precedes expended effort (Z) which precedes test performance (Y) in time and space.

Although measuring these constructs in order of occurrence strengthens mediation inferences, critical assumptions of temporal precedence must be made. Namely, temporal order of measurement must align with the theoretical framework. Consider the following extreme example hypothesizing the effect of SES (X) on physical health (Y) is mediated by ability to cope with stress (Z ; Matthews, Gallo, & Taylor, 2010). If physical health (Y) is measured before ability to cope with stress (Z) and SES (X), one should not infer from the estimated relationship between physical health (Y) and SES (X) that physical health (Y) causes SES (X). Clearly, an inference that physical health (Y) causes SES (X) does not align with theory regardless of when the variables were measured. Accordingly, there are times when the nature of the variable itself (e.g., trait-like characteristics, demographics) may dictate whether measurement of the variable out of temporal order is plausible.

In the absence of sufficiently rigorous research design (e.g., random assignment to X and Z), “the researcher bears the burden of arguing the ordered relationship on logical or theoretical grounds” (Iacobucci et al., 2007, p. 140). Even then, the conclusions made

from analyses conducted on theoretically ordered variables may be untenable. Despite the importance of temporal precedence in the absence of randomization or manipulation of X and Z , researchers conducting observational mediation studies seldom collect data aligning with temporal order of occurrence (Iacobucci et al., 2007).

Summary. The validity of inferences from test scores about mediated processes is enhanced with reliable measurement, experimental design, temporal precedence, and theoretical framework suggestive of mediation. An increasing lack of these measurement and design characteristics contributes to lack of validity of inferences from scores about mediation processes. Although theory and design are necessary for mediation hypotheses, data-analytic techniques are necessary to test these hypotheses.

Assumptions of Mediation

There are several assumptions, both statistical and methodological, when mediation inferences are desired. As with any ordinary least squares (OLS) estimation, assumptions of linearity of the X - Z , X - Y , and Z - Y relationships, normality of residuals, homoscedasticity of residuals, and independence of residuals must be met if OLS estimation is employed (Cohen, Cohen, West, & Aiken, 2003). If maximum likelihood (ML) estimation is employed when estimating the model parameters via SEM, then the assumptions of linearity of the X - Z , X - Y , and Z - Y relationships, multivariate normality of continuous variables, and homoscedasticity of residuals must be met.

Measurement error. An assumption when estimating mediation by means of OLS regression is the predictors are assumed to be measured without error. Predictors measured with error result in biased unstandardized path coefficients. Criteria measured with error result in biased standardized path coefficients and decreased

statistical power (Cohen et al., 2003). Because variables studied in the social sciences are prone to measurement error (Baron & Kenny, 1986; Cohen et al., 2003), this assumption is not often met. In single-mediator mediation analyses, measurement error in the mediator results in underestimates of the effect of the predictor on the mediator (Hoyle & Kenny, 1999; Judd & Kenny, 1981a), the mediator on the criterion (Baron & Kenny, 1986; Hoyle & Kenny, 1999), and tends to overestimate the effect of the predictor on the criterion (Baron & Kenny, 1986). Consequently, the indirect effect tends to be underestimated and thus making it more difficult to detect.

Although meeting the assumption of error free measurement is not likely, remedies do exist. The first and most obvious approach is to measure X , Z , and Y reliably. When highly reliable scores cannot be produced, SEM can be employed to correct for the biasing effects of measurement error. Two approaches to modeling latent variable SEMs that correct for measurement error are use of multiple indicators of each of X , Z , and Y (MacKinnon, 2008) or use of reliability-corrected single-indicator models (Cheung & Lau, 2015; Cole & Preacher, 2014).

Model misspecification. The specification of third variables as mediators, moderators, covariates, confounders, or suppressors is a statistical and methodological concern. Considering statistical mediation analyses are little more than intercorrelations among observed variables and are not often experimentally manipulated, mediation models can easily and unknowingly be prone to misspecification.

Order and direction. With respect to misspecification, one of the most obvious and fundamental assumptions made when making mediation inferences is temporal precedence. That is, the basic conceptual mediation model posits that X causes Z , which

causes Y ($X \rightarrow Z \rightarrow Y$). However, without a properly specified model based on theory, the order and direction of interrelationships between variables could be misspecified. For example, the difference between a completely mediated effect ($X \rightarrow Z \rightarrow Y$) and a confounded effect ($X \leftarrow Z \rightarrow Y$) is simply the direction of one arrow and cannot be discerned by statistical analyses alone (MacKinnon et al., 2000; Thoemmes, 2015). Similarly, the ordering of the predictor and criterion variables ($Y \rightarrow Z \rightarrow X$) results in equivalent model-data fit. Given the example used throughout this chapter (i.e., importance \rightarrow effort \rightarrow performance), depending on theory, it may be plausible that examinee effort causes both perceived importance and test performance (i.e., importance \leftarrow effort \rightarrow performance). Alternatively, it may be plausible that test performance causes reported effort, which causes reported perceived importance (i.e., performance \rightarrow effort \rightarrow importance). Thus, in the absence of a theory specifying a pattern of relationships consistent with mediation, statistical analyses alone cannot differentiate between these statistically equivalent, but conceptually distinct examples.

Several researchers have addressed this issue of equivalent models (Hershberger, 2006; MacCallum, Wegener, Uchino, Fabrigar, 1993; Raykov & Penev, 1999). That is, for any specified model, alternate models that differ in causal order and causal directions may exist with identical model-data fit (Cole & Maxwell, 2003; Hershberger, 2006; Kenny et al., 1998; Little et al., 2009; Raykov & Marcoulides, 2001). Thus, differences in chi-square values, fit indices, p -values, and correlation residuals cannot distinguish between equivalent models (Raykov & Marcoulides, 2001). Despite the equivalence in model-data fit, equivalent models often result in different substantive interpretations. Consequently, equivalent models can be cause for concern when estimating models from

observational research, as statistical estimation cannot supplant sound theoretical foundations and empirical research when specifying models.

Omitted variables. Model misspecification resulting from missing and often unmeasured variables is referred to as omitted variables. When modeling data from observational research, unmeasured variables pose significant problems to interpretation of analyses. This problem of omitted variables can be doubly challenging with a mediation hypothesis as opposed to a more conventional hypothesis (i.e., $X \rightarrow Y$) because of the two effects being estimated. That is, an omitted third variable could account for the relationship between X and Z and/or the relationship between Z and Y . For example, in the example where perceived importance \rightarrow examinee effort \rightarrow test performance, perhaps we should have also measured examinee ability. Academic ability may have a significant positive relationship with importance, effort, and performance. Thus, the omission of ability positively biases the path coefficients because the omitted variable may account for the relationship(s) between X and Z and/or Z and Y .

The dilemma of omitted variables can often be remedied by use of experimental research designs. Randomization of participants into interventions (X) provides control over extraneous sources of variability, mitigating the effects of confounding variables. However, randomly assigning participants to the intervention (X) only mitigates the effects of omitted variables with respect to the X to Z relationship. Because it is rarely possible to randomly assign participants to levels of the mediator, the Z to Y relationship is especially susceptible to omitted variables (e.g., confounders, moderators; Pek & Hoyle, 2016). Furthermore, mediation is often used in research where manipulation or randomization of X may not be feasible or even possible (i.e., observational and quasi-

experimental research). Efforts to understand the bias of confounders using statistical methods have been proposed in situations when mediators are unable to be manipulated (e.g., sensitivity analyses: Imai, Keele, & Tingley, 2010; MacKinnon & Pirlott, 2015). Regardless, in situations where participants are not randomly assigned to levels of X or Z , sound theory and replication studies with different manipulations and controlling for different variables provide strongest evidence of mediated relationships (Holland, 1988; James & Brett 1984; MacKinnon, Cox, & Baraldi, 2012; McDonald, 1997).

Functional form. Another assumption is the correct specification of functional form. Functional form refers to the correct mathematical form of the relationships between variables. For example, mediation models typically assume the X - Z , Z - Y , and X - Y relationships are linear. Of course, any of these relationships may actually be nonlinear which should be assessed and modeled appropriately. If necessary polynomial terms (e.g., quadric, cubic) are not included in the model, this is an issue of omitted variables and thus the same consequences are observed.

Moreover, mediation assumes the relationship of $X \rightarrow Z \rightarrow Y$ is both linear and additive (Fairchild & MacKinnon, 2009; Holland, 1988; Judd & Kenny, 1981; MacKinnon et al., 2007).³ An interaction (non-additive effect) is said to exist if the effect of Z on Y depends on the level of X or the effect of X on Y depends on the level of Z . For example, it is plausible that the effect of examinee effort on test performance depends on the level of examinees' perceived importance. If this moderated effect exists, it should be

³ Of note, methods of mediation including an interaction term have been proposed (e.g., MacArthur approach, Kraemer et al., 2001); however, these methods are not nearly as popular as more traditional approaches.

modeled and not assumed to be nil. Fortunately, the X - Z interaction term can be easily tested and if significant, this effect, not an indirect effect, should be plotted and probed.

Current Measurement of Motivation

Throughout this chapter, I provided a discussion of what mediation is, why it is important to theory and application, and how to provide evidence of mediation. In the remaining portion of this chapter, I discuss the potential issues and implications of prospective and retrospective measurement of perceived importance and examinee effort.

As you recall from the introduction, motivation is operationalized by the SOS as two factors of examinees' perceived test importance and perceived expended effort. As such, according to EV theory, perceived test importance affects low-stakes test performance via the mediator, examinees' perceived expended effort ($I \rightarrow E \rightarrow P$). Herein lies the potential obstacle in inferring an indirect effect of perceived importance on test performance via examinee effort. An assumption of mediation is temporal precedence of the underlying phenomenon. EV theory suggests temporal precedence in that perceived importance (i.e., value) affects examinee effort (i.e., motivation) which affects test performance (Cole, 2007; Cole et al., 2008; Finney, Sundre, et al., 2016; Mathers et al., 2016; Myers et al., 2016; Penk & Schipolowski, 2015; Zilberberg et al., 2014). However, previous research modeling this indirect effect has regularly measured both importance and effort retrospectively after the students have already completed the test (i.e., $P \rightarrow I$, $P \rightarrow E$; Cole, 2007; Cole et al., 2008; Finney et al., 2016; Mathers et al., 2016; Myers et al., 2016; Zilberberg et al., 2014). Although temporal precedence is not necessarily violated by measuring the constructs out of order, this practice leads to questions regarding the implications of prospective versus retrospective measurement of

perceived importance and examinee effort; thus, prompting the current study.

Nonetheless, existing theory and study in the domain of retrospective and prospective measurement of noncognitive constructs helps shed light on the potential implications of the measurement of perceived importance and examinee effort.

Retrospective and Prospective Measurement

Self-report assessments are by far the most popular method of measuring affective examinee characteristics. In contrast to more objective measures such as RTE or galvanic skin response for anxiety, examinees' subjective self-reports have considerable potential of becoming contaminated by sources irrelevant to the construct of interest (i.e., construct-irrelevant variance). Inherent with all self-report measures are potential sources of contamination such as subtle wording idiosyncrasies, inaccurate perception of self, rating scales, social desirability, and acquiescence (Schwarz, 1999; Shadish et al., 2002). When the measure is completed relative to test completion is another potential source of contamination (Hill & Betz, 2005). Researchers have demonstrated examinees' self-reports of performance-related emotions such as anxiety and anger change as a function of when the variables are measured relative to test completion (Folkman & Lazarus, 1985; Goetz, Preckel, Pekrun & Hall, 2007).

Retrospective measurement. Measuring test performance-related variables after examinees complete a test is termed retrospective measurement. When researchers use retrospective measurement methodology, they may be interested in examinees' test anxiety or expended effort *during* the test. Or, perhaps they are gathering information such as perceived importance after the test and using it as a proxy for examinees' pre-test levels of perceived importance. For example, consider the examinees from the example

throughout this chapter. After completing the test, examinees then complete a self-report scale measuring their *expended* effort on the test. Clearly, an examinee expended a certain level of effort completing the test. Examinees reflect back on the effort they expended and respond to effort items accordingly.

Prospective measurement. In contrast, measuring test performance-related variables before examinees complete a test is termed prospective measurement. When researchers use prospective measurement methodology, they may be interested in examinees' *current* interest in an academic subject, perceived ability, anxiety, or self-efficacy (e.g., Freund & Holling, 2011). Prospective measurement is commonly used in academic settings such as classrooms where long-term academic achievement is the criterion (e.g., Hong & Peng, 2008). Prospective measurement can also be used in testing contexts. For example, consider the examinees from the example throughout this chapter. Prior to completing the test, examinees complete a self-report scale measuring their *intended* effort on the test. An examinee likely intends to expend a certain level of effort throughout the test. Examinees consider the effort they intend to expend and respond to effort items accordingly (Higgins, 1997).

Type of Measurement Influences Item Response Patterns

Prospective and retrospective measurement may result in examinees responding differently to self-report measures or the test. First, I discuss some potential effects a test can have on retrospective self-report measures. Second, I discuss potential effects the prospective self-report measures can have on test performance.

Attributional bias. Attributions are used by individuals to explain causes of their behaviors and outcomes (Weiner, 1985, 2000). For example, individuals may attribute

success or failure on a task to their ability, effort, or luck. When asked to self-reflect on a previous behavior, attitude, or belief, ones' retrospective perception of their previous state may be contaminated by attributions of their performance on a task (Mezulis, Abramson, Hyde, & Hankin, 2004; Perry et al., 2008). For example, Avery did not perform very well on a statistics test. Subsequently, Avery attributed the cause of her poor performance to not trying hard on the test. Her retrospective self-report of not trying hard was a function of her poor test performance. Although not explicitly referring to retrospective measurement, the following excerpt from C. Lloyd Morgan's autobiography sums up one's ability to self-reflect on past attitudes, beliefs, and behaviors:

Herein lies a difficulty in any autobiographical sketch which purports to deal with one's mental development. It is a story of oneself in the past, read in the light of one's present self. There is much supplementary inference—often erroneous inference—wherein 'must have been' masquerades as 'was so.' (Morgan, 1932, p. 237-238)

That is, individuals' retrospectively measured self-reports may be functions of their selective attention and perceptions of reality.

Researchers have argued this attributional bias may be a result of individuals attempting to present themselves favorably to others (Arkin, Appelman, & Burger, 1980; Weary, 1979). However, results from a study comparing attributions between a group believing their test results were private and a group believing their test results were public indicated no differences in attributions (Greenberg, Pyszczynski, & Solomon, 1982). Thus, self-serving attributional bias may be individuals' perceptions of the cause of their behavior. Researchers have proposed attributional bias may be a self-serving bias that

functions as a self-protective mechanism (Pekrun et al., 2004; Perry et al., 2008; Thompson, 1996). This phenomenon is self-serving in that individuals systematically distort their self-reported beliefs, attitudes, and behaviors to augment their self-esteem. This phenomenon is a bias because it is not random variability in self-reports, but differential attributions for poor versus good performance. Said another way, examinees who perform relatively poorly may attribute their performance to lack of ability or effort. In contrast, examinees who perform well have no reason to respond inaccurately.

Importantly, results from studies in academic settings indicate expended effort tends to be the most common attribution individuals report to explain task performance (Cheng & Chiou, 2010; Dong et al, 2013; Perry et al., 2008; Weiner, 1985). Personal factors such as anxiety, ability, and test-taking strategy, and situational factors such as test difficulty, quality of professor, and luck are common attributes used to explain academic performance (Cheng & Chiou, 2010; Perry, et al., 2008; Smith, Snyder, & Handelsman, 1982). In applied testing contexts, attributional bias has been posited as a source of systematic error when academic performance-related variables were retrospectively measured (Eklöf, 2007; Folkman & Lazarus, 1985; Hill & Betz, 2005; Zilberberg et al., 2014).

Behavioral commitment. Behavioral commitment has been defined as “those consequences of the initial pursuit of a line of action which constrain the actor to continue that line of action” (Johnson, 1973, p. 397). In other words, once individuals commit to a behavior, they feel a necessity to remain committed to that behavior to avoid discrepancy between their behavioral intentions and their behavior (Carver & Scheier, 2000; Higgins, 1997). For example, Avery says she is going to try hard on her statistics

test; therefore, Avery is committed to a behavior (trying hard) that she must complete. Individuals may feel more obligated to commit to a behavior when bystanders are aware of their commitment; thus, individuals feel some normative expectations to remain committed to their behavior. If Avery told her teacher she was going to try hard on the statistics test, she may feel as if she has to remain committed because her teacher expects her to try hard. Alternatively, individuals' may remain committed to a behavior because of their personal attitudes and beliefs. For example, Avery decides to remain committed to trying hard on the statistics test because she enjoys statistics. Thus, individuals intend to remain committed to a behavior to avoid conflict between their actual behavior and their attitudes, beliefs, and normative expectations (Carver & Scheier, 2000; Gollwitzer, 1993; Higgins, 1997; Schwartz & Tessler, 1972).

Goal setting is thought of as the process where individuals establish a goal based on consideration of their interests, attitudes, and beliefs (Gollwitzer, 1993, 1999; Higgins, 1997). Goals can be conceptualized as individuals' expected level of performance toward an outcome (Garland, 1985; Gollwitzer, 1999). For example, Avery implicitly or explicitly states, "I intend to put forth put forth considerable effort on the statistics test." Next, individuals translate their intended goals into goal-directed behavior to achieve their expected level of performance (Boekaerts, 2002; Gollwitzer, 1993). Thus, Avery's behavioral intention specifies a course of action necessary to attain her goal of putting forth considerable effort on the statistics test. Further, goals are said to influence behavior indirectly via individuals' motivation (Boekaerts, 2002; Locke & Latham, 1990, 2002; Schunk, 1995). Given motivation can be defined as a process where behavior is initiated and sustained (Schunk, Pintrich, & Meece, 2008), behavioral intentions (i.e., goals)

influence individuals to expend and sustain the effort necessary to perform a task. This indirect effect of goal intentions on behavior via motivation has been demonstrated in academic classroom settings (Lee, Lee, & Bong, 2014). Several researchers have demonstrated behavioral intentions to be highly predictive of individuals' future behavior and task performance (e.g., Azjen, 1991; Boekaerts, 2002; Gollwitzer & Sheeran, 2006; Tanaka & Yamauchi, 2001; Vassiou, Mouratidis, Andreou, & Kafetsios, 2016). In sum, individuals' attitudes and beliefs influence goal intentions, which influence motivation, which influence behavior.

Pulling together these concepts and applying to a low-stakes testing contexts, prior to starting a test, examinees first complete a measure assessing their perceived test importance and *intended* effort on a test. Examinees will respond to the perceived importance items according to their attitudes, beliefs, and normative expectations about the test. As follows, examinees consider their perceptions of test importance when responding to the intended effort items. Intended effort may be conceptualized as examinees' behavioral intentions for completing the test (Boekaerts, 2002). Consequently, examinees engage in test-taking behavior that aligns with their specified behavioral intentions (Haggard & Clark, 2003; Johansson, Hall, Sikström, & Olsson, 2005). If this were reality, then self-reported intended effort and expended effort would be essentially equivalent. In short, examinees who tend to report higher levels of intended effort are more likely to expend higher levels of effort throughout the test.

Experience limitation. An important consideration when prospectively measuring task performance-related variables is an individual's experience with the task (e.g., cognitive demand, enjoyment of the task). When a construct or behavior is

measured prospectively without providing individuals with relevant information about the task, an individual's conceptualization of the task may be limited due to inexperience with the task. By measuring constructs retrospectively, individuals' experience with the task will provide for less ambiguity and, consequently, a more accurate conceptualization of the task (Aiken & West, 1990; Boekaerts, 2002; Folkman & Lazarus, 1985; Freund & Holling, 2011). Moreover, after experiencing the task, examinees should have a better conceptualization of the range of the construct's lower and upper bounds (Howard, Dailey, & Gulanick, 1979).

Consider the example used throughout this chapter. If examinees were administered the perceived test importance measure prior to the test (e.g., "This is an important test to me"), it may be difficult for examinees to conceptualize their level of perceived test importance without having any knowledge of the test. As a result, the accuracy of self-reported perceived importance would seemingly increase with examinees' test experience. This experience limitation may be especially problematic in low-stakes, institutional accountability testing where examinees are often unaware of the construct to be assessed or the cognitive demand of the test.

Despite prospective measurement's limitation of examinees' inadequate conceptualization of the construct being measured, researchers have options available to mitigate the effect of inexperience on prospective self-reports. One straightforward approach to providing examinees with contextual information is by displaying sample test items and/or informational cues prior to administration of prospective measures (Ackerman & Wolman, 2007; Aiken & West, 1990; Howard et al., 1979). In a replication study comparing prospective versus retrospective measurement, the accuracy and

stability of self-reported performance-related characteristics (i.e., self-efficacy, self-concept) were substantially improved when explicit descriptions of tests were given instead of brief descriptions (Ackerman & Wolman, 2007). In a prospective measurement study, performance-related characteristics such as interest, perceived ability, and anxiety were more predictive of test performance after examinees had experience with the test ($R^2 = .53$) than when examinees had no experience with the test ($R^2 = .21$; Freund & Holling, 2011). Given these findings, detailed test instructions and an example item will be provided to the examinees to buffer the effect of experience limitation on prospectively measured perceived importance and intended effort.

Implications of retrospective versus prospective measurement. To briefly review, the current practice of retrospectively measuring perceived test importance and examinee effort may result in contaminated perceived importance and examinee effort scores due to attributional bias. If so, we would expect to see different average levels of self-reported perceived importance and examinee effort from before to after the test. However, it is unknown whether the rank-order of perceived importance and intended effort scores measured before the test are different from the rank-order of perceived importance and expended effort scores measured after the test. If the rank-order of those scores are the same from before to after the test, then the relationship between perceived importance, examinee effort, and test performance will remain unchanged. Accordingly, the magnitude of the indirect effect will remain unchanged.

Given the typical practice of retrospective measurement of importance and effort, researchers are unable to determine if examinees' perceived importance or intended effort measured before the test are predictive of test performance. If examinees attribute the

cause of their test performance to perceived importance or expended effort, then the relationships between importance, effort, and test performance are a function of test performance causing their responses to the importance and effort measures. Thus, these relationships may be larger than the relationships when importance and effort are measured prospectively. With the typical retrospective measurement practice, researchers are unable to determine if motivation interventions aimed at influencing examinees' pretest perceived importance are in fact influencing examinees pretest perceived importance (e.g., Finney, Sundre, et al., 2016). If perceived importance scores collected after the test are influenced by test performance, then the effect of motivation interventions on perceived importance is unknown, which may explain the differential results from these interventions (e.g., Finney, Sundre, et al., 2016; Liu et al., 2015). Researchers will also be able to determine if examinees responding to prospective measures of perceived importance and intended effort will exhibit behavior aligning with their intended effort (behavioral commitment) and thus result in more accurate estimates of examinee ability.

Chapter III

Methods

Participants

Data for the current study were collected at a mid-sized, mid-Atlantic university that utilizes a large-scale, institutional accountability assessment program (“assessment day”). Test scores from assessment day are used for state and national accountability reporting. Moreover, assessment results are used to inform evidence-based changes to academic programming with the intent of increasing student learning. The tests administered during assessment day are low-stakes to the students in that there are no personal consequences for performance on the tests. For example, test results do not count toward course grades or graduation.

All students at the university engage in assessment day prior to the first day of class. Based on their student identification numbers, students are randomly assigned to testing rooms where they complete noncognitive and cognitive tests. The samples for this study are random and do represent the university population in terms of gender, ethnicity, age, and SAT verbal and math scores (see Table 3). For the current study, in fall, 2016, a random sample of first-year students were assigned to one of two testing rooms for the combined and prospective conditions. To obtain the sample size necessary for data analyses across three conditions, retrospective data were collected from a random sample of first-year students assigned to one of three testing rooms during fall, 2015. With the exception of the experimental manipulation of measurement condition, all procedures, proctor training, measures, and test instructions were equivalent during both assessment days.

The initial sample consisted of 1518 first-year students: $n = 697$ in the retrospective condition, $n = 474$ in the combined condition, and $n = 347$ in the prospective condition. However, the effective sample size was reduced to 1145 for three reasons. First, 5.67% of the examinees did not have complete data on the NW9 and SOS measures, resulting in 1432 students with complete data. Second, because unbalanced sample sizes across conditions can lead to lack of power to detect model noninvariance (Chen, 2007), a random sample of 400 examinees were selected from the retrospective condition to balance the sample sizes across conditions. Third, prior to conducting data analyses, data were screened for univariate and multivariate outliers using SPSS 23, SAS 9.4, and LISREL 9.20 (Jöreskog & Sörbom, 2015). Univariate outliers were identified by examining boxplots for extreme scores greater than or less than two standard deviations from the mean across each variable. Multivariate outliers were identified by computing Mahalanobis distances, the distance of a set of scores for a case from the multivariate centroid. A break in the Mahalanobis distance values indicates possible multivariate outliers. Cases identified as both univariate and multivariate outliers were examined to assess for invalid responses. Two participants from the combined condition and seven participants from the prospective condition were identified as providing nonsensical response sets (e.g., 1, 2, 3, 4, 5, 4, 3, 2, 1) and were removed from the dataset.

The effective sample sizes per condition were $n = 400$ in the retrospective condition, $n = 437$ in the combined condition, and $n = 308$ in the prospective condition. The demographics characteristics by condition were relatively consistent across conditions and representative of the university's student population (see Table 3).

Measures

Quantitative and scientific reasoning. Performance was assessed using the Natural World Test, version 9 (NW9; Sundre, Thelk, & Wigtil, 2008; Sundre & Thelk, 2010). This test was developed by math and science faculty to assess quantitative and scientific reasoning student-learning objectives from the university's general education program. The NW9 is a multiple-choice test consisting of 66 dichotomously-scored items that takes students approximately one hour to complete. The items were summed to create one total score.

In previous administrations of the NW9, examinees, on average, correctly responded to approximately 70% of the items and approximately 95% of examinees scored between 46% and 89% on the NW9 after filtering out test scores from amotivated examinees. Thus, given the length and typical scores on the test, the NW9 tends to be relatively difficult for first-year examinees. Internal consistency estimates of the NW9 scores for the current sample were adequate across conditions (see Table 4), which is consistent with previous research utilizing the NW9 (Finney, Sundre et al., 2016).

Test importance and examinee effort. The Student Opinion Scale (SOS) was developed to operationalize the task value and motivation components from EV theory (Eccles et al., 1983). The SOS consists of two subscales: 5 items measuring perceived test importance and 5 items measuring expended effort. Examinees respond to items using a Likert scale ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree"). Two items from each subscale are negatively worded and must be reverse scored prior to summing to create subscale scores. The two subscale scores can range from 5 to 25 with higher scores representing higher levels of importance and effort.

The SOS was originally created to assess perceived importance and examinee effort at the end of a testing session (Sundre & Moore, 2002; Thelk et al., 2009). The SOS has been adapted to assess test-specific perceived importance and examinee effort after test completion (see Appendix A for the retrospective SOS; Finney, Mathers, & Myers, 2016). The properties of the SOS scores when gathered after a testing session or specific test have been studied extensively in low-stakes testing contexts (see Sessoms & Finney, 2015). The two-factor structure of the test session-specific SOS has been demonstrated across first-year and upper-class students, gender, and testing medium (i.e., paper-and-pencil and computer-based; Thelk et al., 2009). The two-factor structure of the test-specific SOS has been demonstrated across test-instruction conditions designed to increase the relevance of the test to examinees (Finney, Mathers, & Myers, 2016). Internal consistency of both forms of the SOS has ranged between .79 and .89. Validity evidence has been found in the form of the relationship with test performance ($r = .34$; Thelk et al., 2009) and RTE ($r = .54$; Wise & Kong, 2005).

Because the SOS items were written to be administered retrospectively (e.g., “I gave my best effort on this test”), perceived importance and examinee effort items were modified to be administered before test completion (see Appendix B and C for prospective items). Perceived importance items were easily modified and the meaning of the items remained the same from before to after test completion. For example, the item “Doing well on this test was important to me” was changed to “Doing well on this test is important to me;” thus, both versions of the item represent perceived importance of the test. However, modifying examinee expended effort items necessitated change in the meaning of the item from *expended* to *intended* effort. For example, the item “I gave my

best effort on this test” clearly reflects past behavior, whereas the modified version of the item “I will give my best effort on this test” clearly reflects future behavior. Development of prospective items focused on maintaining nearly identical wording between prospective and retrospective versions. I performed multiple think-alouds to ensure the intended meaning and readability of the prospective items were adequate. Internal consistency estimates for the current sample were adequate for both measures of perceived importance and examinee effort across conditions (see Table 4).

Procedures

At the beginning of the testing session across all three conditions, examinees were shown an instructional video explaining the purpose of accountability testing. After the instructional video, trained proctors read scripted messages to prepare the examinees for the testing session. The proctors remained conspicuous throughout the session, monitored the room for amotivated or disruptive behavior, and answered examinees’ questions.

Retrospective condition. In the retrospective condition, proctors read aloud the standardized test instructions for the NW9 and the retrospective test-specific SOS (see Appendix D). Examinees then completed the NW9 test followed by the retrospective test-specific SOS. Examinees had one hour to complete the NW9 test and retrospective SOS.

Combined condition. In the combined and prospective conditions, examinees were asked to provide prospective ratings of intended effort and/or perceived importance. Because it may be difficult for examinees to conceptualize their perceived importance or intended effort without contextual information about the test they are about to complete, test instructions from the retrospective condition were supplemented with a detailed

description of the NW9. An example item representative of the type and difficulty of the NW9 items was included in the test instructions and on the prospective SOS.

In the combined condition, proctors read aloud the standardized test instructions for the NW9 and prospective SOS (see Appendix E). Given examinees needed to provide prospective ratings of test importance, proctors projected the example NW9 item at the front of the room prior to students completing the SOS. Proctors paused while reading the test instructions to allow examinees to read and formulate an answer to the example item. Examinees then completed the prospective perceived importance measure (see Appendix B). While completed scantrons were collected, the lead proctor read the NW9 test instructions aloud to the examinees (see Appendix F). Examinees then completed the NW9 test followed by the retrospective test-specific SOS within one hour.

Prospective condition. In the prospective condition, proctors read aloud the standardized test instructions for the NW9 and prospective SOS (see Appendix G). Proctors projected the example NW9 item at the front of the room prior to students completing the prospective SOS. After pausing to allow examinees to read and formulate an answer to the example item, examinees were provided the correct answer by the proctor. Examinees then completed the prospective perceived importance and intended effort measure (see Appendix C). While scantrons were collected, the lead proctor read the NW9 test instructions aloud (see Appendix F). Examinees then completed the NW9 test followed by the retrospective test-specific SOS within one hour.

Analytic Approach

Model-data fit. Several latent variable models were specified to answer the research questions (see Table 2). Alignment between these hypothesized models and the

data was evaluated via global and local fit indices. Global fit indices describe overall model-data fit, whereas local fit indices indicate specific areas of model-data misfit. The χ^2 test, the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) global fit indices were used. The χ^2 goodness-of-fit test is an absolute fit index that tests the discrepancy between observed and model-implied covariance matrices and means. Because the χ^2 test evaluates the strict hypothesis of perfect fit, the χ^2 test was supplemented with approximate model-data fit indices intended to assess fit on a continuum.

The CFI provides an estimate of fit of the specified model relative to a baseline model that specifies all variables are uncorrelated and means are freely estimated. Low CFI values may be due to weak correlations among the variables, because the baseline model can fit these data well. CFI values range from 0 to 1, with larger values indicating better model-data fit. The CFI tends to be sensitive to misspecified factor pattern coefficients (Hu & Bentler, 1998). The RMSEA provides an estimate of misfit due to model misspecification controlling for sampling error. It can be interpreted as the amount of misfit per degree of freedom. RMSEA values range from 0 to 1, with smaller values indicating better model-data fit. The RMSEA tends to be sensitive to misspecified factor pattern coefficients. The SRMR is essentially the average difference between the observed and the model-implied relationships on a correlation metric. The SRMR tends to be sensitive to misspecified latent factor correlations. SRMR values range from 0 to 1, with smaller values indicating better model-data fit.

Suggested fit index values indicating good model fit (approximately, CFI \geq .95, RMSEA \leq .06, and SRMR \leq .08) were considered (Hu & Bentler, 1999; Yu & Muthén,

2002). However, fit indices are intended to be used to assess model-data fit on a continuum, not to make dichotomous decisions about model-data fit (Marsh, Hau, & Wen, 2004). Accordingly, global fit indices were supplemented with evaluation of local misfit to guide evaluation of model-data fit. Local misfit was assessed by examining correlation residuals, which are the differences between observed correlations and model-implied correlations. Correlation residual values greater than $|.15|$ suggested misfit and prompted reevaluation of the model. When modeling means, local misfit was assessed by examining unstandardized mean residuals, which are the differences between observed means and model-implied means.

Nested model comparisons. Nested model comparisons were conducted by estimating the more complex (less constrained) model and the simpler (more constrained) model. The difference in χ^2 test values ($\Delta\chi^2$) and degrees of freedom (Δdf) were computed. The $\Delta\chi^2$ values were compared against critical values from χ^2 distributions with Δdf . A significant $\Delta\chi^2$ test indicated the more constrained model fit significantly worse than the less constrained model. Given the $\Delta\chi^2$ test is sensitive to large sample sizes (Cheung & Rensvold, 2002), practical significance of the difference between nested models was assessed by evaluating the change in CFI (ΔCFI). $\Delta CFI > .01$ indicated the more constrained model fit practically significantly worse than the less constrained model (Chen, 2007; Cheung & Rensvold, 2002). Because the ΔCFI and the $\Delta\chi^2$ test may provide conflicting information (French & Finch, 2006), correlation and mean residuals $>|.15|$ were examined to supplement evaluation of model-data fit.

Research questions 1 and 3. Recall, research questions one and three focus on latent mean differences in perceived importance and examinee effort scores across

conditions. Research question one examines mean differences between retrospectively gathered importance and effort across conditions whereas research question three examines mean differences between prospectively and retrospectively measured importance and effort across conditions. Prior to assessing cross-condition mean differences, measurement invariance was tested using a structured means model (SMM) to determine the cross-condition invariance of parameters (e.g., pattern coefficients).

The first SMM test was configural invariance, which indicates if the same number of factors are present across conditions and if the same items are indicators for their respective factors across conditions. Two competing factor models were specified to determine if the interrelationships could be best represented by a one-factor model representing the broader construct of motivation or a correlated two-factor model of perceived importance and examinee effort.

Single-condition one-factor and two-factor models were fit to the data from each condition to identify condition-specific sources of misfit. The latent factors were scaled by constraining the unstandardized factor pattern coefficient of one indicator to a value of one. This scaling technique set the metric of the latent factor to the metric of the respective item (i.e., 1 to 5) and was used for all subsequent CFA models⁴. Configural invariance was established if the same model fit the data adequately across data from independent conditions. Next, a multiple-condition configural model with no equality constraints was fit to data across all conditions to obtain baseline global fit indices. The

⁴ Importance item one and effort item two were chosen as the referent variables (unstandardized factor pattern coefficients constrained to one) for the two-factor models. Because these pattern coefficients are not estimated, it is important to verify the coefficients are, themselves, invariant across conditions. This was checked by setting the factor pattern coefficient associated with item one and item two to be invariant across conditions when each of the other items served as the referent variable in the metric and scalar invariance models (Rensvold & Cheung, 2001).

χ^2 statistic and degrees of freedom of this multiple-condition model were the sum of the three individual χ^2 statistics and degrees of freedom of the models fit to each condition.

Given adequate fit of the configural model, a metric invariance model was fit to the data. Testing metric invariance involved constraining the unstandardized factor pattern coefficients to be equal across the three independent conditions. Establishing metric invariance indicates the items are equally salient to their respective latent factors across conditions. At least partial, but preferably full, metric invariance is required prior to assessing the invariance of the completely mediated models across conditions.

If the metric invariance model did not fit statistically significantly worse than the configural model (i.e., metric invariance is established), then scalar invariance is assessed. The scalar invariant model was estimated by constraining the unstandardized factor pattern coefficients and item intercepts to be equal across the three test instruction conditions. Establishing scalar invariance indicates average differences in observed score are attributable to different average levels of the respective latent factors; thus, latent factor means can be compared across conditions. At least partial, but preferably full, scalar invariance is required prior to examining latent mean differences.

Given satisfactory fit of the scalar invariance model (i.e., scalar invariance is established), a scalar and measurement error invariant model was fit to the data. The measurement error invariance model was estimated by constraining the unstandardized factor pattern coefficients, item intercepts, and error variances to be equal across the three conditions. Given configural, metric, and scalar invariance, measurement error invariance indicates that any average differences in observed scores are not due to differences in measurement error across conditions. Thus, measurement error invariance is of practical

importance because it allows for comparisons of average observed score differences across conditions (Millsap & Meredith, 2007).⁵

Latent mean differences in perceived importance and examinee effort across conditions were estimated in the scalar invariance model by constraining the retrospective condition's latent mean to zero and estimating the combined and prospective condition's latent mean differences from the retrospective condition. Given three conditions, the model was re-estimated to produce the latent mean difference between the combined and prospective conditions by constraining the latent mean of the combined condition to zero. The unstandardized latent mean difference is on the metric of the items (i.e., 1 to 5). Practical significance was assessed via a standardized latent mean effect size (latent d):

$$\text{Latent } d = \frac{\kappa}{\sqrt{\Psi_{pooled}}}, \quad (1)$$

where κ is the unstandardized latent mean difference between two conditions and Ψ_{pooled} is the pooled latent variance. The pooled latent variance is:

$$\sqrt{\Psi_{pooled}} = \sqrt{\frac{n_1\Psi_1 + n_2\Psi_2}{n_1 + n_2}}, \quad (2)$$

where n_1 and n_2 are the sample sizes from independent conditions and Ψ_1 and Ψ_2 are latent variances from conditions.⁶

⁵ Note, configural, metric, scalar, and measurement error invariance is a common assumption made in many statistical analyses of observed means (e.g., t -tests and analysis of variance).

⁶ The formulas used to compute the standardized latent mean difference effect size are analogous to the formulas used to compute the standardized observed mean difference effect size.

$$\text{Cohen's } d = \frac{M_2 - M_1}{SD_{pooled}} \quad \text{and} \quad SD_{pooled} = \sqrt{\frac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{(n_1 + n_2 - 2)}}$$

Research questions 2 and 4. Recall, research questions two and four focus on the stability and magnitude of the indirect effect of perceived importance on test performance via examinee effort across conditions. Research question two asks if the indirect effect of perceived importance on test performance via examinee effort is stable when modeling retrospectively measured importance and effort scores. In contrast, research question four asks if the indirect effect of importance on performance via effort is stable when modeling prospectively and retrospectively measured importance and effort scores. The indirect effect of perceived importance on test performance via examinee effort was assessed across independent conditions via multiple-condition moderated mediation analyses to answer both research questions.

Given the number of parameters estimated in a SEM that incorporates the measurement models associated with the importance, effort, and performance along with the structural components of the model (direct effects and disturbance terms), coupled with the (small) within-condition sample size in the current study ($n = 400, 437, \text{ and } 308$), a single-indicator latent variable approach was used to mitigate the effects of random measurement error (Cheung & Lau, 2015; DeShon, 1998). The single-indicator latent variable approach involved the five steps (Cheung & Lau, 2015). First, composite scores of importance, effort, and NW9 were computed by summing the scored items. Second, the proportion of the composite scores' variance attributed to measurement error was calculated by multiplying the proportion of error variability in the scores ($1 - \text{reliability estimate}$) by each composite variable's respective variance ($[1 - r_{xx}] * s_x^2$). Third, the observed composite scores were used as single indicators of the respective latent variables (see Figure 9). Fourth, the paths from the latent factors to their respective

single indicators were fixed to one. Fifth, the proportions of reliable variance from step two for each composite score were fixed as the unstandardized error variance of the respective indicators. Thus, essentially error free composite variances were modeled.

Because the conditions are categorical in nature, moderated mediation analyses was performed via the multiple-condition approach (Marsh, Wen, Hau, & Nagengast, 2013) to assess the invariance of the indirect effect. The first step involved fitting the completely mediated model to the data within each independent condition. Given the completely mediated model fit the data adequately within each conditions, a multiple-condition model fit to all three conditions with no constraints across conditions was estimated to obtain baseline fit indices. In step two, the two unstandardized path coefficients ($I \rightarrow E$ and $E \rightarrow P$) were constrained to be equal across the three independent conditions. Given that constraining the two direct paths to be equal across the conditions did not result in significantly worse fit than the unconstrained model, the two unstandardized disturbance terms (i.e., unexplained variances associated with importance and effort) were constrained be equal across independent conditions. If the fully constrained (unstandardized path coefficient and disturbance term) model did not fit significantly worse than the constrained path coefficient model, then examinee effort and NW9 performance are explained equivalently well across conditions.

Given the completely mediated models of perceived importance on test performance via examinee effort fit the data for each independent condition, the indirect effect was assessed for statistical and practical significance. Statistical significance of the indirect effect was tested via the distribution of the product method (Tofighi & MacKinnon, 2011). Practical significance of the indirect effect was assessed via the

following crude guidelines: .01, .09, and .25 for small, medium, and large effect sizes, respectively (Kenny, 2016; Kenny & Judd, 2014).

Research question 5. Research question five focuses on latent mean change between prospectively measured perceived test importance and examinee effort scores and retrospectively measured importance and effort scores within the prospective condition. Prior to assessing repeated-measures latent mean differences, the longitudinal invariance of model parameters (e.g., factor pattern coefficients and item intercepts) was assessed using longitudinal mean and covariance structure analysis (LMACS).

Configural invariance was assessed by fitting the theorized two-factor model to prospective importance and effort scores and to retrospective importance and effort scores independently. Configural invariance indicates no change in the number of factors or which items are indicators for each factor from before to after test completion.

Given adequate model-data fit of the two-factor model across time points, a baseline configural model with no equality constraints was fit to importance and effort items from both time points simultaneously (i.e., 20 items) to obtain baseline global fit indices. Item error variances from time one and time two were allowed to covary (e.g., effort item 1 at time 1 and effort item 1 at time 2) to account for shared error variance between corresponding items across time points. The latent factors were scaled by constraining the unstandardized factor pattern coefficients of one indicator from each

factor to one across time. This scaling technique set the metric of the latent factor to the metric of the respective indicator.⁷

Given adequate fit of the configural model (i.e., configural invariance), a metric invariance model was fit to the data. Testing metric invariance involves constraining the unstandardized factor pattern coefficients of corresponding indicators across time points to be equal. Establishing metric invariance indicates corresponding items are equally salient to their respective latent factors across time points. Given configural and metric invariance are established, rank-order stability (i.e., test-retest reliability) of scores across time points can be assessed via the latent correlation between prospective and retrospective scores. High correlations indicate examinees stay in the same relative rank-order of scores across time points and low correlations indicate examinees change their relative rank-order of scores across time.

If the metric invariance model did not fit statistically significantly worse than the configural model (i.e., metric invariance is established), then scalar invariance was assessed. The scalar invariance model was estimated by constraining the unstandardized pattern coefficients and item intercepts of corresponding indicators to be equal across time points. Establishing scalar invariance indicates average differences in observed scores across the two times are attributable to average differences in the respective latent factors across time; thus, average latent change in effort and importance were interpreted.

⁷ Importance item one and effort item two were chosen as the referent variables (unstandardized factor pattern coefficients constrained to one) for the two-factor model. Because these factor pattern coefficients are not estimated, it is important to verify the factor pattern coefficients are, themselves, invariant across time. This was checked by setting the factor pattern coefficient associated with item one and item two to be invariant across time when each of the other items served as the referent variable in the metric and scalar invariance models (Rensvold & Cheung, 2001).

At least partial, but preferably full, scalar invariance was required prior to examining latent mean change. Practical significance of latent mean change was assessed via a standardized latent mean change effect size for longitudinal comparisons:

$$\text{Latent } d = \frac{(\kappa_2 - \kappa_1)}{\sqrt{\Psi_{pooled}}}, \quad (3)$$

where κ_1 and κ_2 are the unstandardized latent means from time one and time two. Ψ_{pooled} is the pooled latent variance across time:

$$\sqrt{\Psi_{pooled}} = \sqrt{\frac{n_1\Psi_{1j} + n_2\Psi_{2j}}{n_1 + n_2}} * \sqrt{2(1 - \phi_{12})}, \quad (4)$$

where n_1 and n_2 are the sample sizes from time one and time 2 two, Ψ_1 and Ψ_2 are latent variances from time one and time two, and ϕ_{12} is the latent factor correlation between scores at time one and time two.⁸

Research question 6. Research question six focuses on the relationship between change in examinee effort from before to after the test and test performance. Likewise, the change in perceived importance may also be related to test performance. Given longitudinal scalar invariance was established for importance and effort scores from the prospective condition, the LMACS model was reparameterized as a two time point latent

⁸ The formulas used to compute the standardized longitudinal latent mean difference effect size are analogous to the formulas used to compute the standardized observed mean difference effect size longitudinal comparisons.

$$\text{Cohen's } d = \frac{M_2 - M_1}{SD_{pooled}} \text{ and } SD_{pooled} = \sqrt{\frac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{(n_1 + n_2 - 2)}} * \sqrt{2(1 - r_{12})}$$

growth model (LGM) in order to estimate individual differences in change scores and predict them from test performance.⁹

Estimation of the two time-point latent growth model was accomplished by specifying a latent intercept factor and a latent slope factor (see Figure 10). The intercept factor mean and variance simply represent the mean and variance for prospective scores. The slope factor mean and variance represent the mean of the change scores (equal to the mean change from the LMACS model) and the individual differences in change over time, the main parameter of interest in this model. Given significant slope variance, I examined if test performance explained these individual differences in change in effort and importance ratings. Moreover, the covariance between the intercept and slope factors indicated if prospective scores related to individual differences in change scores. A single-indicator latent variable approach was used to estimate the relationship between performance (NW9 composite scores) and change in importance and effort. By simultaneously estimating the LGM for importance and effort (see Figure 10), I assessed if change in importance is related to change in effort, after controlling for performance.

Research question 7. To assess if average levels of NW9 scores differed across conditions, a one-way between-subjects analysis of variance (ANOVA) was conducted followed by Tukey's HSD test for pairwise comparisons. Practical significance of differences in performance across conditions was assessed via standardized effect size indices (eta-squared and Cohen's *d*) and unstandardized effect size (mean difference

⁹ The two time point LGM parameterization is an equivalent model (e.g., equivalent fit, degrees of freedom, and parameter estimates) to the LMACS model. However, the conditional LGM can accommodate latent predictors of change. Thus, the relationships between change in importance and change in effort and test performance are able to be assessed. Further, LGM parameterization allows for estimation of the relationship between the change in perceived importance and the change in examinee effort (i.e., simultaneous growth).

between conditions). Eta-squared (η^2) was used to assess the practical significance of the omnibus test and was interpreted as the proportion of variance explained in the NW9 scores by condition. Cohen's d was used to assess the practical significance of the simple main effects and was interpreted as the difference between mean NW9 scores between conditions in standard deviation units.

Chapter IV

Results

Retrospective Importance and Effort Analyses

Retrospective perceived importance and examinee effort scores were modeled across conditions to assess the effects of behavioral commitment via priming.

Research question 1. On average, do examinees report different levels of retrospectively measured perceived test importance and examinee expended effort across measurement conditions? Recall, retrospective importance and effort scores from the combined and prospective conditions may be influenced by simply asking examinees to engage in rating their perceived importance and intended effort before test completion. Item-level statistics (means, variability, normality) and bivariate relationships are presented to foreshadow the fit of the competing models. Results of the measurement invariance tests followed by latent mean difference tests are then presented.

Descriptive statistics. Frequency distributions of importance and effort responses indicated all response options (i.e., 1-5) were utilized by examinees for each item in each of the three conditions. Item means ranged between 3.15 and 4.18 across items and were similar across conditions, indicating respondents typically responded neutral to agree with the items (see Table 5). Standard deviations ranged between 0.73 and 1.08 across items and were similar across conditions, indicating no floor or ceiling effects associated with each item. Item-level skewness and kurtosis values across conditions were less than $|1.01|$ and $|2.10|$. Assessing multivariate non-normality, Mardia's normalized multivariate kurtosis values of 18.11, 27.44, and 22.69 for retrospective, combined, and prospective conditions, respectively indicated some multivariate non-normality.

Estimation. The SMM used to test measurement invariance and latent mean differences in importance and effort scores across conditions requires the selection of an appropriate estimation technique. When employing a normal-theory estimator, such as ML, categorical or non-normal data can result in biased parameter estimates, standard errors, approximate fit indices, and χ^2 statistic. Examinees' responses to the importance and effort items are ordinal in nature (Likert scale); however, treating responses with five or more ordered categories as continuous tends to result in accurate parameter estimates (Finney & DiStefano, 2013). Thus, item responses were treated as continuous. Univariate skewness and kurtosis values less than $|2|$ and $|7|$, respectively, can be considered sufficiently univariately normal and have little effect on standard errors and fit indices (Finney, DiStefano, & Kopp, 2016). Mardia's normalized multivariate kurtosis values larger than 20 may be problematic (Peter Bentler, 1998, post on SEMNET). However, Mardia's multivariate kurtosis values have no clear cutoff indicating multivariate non-normality. Given the continuous nature of the item responses and small values of univariate skewness and kurtosis, all analyses employed ML estimation.

Inter-item correlations were examined to foreshadow the fit of the one- and two-factor models (see Table 5). In general, importance items were more strongly related to each other than to effort items and effort items were more strongly related to each other than to importance items across conditions. The pattern of inter-item correlations appears to align with a two-factor model and suggests gross misfit of the one-factor model. Moreover, the inter-item correlations appear to be relatively equivalent across conditions, suggesting metric invariance across conditions is plausible. Importance item three ("I am not curious about how I did on this test relative to others.") tended to relate more strongly

to importance items four (“I am not concerned about the score I receive on this test.”) and eight (“I would like to know how well I did on these tests.”) and less strongly to importance item five (“This was an important test to me.”) than to other importance items across conditions. This pattern of correlations associated with item three suggests item three may be problematic when fitting the two-factor model to scores across conditions.

Measurement invariance across conditions. As foreshadowed by the pattern of inter-item correlations, global model-data fit indices indicated gross misfit of the one-factor model within each condition (see Table 6). Moreover, several correlation residuals greater than $|.15|$ indicated the one-factor model did not represent the factor structure of the retrospective importance and effort scores.

Although model-data fit of the two-factor model was better than the one-factor model, global fit indices suggested areas of misfit (see Table 6). Moreover, examination of correlations residuals indicated non-ignorable local misfit. As foreshadowed by the inter-item correlations associated with importance item three, a large, positive correlation residual was associated with the relationship between importance items three and four across retrospective, combined, and prospective conditions (.20, .17, .16). The model underestimated the relationship between importance items three and four. Moreover, the correlation residual between importance items three and eight (.18) indicated the model underestimated the relationship in the prospective condition. The shared variance between importance items three and four after controlling for the importance factor (i.e., correlation residual) may be due to proximity of items or negative wording effects (Podsakoff, MacKenzie & Podsakoff, 2012). Shared variance between importance items three and eight, after controlling for the importance factor, may be due to item wording

redundancy (Podsakoff et al., 2012). This shared variance between importance items three and eight has been found in previous studies (Thek et al., 2009).

To account for the shared variance, the two-factor model was estimated with error covariances between importance item three and importance items four and eight. As expected, the two-factor model with two error covariances fit the importance and effort scores significantly better than the two-factor model without error covariances (see Table 6). The error covariances between items three and eight were statistically significant and moderate in size across conditions (residual correlations = .38, .23, .49) as were the residual correlations for items three and four (.30, .23, .26).

Even after this modification, the correlation residual between importance item eight and effort item ten revealed the model underestimated this relationship in the retrospective and combined conditions (.16 and .16). Further, a correlation residual of .16 was found between importance items four and eight in the prospective condition. Given 99% of the correlation residuals were less than $|\cdot 10|$ and global model-data fit indices were adequate within each condition, the two-factor model with error covariances between importance item three and importance items four and eight model (i.e., model 2-factor with I3, I4, & I8 model in Table 6) was championed.

Given adequate model-data fit of the two-factor model with the two error covariances, the factor correlation, factor pattern coefficients, variance explained by each factor in the set of items, and reliability estimates were examined. The factor correlation indicated the importance factor and effort factors are moderately related across retrospective, combined, and prospective conditions (.65, .69, and .65), respectively. The unstandardized factor pattern coefficients were relatively equivalent across conditions.

The standardized factor pattern coefficients associated with the importance factor tended to be small ($< .70$), with the exception of item one ($> .80$; see Table 7 for unstandardized and standardized factor pattern coefficients). The importance factor explained, on average, 37%, 41%, and 47% of the variance in the items for the retrospective, combined, and prospective conditions, respectively. Thus, measurement error accounts for greater than 50% of the variance in the importance items. The effort factor explained an average of 54%, 56%, and 54% of the variance in the items for the retrospective, combined, and prospective conditions, respectively. Estimates of reliability of the subscale scores indicated reliability was adequate across conditions (see Table 4).

A multiple-condition configural two-factor model with error covariances was estimated to obtain baseline fit indices. Given configural invariance was supported, metric invariance was tested. As foreshadowed by the similar unstandardized pattern coefficients across conditions, the metric invariant model with error covariances fixed to be equal across conditions fit the retrospective importance and effort scores well and did not fit statistically significantly worse than the configural model (see Table 6).

Given metric invariance was supported, measurement error invariance was examined. The measurement error invariant model fit well, but did fit the scores statistically significantly worse than the metric invariant model (see Table 6). However, the measurement error invariant model did not fit practically significantly worse than the metric invariant model, as indicated by the negligible ΔCFI . Moreover, the number and magnitude of correlation residuals did not change from the metric and scalar invariant model. Thus, measurement error invariance was supported.

Given metric invariance was supported, scalar invariance was examined. The scalar invariance model fit the retrospective importance and effort scores well, but did fit statistically significantly worse than the metric invariance model (see Table 6). Thus, local misfit was closely examined. All importance and effort items had nonsignificant standardized mean residuals and the largest unstandardized mean residual was $|.15|$. This mean residual was associated with importance item three. The observed means for importance item three for retrospective, combined, and prospective conditions were 3.37, 3.58, and 3.57, whereas, the model-implied means were 3.54, 3.50, and 3.54. Note the observed and model-implied means in the combined and prospective conditions are similar, whereas the model predicted the retrospective mean to be larger than observed. That is, because the factor pattern coefficients were equivalent across conditions and there were no latent mean differences across groups (as described below), the model over-estimated item three's mean in the retrospective condition to reflect the model-implied similarity in item-level means across conditions.

Although the mean residual (3/100 of the 5-point scale) appears to be negligible, a partial invariant model was tested by freely estimating importance item three's intercept. As expected, the partial invariance model fit statistically significantly better than the full scalar invariance model. Moreover, the largest unstandardized mean residual reduced to $|.08|$ from $|.15|$ in the full scalar invariance model. Given the results from the scalar and partial scalar models, latent mean differences were estimated with and without equality constraints on importance item three's intercepts to fully evaluate the impact of the scalar noninvariant item on the magnitude of the latent mean differences across conditions.

Latent mean differences. When estimating the full scalar invariance model, the retrospective condition perceived importance latent mean was not statistically significantly different than the combined ($\kappa = -0.09, t = 1.54, p > .01$) or prospective ($\kappa = 0.01, t = 0.16, p > .01$) conditions. The combined condition latent mean was not statistically significantly different than the prospective condition ($\kappa = 0.10, t = 1.53, p > .01$). Moreover, the standardized latent mean difference in importance scores between the retrospective condition and the combined ($d = 0.12$) and prospective ($d = 0.01$) conditions and between the combined and prospective conditions ($d = 0.12$) were not practically significant. That is, the retrospective condition's average latent importance score is 0.12 standard deviation units smaller than the average score from the combined condition.¹⁰

When estimating the partial scalar invariance model with no equality constraint on importance item three's intercept across conditions, the latent mean differences were essentially equivalent to those of the full scalar invariant model. The latent differences between the retrospective and combined ($\kappa = -0.09, t = 1.56, p > .01$) and prospective ($\kappa = 0.01, t = 0.17, p > .01$) conditions were not statistically significant. The combined condition latent mean was not statistically significantly different than the prospective condition ($\kappa = 0.10, t = 1.56, p > .01$). The standardized latent mean difference between the retrospective and combined ($d = 0.12$) and prospective ($d = 0.01$) conditions and between the combined and prospective conditions ($d = 0.13$) were essentially equivalent

¹⁰ Practically similar results are obtained when computing the standardized effect size on observed composite importance scores between the retrospective condition and the combined ($d = 0.04$) and prospective ($d = 0.10$) conditions and between the combined and prospective conditions ($d = 0.06$). The equivalency of latent and observed composite mean differences in importance scores is not surprising given the relatively high coefficient alpha values for the importance factor across conditions.

to the scalar invariance model with equality constraints on all item intercepts. Thus, the more parsimonious fully scalar invariant model was championed. Contrary to the hypothesis that average importance scores would be higher in the prospective condition due to the effect of behavioral commitment via priming, average importance scores were not statistically nor practically significantly different across conditions.

Similar results were found for average examinee effort across conditions. When estimating the full scalar invariant model, the retrospective condition latent mean was not statistically significantly different than the combined ($\kappa = -0.04$, $t = 0.78$, $p > .01$) or prospective ($\kappa = 0.08$, $t = 1.44$, $p > .01$) conditions.¹¹ The combined condition latent mean was not statistically significantly different than the prospective condition ($\kappa = 0.12$, $t = 2.25$, $p > .01$). The standardized latent mean difference in effort scores between the retrospective and combined ($d = 0.06$) and prospective ($d = 0.11$) conditions and between the combined and prospective conditions ($d = 0.18$) were not practically significant.¹²

Research question 2. Is the indirect effect of retrospectively measured perceived test importance on test performance via retrospectively measured examinee effort affected by asking examinees to rate their importance and effort before test completion? Descriptive statistics are presented to foreshadow the fit of the mediation model. Then,

¹¹ Changes to the partial scalar invariance model were limited to releasing the equality constraint on importance item three's intercept. No changes were made to the effort factor or its indicators. Thus, all latent mean differences and standardized latent difference estimates were equivalent across the full scalar and partial scalar invariance models.

¹² Nearly equivalent results are obtained when computing the standardized effect size on observed composite effort scores between the retrospective condition and the combined ($d = 0.06$) and prospective ($d = 0.12$) conditions and between the combined and prospective conditions ($d = 0.17$). The equivalency of latent and observed mean differences in effort scores is not surprising given the high coefficient alpha values for the effort factor across conditions.

the fit of the mediation model within each condition is presented. Finally, the test of the invariance of the indirect effect across conditions is presented.

Descriptive statistics. Given support for measurement invariance across conditions, composite effort and importance scores were computed and mediation path models were estimated using single-indicator latent variables. Composite importance, effort, and NW9 were examined to determine how examinees responded on average and the variability in scores around these means (see Table 4). Aligning with the SMM results above, average importance and effort composite scores were essentially equivalent across conditions and indicated examinees tended to report, on average, moderate importance and effort. Given the importance and effort composite scores are on the same metric (i.e., 5-25), note the relatively higher variability in importance scores. Further, this pattern of variability is essentially equivalent across conditions. NW9 composite scores and their variability were essentially equivalent across conditions. Consistent with previous findings, examinees, on average, correctly responded to approximately 68% of the NW9 items. Univariate skewness and kurtosis values indicated the importance, effort, and NW9 composite scores were essentially univariately normally distributed (Finney, DiStefano, & Kopp, 2016). Normalized Mardia's multivariate kurtosis coefficients of 3.37, 3.17, and 2.33 for retrospective, combined, and prospective conditions, respectively, suggested the importance, effort, and NW9 composite scores could be treated as essentially multivariately normally distributed and thus ML estimation was employed when estimating the mediation path model.

Examination of intercorrelations among importance, effort, and NW9 composite scores foreshadows the fit of the partial and completely mediated models (see Table 4).

As expected, given the hypothesized indirect effect, the relationships between importance and effort scores and between effort and NW9 scores were stronger than the relationship between importance and NW9 scores across conditions. Thus, it is plausible that the completely mediated model will fit the composite scores within each condition. Moreover, the correlations between importance, effort, and NW9 scores are essentially equivalent across conditions. This finding, coupled with the similar variances for these variables, suggests the indirect effect may be invariant across conditions.

Single-condition analyses. As foreshadowed by the intercorrelations among observed variables, complete mediation was supported within each of the three conditions (see Table 8). Thus, the relationship between importance and performance is solely accounted for by the indirect effect between importance and performance via effort.¹³

The direct paths from importance to effort and effort to performance were statistically and practically significant ($p < .01$) across conditions (see top of Figure 9, retrospective SOS model). In the retrospective condition, the standardized direct effect between perceived importance and effort and effort and test performance is interpreted as for every standard deviation unit change in importance, effort increases by .62 standard deviation units and for every standard deviation unit change in effort, test performance increases by .46 standard deviation units. The variance explained in effort by importance (1 - standardized disturbance term) was 39%, 45%, and 34% across retrospective, combined, and prospective conditions, respectively. The variance explained in test performance by effort was 22%, 17%, and 15% across retrospective, combined, and

¹³ The nonsignificant χ^2 value associated with the completely mediated model indicates the partial mediation model, which is just-identified, cannot fit significantly better than the completely mediated model.

prospective conditions, respectively. The essentially equivalent unstandardized direct paths and unexplained variances (disturbance terms) across conditions suggest these parameter estimates may be invariant across conditions.

The indirect effects of perceived importance on test performance via examinee effort were statistically significant for the retrospective (unstandardized = 0.637, 99% CI [0.400, 0.914], standardized = .287), combined (unstandardized = 0.666, 99% CI [0.422, 0.943], standardized = .280), and prospective (unstandardized = 0.450, 99% CI [0.230, 0.711], standardized = .223) conditions. The standardized indirect effect can be interpreted as for every standard deviation unit change in perceived importance, test performance increases by .29 standard deviations in the retrospective condition. As indicated by the standardized indirect effects, the practical significance was moderate to large in magnitude across conditions (Kenny, 2016; Kenny & Judd, 2014).

Multiple-condition analyses. Invariance of the parameters estimated in the completely mediated model (i.e., direct effects, variances, and unexplained variances) were examined across conditions. First, a multiple-condition, unconstrained, completely mediated model was estimated using retrospective importance and effort composite scores and NW9 scores from the three conditions to obtain baseline global fit indices (see Table 8, unconstrained model). Next a multiple-condition, completely mediated model was estimated constraining the unstandardized direct paths from perceived importance to examinee effort and examinee effort to test performance to be equivalent across conditions (see Table 8, constrained direct paths). The constrained unstandardized direct paths model fit well and did not fit significantly worse than the unconstrained multiple-condition model. Finally, in addition to the constrained unstandardized direct paths, the

variance of perceived importance and the two unstandardized disturbance variances were constrained to be equal across conditions (see Table 8, completely constrained). The completely constrained model fit well and did not fit significantly worse than the constrained direct paths multiple-condition model. Thus, condition did not moderate the interrelationships between perceived importance, examinee effort, and test performance or their variances. In short, the simple act of indicating one's perceived test importance and intended effort prior to the test did not influence the interrelationships between retrospectively gathered importance and effort scores, and NW9 scores.

When constrained across conditions, the direct paths from perceived importance to examinee effort ($b = 0.697, p < .01, \beta = .627$) and from examinee effort to test performance ($b = 0.846, p < .01, \beta = .424$) were statistically and practically significant. When constrained across conditions, the indirect effect of perceived importance on test performance via examinee effort was statistically and practically significant (unstandardized indirect effect = 0.590, 99% CI [0.450, 0.742], standardized indirect effect = .266). When constrained across conditions, the variance explained in effort by importance was 40% and the variance explained in performance by effort was 17%.

Summary of Retrospective Importance and Effort Analyses. No mean differences in perceived importance and examinee effort scores were found across conditions. Further, the interrelationships between perceived importance, examinee effort, and test performance, their variances, and unexplained variances were not moderated by test condition. Asking examinees to report their perceived importance and intended effort before test completion does not appear to have an effect on average importance, effort, or test scores nor their interrelationships. Thus, there is no evidence

that behavioral commitment via priming positively impacts importance and effort scores or the interrelationships between importance, effort, and test scores.

The results thus far have not helped elucidate if examinees attribute the cause of their test performance to reported importance and effort. Answering the remaining research questions may help determine if examinees engage in attributional bias by reporting low expended effort to justify their performance on the test.

Prospective and Retrospective Importance and Effort Analyses

Recall, research questions three and four utilize a combination of retrospective and prospective importance and effort scores across conditions. Retrospective scores within the retrospective condition will be free from any influence of priming; thus, there is no possibility of behavioral commitment. However, these scores may reflect post-test attributions of test performance rather than true perceptions of importance and effort. Prospective importance and effort scores in the prospective condition should reflect true perceptions of importance and intended effort and will be free from any contamination due to attributional bias. Moreover, prospective ratings may induce behavioral commitment and influence test performance. Answering these questions may begin to uncover the effects of behavioral commitment and attributional bias (see Table 2).

Research question 3. Do examinees reporting retrospective perceived importance and expended effort have lower scores, on average, than examinees reporting prospective importance and intended effort for the same test? Whereas results from research question one indicated presence or absence of behavioral priming had no effect on retrospective importance and effort scores, this comparison will indicate if prospective importance and effort scores differ, on average, from retrospective scores.

Descriptive statistics. Frequency distributions of importance and effort scores indicated all response options (i.e., 1 through 5) were utilized by examinees for each item across conditions. Item means ranged between 3.15 and 4.36 across conditions, indicating respondents tended to respond neutral to agree to the items (see Table 9). Standard deviations ranged between 0.72 and 1.07 across conditions. Item-level skewness and kurtosis values were less than $|1.60|$ and $|4.34|$. Mardia's normalized multivariate kurtosis values of 18.11, 25.66, and 32.77 for retrospective, combined, and prospective conditions indicated some multivariate non-normality. Given the small univariate skewness and kurtosis values, ML estimation was used for the following analyses.

Inter-item correlations were examined to foreshadow the fit of the one- and two-factor models. Importance items were generally more strongly related to each other than to effort items and effort items were more strongly related to each other than to importance items across conditions. The pattern of inter-item correlations appears to align with adequate fit of the two-factor and gross misfit of the one-factor model. Inter-item correlations appear to be relatively equivalent across conditions, however, effort item seven in the prospective condition ("After taking this test, I expect I could have worked harder on it.") has relatively small correlations with the importance items. This suggests possible metric noninvariance across conditions. As reported above, importance item three tended to relate more strongly to importance items four and eight and less strongly to importance item five than to other importance items across conditions. This pattern of correlations associated with item three suggests item three may be problematic when fitting the two-factor model to scores across conditions.

Measurement invariance across conditions. As found in the retrospective CFA analyses and foreshadowed by the pattern of inter-item correlations, model-data fit indices indicated gross misfit of the one-factor model fit to retrospective importance and effort scores within the retrospective condition, prospective importance and retrospective effort scores within the combined condition, and prospective importance and effort scores within the prospective condition (see Table 10). As foreshadowed by relatively larger inter-item correlations associated with importance item three, large, positive correlation residuals indicated the model underestimated the relationship importance items three and four across conditions (.20, .17, .16) and underestimated the relationship between importance items three and eight in the prospective condition (.15).

To account for the shared variance, the two-factor model was estimated with error covariances between importance item three and importance items four and eight. As expected, the two-factor model with two error covariances fit the importance and effort scores significantly better than the two-factor model without error covariances (see Table 10). The residual correlations between items three and eight were statistically significant and moderate in size across conditions (.38, .22, .47) as were the residual correlations for items three and four (.30, .21, .24). Within the retrospective condition, a correlation residual between importance item eight and effort item ten (.16) indicated the model underestimated this relationship and within the prospective condition, a correlation residual between effort item seven and importance item one (-.15) indicated the model overestimated this relationship. Given 99% of the correlation residuals were less than |.10| and global model-data fit indices were adequate within each condition, the two-

factor model with error covariances between importance item three and importance items four and eight model was determined to fit the data well.

Given adequate model-data fit of the two-factor model with two error covariances, the factor correlation, factor pattern coefficients, variance explained, and reliability estimates were examined. The correlation between importance and effort factors measured after test completion within the retrospective condition ($r = .69$) was similar in magnitude to the correlation between the importance and effort factors measured prior to test completion within the prospective condition ($r = .65$). The correlation between the importance factor measured before test completion and the effort factor measured after test completion in the combined condition ($r = .51$) was relatively smaller in magnitude. The standardized factor pattern coefficients associated with the importance factor were generally small ($< .70$), with the exception of item one ($> .80$; see Table 11 for unstandardized and standardized factor pattern coefficients). The importance factor explained, on average, 37%, 38%, and 40% of the variance in the items for the retrospective, combined, and prospective conditions, respectively. Thus, measurement error accounts for greater than 50% of the variance in the importance items. The effort factor explained an average of 54%, 56%, and 50% of the variance in the items for the retrospective, combined, and prospective conditions, respectively. Estimates of reliability of both subscale scores indicated reliability was adequate across conditions (see Table 4). With the exception of the factor pattern coefficient associated with effort item seven in the prospective condition, the unstandardized factor pattern coefficients were relatively equivalent across conditions, which foreshadows measurement invariance discussed now.

A multiple-condition configural two-factor model with error covariances was estimated to obtain baseline global-fit indices. Given configural invariance was supported, metric invariance was tested. As foreshadowed by the smaller unstandardized factor pattern coefficient associated with effort item seven in the prospective condition, the metric invariant model with error covariances fit statistically and practically significantly worse than the configural model (see Table 10). Large correlation residuals (-.24 and -.17) associated with effort item seven and importance item one and four, respectively, indicated the model overestimated the relationships. This is not surprising, given the near zero correlations associated with item seven and the importance items.

Given metric noninvariance, a partial metric invariance model was examined by freely estimating the unstandardized factor pattern coefficient associated with effort item seven in the prospective condition. The partial metric invariance model did not fit the data statistically significantly worse than the configural model. Two correlations residuals remained, one between effort item seven and importance item one (-.16) and one between importance item eight and effort item ten (.19).

Given 99% of the correlation residuals were less than $|.10|$ and global model-data fit indices were adequate, a partial scalar invariance model with no constraints on effort item seven's unstandardized factor pattern coefficient or intercept in the prospective condition was estimated. The partial scalar invariant model fit the data well, but fit statistically significantly worse than the partial metric invariant model (see Table 9). All importance and effort items had nonsignificant standardized mean residuals and the largest unstandardized mean residual was $|.06|$. Given the mean residual was 1/100 of the 5-point response scale, the mean residual appears to be negligible.

Latent mean differences. Given partial scalar invariance, latent mean differences were estimated. Average retrospective importance scores from the retrospective condition were not statistically significantly different than average prospective importance scores from the combined ($\kappa = 0.03$, $t = 0.49$, $p > .01$) or prospective conditions ($\kappa = 0.07$, $t = 1.27$, $p > .01$). The combined condition latent mean was not statistically significantly different than the prospective condition latent mean ($\kappa = 0.10$, $t = 1.70$, $p > .01$). Moreover, the standardized latent mean difference in importance scores between the retrospective and combined ($d = 0.04$) and prospective ($d = 0.11$) conditions and between the combined and prospective conditions ($d = 0.15$) were not practically significant.¹⁴

With respect to examinee effort scores, prospective intended effort scores from the prospective condition were statistically significantly higher than retrospective expended effort scores from the retrospective ($\kappa = 0.28$, $t = 5.60$, $p < .01$) and combined ($\kappa = 0.31$, $t = 6.52$, $p < .01$) conditions. Moreover, the standardized latent mean differences indicated the prospective condition latent prospective effort scores were approximately one-half of a standard deviation larger than the retrospective effort scores from the retrospective ($d = 0.45$) and combined ($d = 0.51$) conditions. The retrospective

¹⁴ Practically similar results are obtained when computing the standardized effect size on observed composite importance scores between the retrospective condition and the combined ($d = 0.02$) and prospective ($d = 0.08$) conditions and between the combined and prospective conditions ($d = 0.06$). The equivalency of latent and observed composite mean differences in importance scores is not surprising given the relatively high coefficient alpha values for the importance factor across conditions.

condition latent mean was not statistically or practically significantly different than the combined condition ($\kappa = 0.04$, $t = 0.78$, $p > .01$; $d = 0.06$).¹⁵

The results partially support the hypothesis that average prospective importance and effort scores would be higher in the prospective condition. Recall, results of research question one indicated no differences in average retrospective importance and effort scores across conditions after asking examinees to engage in reporting their perceived importance and intended effort. In contrast, results associated with question three indicate prospective effort scores are higher than retrospective effort scores across conditions. Thus, retrospective effort scores may be affected by attributional bias, but, importantly these results indicate completion of the test negatively influences ratings of effort. Next, I will examine if the interrelationships between importance, effort, and performance also depend on whether examinees experience of the test or not.

Research question 4. Does the theoretically and empirically suggested indirect effect of perceived test importance on test performance via examinee effort differ depending on if examinees report importance and effort prospectively versus retrospectively? Whereas research question two indicated the indirect effect estimated from retrospective importance and effort scores was not influenced by merely asking examinees to engage in prospective ratings of importance and effort, this comparison will indicate if the indirect effect depends on when importance and effort are measured relative to test completion. Descriptive statistics are presented to foreshadow the fit of the

¹⁵ Nearly equivalent results are obtained when computing the standardized effect size between observed prospective effort scores in the prospective condition and retrospective effort scores in the retrospective ($d = 0.39$) and combined ($d = 0.45$) conditions and between retrospective effort scores in the retrospective and combined ($d = 0.06$) conditions. The equivalency of latent and observed mean differences in effort scores is not surprising given the high coefficient alpha values for the effort factor across conditions.

mediation model. Then, the fit of the mediation model within each condition is presented. Finally, the invariance test of the indirect effect across conditions is presented.

Descriptive statistics. Given support for measurement invariance across conditions, effort and importance scores were computed and mediation path models were estimated using single-indicator latent variables (see Table 4). As indicated by the SMM results, average importance scores were essentially equivalent across conditions. However, prospective effort scores were higher than retrospective scores. Importance and effort scores indicated examinees tended to report, on average, moderate importance and effort. Importance scores had similar variability across conditions ($SDs = 3.23, 3.33, 3.30$). Interestingly, prospective effort scores in the prospective condition had less variability ($SD = 2.87$) than retrospective effort scores in the retrospective and combined conditions ($SDs = 3.61, 3.58$).¹⁶ Total NW9 scores and their variability were essentially equivalent across conditions. Univariate skewness and kurtosis values indicated the importance, effort, and NW9 composite scores were essentially univariately normally distributed. Normalized Mardia's multivariate kurtosis coefficients of 3.37, 4.49, and 2.67 for retrospective, combined, and prospective conditions, respectively, suggested data could be treated as essentially multivariately normally distributed and thus ML estimation was employed when estimating the mediation path model.

¹⁶ A formal test of the equivalence of the variance in effort scores was performed by estimating a multiple-condition 2-factor CFA metric invariance model with error covariances (RQ 4). Constraining the effort factor variance to be equal across conditions resulted in statistically significantly worse fit than constraining the effort factor variance to be equal in the retrospective and combined conditions (retrospective effort scores) and freely estimating it in the prospective condition (prospective effort scores), $\Delta\chi^2(1) = 12.54$. Thus, the variance in prospective effort scores was statistically significantly less than the variance in retrospective effort scores.

Examination of importance, effort, and NW9 intercorrelations foreshadows the fit of the completely mediated models within each condition (see Table 4). As expected, the relationships between importance and effort scores and effort and NW9 were stronger than the relationship between importance and NW9 scores within each condition. Thus, it is plausible that the completely mediated model will fit the scores within each condition.

Single-condition analyses. As foreshadowed by the intercorrelations, complete mediation was supported across the three conditions (see Table 12). The direct paths from importance to effort and effort to performance were statistically ($p < .01$) and practically significant across conditions (see Figure 9). The variance explained in examinee effort by importance was 39%, 28%, and 27% across retrospective, combined, and prospective conditions, respectively. The variance explained in test performance by effort was 22%, 18%, and 7% across retrospective, combined, and prospective conditions, respectively.

Recall the indirect effect of retrospective importance on performance via retrospective effort was statistically significant for the retrospective condition (unstandardized = 0.637, 99% CI [0.400, 0.914], standardized = .287). The indirect effect of prospective importance on performance via retrospective effort was significant for the combined condition (unstandardized = 0.534, 99% CI [0.325, 0.783], standardized = .224). The indirect effect of prospective importance on performance via prospective effort was significant for the prospective condition (unstandardized = 0.338, 99% CI [0.096, 0.617], standardized = .156). As indicated by the standardized indirect effects, the practical significance ranged from large to medium in magnitude across conditions. These differences in the standardized indirect effect reflect the significantly smaller variance in prospective effort scores in the prospective condition than in the retrospective

and combined conditions. Moreover, the smaller variance in prospective effort scores leads to a smaller correlation between effort and NW9 scores in the prospective condition than in the retrospective and combined conditions (see Table 4). This difference in unexplained effort variance in the prospective condition suggests possible noninvariance.

Multiple-condition analyses. Invariance of the parameters estimated in the completely mediated model were examined across conditions. First, a multiple-condition, unconstrained, completely mediated model was estimated to obtain baseline fit indices using: retrospective importance and effort scores from the retrospective condition, prospective importance and retrospective effort scores from the combined condition, and prospective importance and effort scores from the prospective condition (see Table 12, unconstrained model). A multiple-condition, completely mediated model was then estimated constraining the unstandardized direct paths from importance to effort and effort to performance to be equivalent across conditions (see Table 12, constrained direct paths). The constrained unstandardized direct paths model fit well and did not fit significantly worse than the unconstrained model. Next, in addition to the constrained unstandardized direct paths, the variance of importance and the two unstandardized disturbance variances were constrained to be equal across conditions (see Table 12, completely constrained). The completely constrained model fit the data adequately, but did fit significantly worse than the constrained direct paths multiple-condition model.

Recall the prospective effort variance was significantly smaller than the retrospective variance. Thus, given importance has an equal unstandardized effect on effort across conditions, the unexplained variance (disturbance) has to differ across conditions. That is, there is less effort variance to be explained, therefore less

unexplained variance. Given effort's relatively smaller unstandardized disturbance variance in the prospective condition (see bottom of Figure 9), a multiple-condition model was estimated allowing effort's disturbance variance to be freely estimated for the prospective condition. The partially constrained model allowing effort's disturbance to be freely estimated in the prospective condition fit the data well and did not fit significantly worse than the constrained direct paths multiple-condition model. Thus, condition did not moderate the unstandardized relationships between importance, effort, and performance, but did moderate the variance explained in effort in the prospective condition.

When constrained across conditions, the direct paths from importance to effort ($b = 0.624, p < .01, \beta = .541$) and from effort to performance ($b = 0.820, p < .01, \beta = .414$) were statistically and practically significant. When constrained across conditions, the indirect effect of importance on performance via effort was statistically and practically significant (unstandardized indirect effect = 0.512, 99% CI [0.380, 0.657], standardized indirect effect = .224). Although the unstandardized indirect effect is invariant across conditions, the relatively smaller effort variance in the prospective condition results in an overestimation of the standardized indirect effect in the prospective condition when estimating the partially constrained model. When unstandardized direct paths and variances (effort's disturbance freely estimated in prospective condition) were constrained across conditions, the variance explained in effort was 29% in the retrospective and combined conditions and 44% in the prospective condition. The variance explained in performance by effort was 17% in the retrospective and combined conditions and 12% in the prospective conditions.

Summary of Prospective and Retrospective Importance and Effort Analyses.

Examinees did not report different levels of prospective and retrospective perceived importance across conditions, yet examinees tended to report lower retrospective effort than prospective effort. The unstandardized interrelationships between perceived importance, examinee effort, and test performance were not moderated by the timing of measurement of importance and effort. However, given the variance in effort not explained by importance was smaller in the prospective condition, the standardized interrelationships between importance, effort, and performance differed depending on when effort was measured relative to test completion. Thus, when examinee effort is measured relative to test completion appears to influence average levels of effort and its standardized interrelationships with importance and performance.

Changes in Test Importance and Effort Scores

Change in perceived importance and effort from before to after test completion was modeled using prospective and retrospective importance and effort scores collected from examinees in the prospective condition. Latent mean change in importance and effort scores from before to after test completion was estimated after establishing longitudinal measurement invariance (research question 5). Given individual differences in change in importance and change in effort, I examined if test performance predicted change in examinees' reported importance and effort (research question 6).

Research question 5. Do examinees completing both prospective and retrospective measures report different levels of prospective perceived importance and examinee effort, on average, than retrospective importance and effort? Examinees may make post-test attributions (e.g., low expended effort, disinterest in test) to explain their

test performance. Consequently, lower retrospective importance and effort scores may result in substantially more examinees being removed when employing motivation filtering procedures than when using prospectively measured importance and effort scores. Likewise, a simple description of the level of examinee motivation may be dramatically different when the description is based on average prospective versus retrospective importance and effort scores.

Recall the two-factor model with error covariances between importance items three and four and three and eight was fit individually to prospective importance and effort scores (research question 3) and retrospective importance and effort scores (research question 1) within the prospective condition (see Tables 6 and 10). To assess change in ratings from before to after the test, this two-factor model was now fit simultaneously to prospective and retrospective importance and effort scores within the prospective condition utilizing LMACS analysis (see Table 13). Corresponding item-level error covariances were modeled across the two time points, and importance items three, four, and eight were allowed to covary within as well as across time. The configural model fit the data well globally (see Table 14). Examination of correlations residuals revealed the model overestimated the correlation ($|.15|$) between prospective importance item one (“Doing well on this test is important to me.”) and prospective effort item seven (“After taking this test, I expect could have worked harder on it.”). Given adequate model-data fit of the configural model, the unstandardized factor pattern coefficients were examined (see Table 15). With the exception of importance item eight, the unstandardized factor pattern coefficients were roughly equivalent across time points.

As foreshadowed by examination of the unstandardized factor pattern coefficients, the metric invariance model did not fit the scores well globally or locally (see Table 14). Examination of correlation residuals revealed five residuals greater than $|.15|$ and one correlation residual of $|.20|$, three of which were associated with importance item eight. Given the large correlation residuals and discrepancy in unstandardized factor pattern coefficients across time associated with importance item eight, a partial metric invariance model was estimated freeing the equality constraint imposed on importance item eight's unstandardized factor pattern coefficient. The partial metric model fit the data significantly better globally and locally. Examination of correlation residuals revealed two residuals greater than $|.15|$. The pattern of correlation residuals did not indicate an apparent reason for metric noninvariance; thus, each item was tested for noninvariance by individually releasing each item's factor pattern coefficient equality constraint. None of the remaining items were associated with significant misfit when the item's factor pattern coefficient equality constraint was released across time, in addition, no large residuals appeared when applying the constraint.

Given partial metric invariance was supported, partial scalar invariance was examined. The partial scalar invariant model with no factor loading or intercept equality constraints on importance item eight fit the data well globally and locally. Given the partial scalar invariance model fit statistically significantly worse than the partial metric invariance model, local misfit was assessed. All importance and effort items had nonsignificant standardized mean residuals and the largest unstandardized mean residual was $|.10|$. Given the metric of the items (1 to 5), the size of the mean residual is

negligible. Thus, partial scalar invariance was supported and latent mean change in effort and importance was estimated.

Latent mean change. The average prospectively measured perceived importance latent score was not statistically significantly different than the average retrospectively measured importance latent score ($\kappa = -0.03$, $t = 1.07$, $p > .01$). Moreover, the standardized latent mean change in importance ($d = 0.13$) was not practically significant. On average, latent importance scores decreased 0.13 standard deviation units from before to after test completion.¹⁷ Average prospectively measured examinee effort was statistically and practically significantly higher than the average retrospectively measured examinee effort ($\kappa = 0.21$, $t = 6.00$, $p < .01$). On average, latent examinee effort scores decreased 0.41 standard deviation units from before to after test completion.¹⁸

Thus, contrary to my prediction, importance scores did not change from before to after test completion. However, aligning with my prediction, retrospectively measured effort scores were statistically and practically significantly lower than prospectively measured effort scores, suggesting that examinees may attribute the cause of their test performance to lack of expended effort.

Latent score stability. Stability of importance and effort scores over time was assessed (test-retest reliability). The latent correlation between prospective and

¹⁷ Practically similar results were obtained when computing the standardized effect size between observed composite prospective and retrospective importance scores ($d = 0.03$). The equivalency of latent and observed composite mean differences in importance scores over time is not surprising given the relatively high coefficient alpha values for the importance factor over time.

¹⁸ Practically similar results were obtained when computing the standardized effect size between observed composite prospective and retrospective effort scores ($d = 0.32$). The equivalency of latent and observed composite mean differences in effort scores over time is not surprising given the relatively high coefficient alpha values for the effort factor over time.

retrospective perceived importance scores was .90, which indicates examinees tend to stay in the same relative rank-order from before to after test completion. The latent correlation between prospective and retrospective examinee effort scores was .67, which indicates examinees tend to change relative rank-order from before to after test completion. Given the variability in the way examinees change their ratings over time, I will next examine this variability in change in effort and if this change can be predicted by test performance.

Research question 6. Is change in importance or effort related to performance on the test? Attributional bias may serve as a self-protective mechanism to a greater degree for examinees who perform poorly on the test than for examinees who perform well on the test. Thus, examinees who perform poorly may tend to decrease their reported importance and effort from before to after test completion to a greater extent than examinees who perform well on the test.

The LMACS model was reparameterized as a simultaneous growth model to obtain the variance in importance and effort latent change scores. Given the variability in latent importance and effort change scores was significantly different from zero (0.11 and 0.26), a conditional simultaneous growth model with a time-invariant latent test performance (NW9) predictor was estimated to potentially explain this variability (see Figure 10). Global fit indices indicated the model fit the data adequately; thus, parameter estimates were examined.

Test performance did not statistically significantly predict change in importance (see Figure 11; $b = 0.02$, $p > .01$; $R^2 < .01$), or change in effort (see Figure 12; $b = 0.15$, $p > .01$; $R^2 = .03$). Thus, examinees who performed poorer on the test did not change their

importance or effort ratings differently than examinees who performed better on the test. Test performance statistically significantly predicted baseline (prospective) effort scores ($b = 0.21, p < .01; R^2 = .05$). This relationship can be interpreted as, for every 10 unit (correct item) increase in test performance, examinees increased their effort ratings (1-5 scale) by an average of 0.21 units.¹⁹ Test performance did not statistically significantly predict prospective importance scores ($b = 0.14, p > .01; R^2 = .02$).

Contrary to my hypothesis, change in importance and effort scores is not related to test performance. However, change in importance had a moderate to strong relationship with change in effort, even after controlling for test performance (partial $r = .66$). There was a nonsignificant decrease in average importance ratings over time ($M_{\text{slope}} = -0.03$) and a significant decrease in effort ratings ($M_{\text{slope}} = -0.21$). Thus, a positive relationship between change importance and effort can be interpreted as examinees' importance ratings decreased over time, their effort ratings decreased over time.

Test Performance Differences: Research question 7.

Does test performance differ, on average, as a function of measurement condition? A one-way between-subjects ANOVA was conducted to evaluate the relationship between test scores and condition (retrospective, combined, and prospective). The test scores were statistically significantly different across conditions, $F(2, 1142) = 6.40, p = .002$. However, only 1% of the variance in test scores ($\eta^2 = .01$) could be accounted for by condition. Tukey HSD post hoc analyses were conducted to evaluate mean differences across pairs of conditions. The average total test score in the

¹⁹ Estimation problems arose due to differences in the metric of NW9 composite scores (0-66) and importance and effort item scores (1-5). Therefore, the model was estimated after the total NW9 composite scores were scaled by a factor of 10 (0-6.6).

prospective condition was statistically significantly higher than the average score in the retrospective condition, $t(706) = 3.46, p = .002$ (see Table 4 for means and standard deviations). Average prospective condition test scores were 0.26 standard deviations higher than retrospective condition test scores ($d = 0.26$, using the pooled standard deviation). The arguably negligible unstandardized effect size of an approximately two point difference in observed average test scores between these two conditions further indicates the small effect. Average test scores in the combined condition were not statistically or practically significantly different than average test scores in the retrospective, $t(835) = 2.42, p = .042, d = 0.17$, or prospective conditions, $t(743) = 1.29, p = .401, d = 0.10$. Thus, the hypothesis that test performance in the prospective condition would be higher than scores in the retrospective and combined conditions was supported, yet, the effect of condition on test performance was small. Given prior math ability—as measured by SAT math scores—was essentially equivalent across conditions, behavioral priming may have a small effect on examinees' test performance.

Chapter V

Discussion

The purpose of this study was to empirically examine the differential effects of measuring perceived importance and examinee effort before and after test completion. The results in light of previous research and implications are discussed below. Limitations of the current study and suggestions for future research are then discussed.

No Impact of Behavioral Commitment on Retrospective Importance and Effort

Measurement invariance and no average differences across conditions. On average, do examinees report different levels of retrospectively measured perceived test importance and expended effort across conditions that primed or did not prime behavioral commitment? Prior to evaluating average differences of importance and effort scores across conditions, measurement invariance (i.e., configural, metric, scalar, and measurement error invariance) was established for the scores across conditions. The SOS appears to be measuring the same construct in the same way across three independent conditions. Importantly, this study is the first to evaluate the measurement error invariance of the SOS. Measurement error invariance is of practical importance because it allows for one to make comparisons of observed score differences across groups (Millsap & Meredith, 2007). The measurement error invariance of the retrospective SOS coupled with the practically equivalent latent and observed mean differences of retrospective importance and effort scores found in this study provide further justification for researchers and testing practitioners to utilize and compare observed SOS scores. These measurement invariance results add to previous research on the test-specific retrospective SOS (e.g., Finney, Mathers, & Myers, 2016).

Contrary to my hypothesis, there were no statistically or practically significant differences in average levels of retrospective importance or effort scores across the three measurement conditions. Recall, examinees in the retrospective condition were not asked to engage in reporting their perceived importance or intended effort before test completion. Thus, these retrospective importance and effort scores are theoretically free of the effects of behavioral commitment. Comparing scores from the retrospective condition to the other conditions allows for an explicit test of the effect of the presence or absence of behavioral priming on average retrospective importance and effort scores and their variability. Given average levels and variability in importance and effort scores were essentially equivalent across conditions, behavioral commitment did not appear to have any influence on or mitigate a possible effect of attributional bias on average ratings of retrospective importance or effort. These results imply examinees may attribute the cause of their performance equivalently in the presence and absence of behavioral priming. Thus, contrary to previous research, reporting behavioral intentions before the test did not appear to influence later behavior or attitudes as operationalized by self-reported expended effort and perceived test importance (e.g., Carver & Scheier, 2000).

The implications of these results for testing practice are clear. Unfortunately, developing potential motivation interventions aimed at increasing effort put forth on a test via priming mechanisms may not have the desired result of increasing the level of retrospective ratings of expended effort and thus increasing the level of test performance. Although examinees asked to report their perceived importance and intended effort before the test did not differ in their reported levels of expended effort from examinees who only reported effort after the test, it may be prudent for researchers to

explore other behavioral priming interventions that may be more salient to examinees, which may in turn have a stronger impact on examinee behavior. For example, rather than having examinees state their intentions by responding to vague statements (e.g., “I will give my best effort on this test”), having examinees more explicitly state the how, when, or why they intend to reach their goal (e.g., “Because this test is important, I am going to try hard on every question;” Gollwitzer & Sheeran, 2006).

No difference in indirect effect across conditions. Is the indirect effect of retrospectively measured perceived test importance on test performance via retrospectively measured examinee effort affected by asking examinees to rate their perceived importance and intended effort before test completion? As hypothesized, the completely mediated model fit the scores within each condition. However, contrary to my hypothesis, the magnitude of the unstandardized and standardized indirect effects was essentially equivalent across conditions and was similar in magnitude to the indirect effect found in previous studies (e.g., Cole et al., 2008; Penk & Schipolowski, 2015).

This finding extends previous research uncovering the stability of the indirect effect of importance on performance across college students taking math and science tests (e.g., Cole et al., 2008; Penk & Schipolowski, 2015; Zilberberg et al., 2014) and across test instruction conditions (e.g., Myers et al., 2016). Given the extent to which examinees perceive low-stakes tests as important directly influences effort put forth on tests, which, in turn, directly influences test performance remains unchanged in the presence or absence of behavioral priming, EV theory remains a useful framework to explain the variability in motivation and variability in test scores in low-stakes testing contexts (Wise & DeMars, 2005; Wolf & Smith, 1995). Considering the *Standards*’ recommendation to

interpret test scores in light of information that may result in inaccurate interpretations of test scores (AERA, APA, & NCME, 2014), it is imperative that researchers consider importance and effort when examining factors that may influence test performance on low-stakes tests. Another implication of this finding is that behavioral priming may not be an effective motivation intervention for increasing test performance via increasing retrospectively reported test importance and expended effort. That is, asking examinees to report their perceived importance and intended effort before the test does not increase average retrospective importance and effort, nor does it increase the strength of the relationships between retrospective importance, effort, and test performance.

Differences in Prospectively and Retrospectively Measured Effort

Measurement invariance and differences in effort across conditions. Do examinees reporting retrospective importance and effort have lower scores, on average, than examinees reporting prospective importance and effort for the same test? Prior to evaluating average differences of importance and effort ratings across conditions, measurement invariance was established for prospective and retrospective scores across conditions. The test-specific prospective SOS functioned well in terms of structure, reliability, and scaling relative to the test-specific retrospective SOS. These results have an important implication for practice. Given its psychometric properties, the prospective SOS provides researchers and testing practitioners a viable option for measuring importance and effort prior to test completion. Moreover, invariance in the scaling of the prospective and retrospective measures enables comparisons of prospective and retrospective importance and effort scores. One caveat is that future research should

evaluate the functioning of effort item seven from the prospective SOS. If noninvariance of effort item seven scores replicates, revision of the item may be necessary.

As hypothesized, average levels of prospective effort scores from the prospective condition were statistically and practically significantly higher than retrospective effort scores from the retrospective and combined conditions. However, there were no significant differences in average levels of importance scores across conditions. The finding of lower average retrospective effort ratings is not surprising given effort is a common attribution of performance in academic settings (Dong et al., 2013; Perry et al., 2008); thus, these findings support this research showing this effect.

Given prospective and retrospective importance scores are essentially interchangeable with respect to average levels and their interrelationships with effort and test performance, a practical implication of no difference in prospective and retrospective importance scores is that researchers need not be concerned with when importance data is collected relative to test completion. Thus, testing practitioners can collect importance data when it is most convenient.

In contrast, significant implications associated with measurement of examinee effort arise. First, broadly speaking, it does matter when examinee effort is measured relative to test completion. More specifically, depending on when effort data is collected, researchers and testing practitioners may arrive at different substantive conclusions when employing motivation filtering techniques (e.g., Swerdzewski et al., 2011). Specifically, assuming the same cut point is used regardless of when effort scores are collected, filtering examinee data using prospective effort scores would result in removal of fewer examinees from the data set, some of which may not have put forth effort when

completing the test. As a result, test scores may be negatively biased (Wise et al., 2006). Thus, given the high-stakes nature of inferences institutions make from low-stakes test scores, one cannot infer that intended effort equates to expended effort. In fact, these results indicate intended effort may be an overestimate of expended effort, which should be mentioned by testing practitioners when interpreting test scores.

Smaller standardized indirect effect modeling prospective importance and effort. Does the theoretically and empirically suggested indirect effect of perceived importance on test performance via examinee effort differ depending on if examinees report importance and effort prospectively versus retrospectively? As hypothesized, the completely mediated model fit the scores within each condition. The unstandardized indirect effect was statistically equivalent when modeling prospective scores and retrospective scores. With respect to the equivalent unstandardized indirect effect, this finding extends previous research (e.g., Penk & Schipolowski, 2015) by demonstrating the unstandardized indirect effect is stable when importance and effort are measured prospectively and retrospectively. A major implication of this finding concerns the reporting and interpretation of the indirect effect. Researchers and testing practitioners may use prospective and retrospective importance and effort scores interchangeably when one is concerned with the unstandardized indirect effect.

However, prospective and retrospective importance and effort scores are not interchangeable when one is concerned with the standardized indirect effect. The difference in the magnitude of the standardized indirect effect when using prospective effort scores (.16) versus retrospective scores (.29) is not ignorable. This difference can be explained as follows. As expected, retrospectively rated effort scores ($SD = 3.60$) were

more variable than prospectively rated effort scores ($SD = 2.87$). Given the unstandardized relationship between importance and effort was statistically equivalent across conditions, importance explains essentially equivalent amounts of variance in both prospective and retrospective effort. As such, equal variance explained in prospective and retrospective effort scores coupled with less total variance in prospective versus retrospective scores necessitates less unexplained variance in prospective effort scores and more unexplained variance in retrospective effort scores. The increase in variability in retrospective effort scores could be a function of variables in addition to importance influencing retrospective effort scores and thus the inclusion of those variables in a model would explain this additional variability. On the other hand, the increase in variability in retrospective effort scores could be random measurement error and if so, the unexplained variance would remain unexplained with the inclusion of additional variables.

As illustrated by Pedhazur (1997), and consistent with these findings, holding all else equal, decreased variability in a predictor (i.e., prospective effort in this case) results in an attenuated standardized effect on the outcome (i.e., test performance) and a smaller proportion of variance explained in the outcome (i.e., R^2). This finding becomes perplexing when considered in terms of motivation interventions. The smaller standardized indirect effect for the prospective condition may lead one to believe the manipulation of intended effort via importance may be less beneficial than for the retrospective condition. However, the difference in variability of prospective effort scores (smaller) compared to retrospective effort scores (larger) must be taken into account when interpreting this standardized indirect effect.

Changes in Effort Scores Due to Experiencing the Test

Lower retrospective effort than prospective effort. Do examinees completing both prospective and retrospective measures report different levels of prospective perceived importance and examinee effort, on average, than retrospective importance and effort? Contrary to my hypothesis, average prospective and retrospective importance scores were not significantly different. Thus, researchers and testing practitioners need not be concerned about when importance data is collected relative to test completion. In contrast, consistent with the findings of lower average retrospective than prospective effort scores from the cross-condition comparison (research question 3), examinees tend to report lower average retrospective effort than prospective effort. Accordingly, it clearly matters when examinee effort data is collected relative to test completion.

This conclusion has serious implications for researchers and testing practitioners who make decisions based on effort scores. For example, researchers applying motivation filtering techniques have filtered out examinee data associated with SOS effort scores less than thirteen (Liu et al., 2015; Rios et al., 2014; Wise & Kong, 2005), fourteen (Hathcoat, et al., 2015), and fifteen (Swerdzewski et al., 2011; Wise et al., 2006). As discussed above, motivation filtering based on higher average prospective effort scores may result in removal of less examinee data and, therefore, may negatively bias test scores (Wise et al., 2006). Thus, different conclusions will be made depending on when effort is measured relative to test completion. A possible explanation as to why effort decreases after the test is because examinees may tend to overestimate their ability when reporting intended effort before the test. Recall, EV theory posits expectancy of success on a task influences effort put forth on a task. As such, examinees' intended effort ratings may be a

proxy for their expectancy of success on the test. After experiencing the rigor of the test, examinees may tend to decrease the effort they put forth on the test. Thus, it may be that examinees truly decrease their expended effort throughout the test.

Another serious implication for testing practitioners is the lack of stability of effort scores from before to after the test. Given examinees tend to change relative rank-order of effort scores from before to after test completion, motivation filtering may result in different examinees filtered out of the data set depending on when effort is measured relative to test completion. Thus, testing practitioners cannot be sure if examinees removed from the data set are truly amotivated or if examinees are being removed for factors unrelated to examinee motivation.

It is also important to consider the differences in variability associated with prospective and retrospective effort scores. Recall, intended effort scores measured before the test had significantly less variability than expended effort scores measured after the test. The relatively smaller amount of variability in intended effort may be because intended effort is more representative of examinees' true effort (i.e., less measurement error), whereas the larger amount of variability in retrospective effort scores may be a sign of increased random and systematic measurement error. Nonetheless, given the relationship is stronger between test performance and retrospective scores than prospective scores, it may be that retrospective scores are more aligned with examinees' true effort put forth on the test than prospective scores.

No effect of test performance on change in importance or effort. Are changes in perceived importance and examinee effort related to performance on the quantitative and scientific reasoning test? Recall, I hypothesized that examinees who perform

relatively poorly on the test may be more susceptible to the effect of attributional bias on retrospective importance and effort scores (Perry et al., 2008), whereas this bias may be nil for examinees who perform relatively well. Contrary to my hypothesis, test performance was not significantly related to change in importance or effort over time.

It is important to note the nonsignificant relationship between test performance and change in effort does not indicate test performance is not related to effort at a single time point. Instead, the nonsignificant relationship indicates the rate of change does not depend on test performance. In other words, change in effort (average decrease of effort ratings) is essentially the same across all levels of test performance. Thus, contrary to the assertion that examinees make self-protective post-test attributions to explain their test performance (e.g., Dong et al., 2013; Pekrun, et al., 2004; Perry et al., 2008), examinees who performed poorly did not attribute their performance to expended effort differentially than examinees who did not perform poorly.

These findings have several implications for researchers and test practitioners. Given change in importance and effort scores is not related to test performance, we can be reassured that retrospective ratings are not biased due to self-protective attributions of test performance. However, given differences in average levels of effort, variability of effort ratings from before to after the test, and lack of stability of these ratings, further research is needed to examine the cause of this variability. One such factor that may explain this variability in change in effort is expectancy of success. Although actual test performance was not related to change in importance or effort scores, perhaps examinees' perceptions of competence can explain this variability in change in effort ratings. Examinees' perceptions of their test performance may not be aligned with their actual test

performance. It may be that examinees who tend to decrease their effort ratings from before to after the test perceived their performance to be poorer and examinees who tend to rate their effort equivalently from before to after the test perceived their performance to be better. Thus, future research should evaluate if examinees' perceptions of success account for the variability in change in effort.

Small Difference in Test Performance across Conditions

Does average test performance differ as a function of measurement condition? Recall, examinees in the retrospective condition were not asked to engage in reporting perceived importance or intended effort, examinees in the combined condition were asked to engage in reporting importance, and examinees in the prospective condition were asked to engage in reporting importance and effort prior to test completion. Consistent with previous behavioral commitment research (e.g., Carver & Scheier, 2000; Higgins, 1997), priming students by asking them to engage in reporting their intended effort before the test had a small, but arguably practically significant impact on test performance. Interestingly, the effect of priming examinees does not affect average levels of importance or effort or the indirect effect across conditions, yet priming is affecting average levels of test performance. Thus, priming examinees is affecting test performance via some mechanism (e.g., mediator) other than perceived importance and examinee effort.

These findings have important implications for test practitioners. Given the arguably small unstandardized effect size (approximately 2 points), it is important to interpret these results in light of previous research on the same test. Previous longitudinal studies utilizing students from the same university, the same test (i.e., NW9), and the

same test administration procedures have found nearly equivalent increases in test scores after examinees completed 45 to 70 credit hours (Finney, Sundre et al., 2016; Hathcoat et al., 2015). Considering examinees in the current study had essentially equivalent prior ability (i.e., SAT math and verbal), the simple act of reporting intended effort before the test resulted in a nearly equivalent increase in test scores as the learning gain associated with students who completed 45-70 credit hours. Thus, the two point increase in test scores may be important to faculty members who make decisions based on these scores.

Another important implication for researchers and testing practitioners is associated with motivation interventions. Some motivation interventions have failed to increase performance when offering monetary incentives (Baumert & Demmrich, 2001; O'Neil et al., 2005) and when manipulating test instructions (Finney, Sundre et al., 2016), while manipulation of test instructions in other studies has resulted in increased average test scores similar in magnitude to the current study ($d = 0.26$ vs. $d = 0.31-0.41$); Liu et al., 2012) and larger than the current study ($d = 0.26$ vs. $d = 0.63$; Liu et al., 2015). Given the obvious disadvantages of offering monetary incentives and the possibility of test anxiety negatively influencing test performance associated with manipulation of test instructions (Mathers et al., 2016), simply measuring intended effort before the test may be a cheap and moderately effective method of increasing test performance.

Summary of Findings

Based on this one study examining first-year college students who completed a quantitative and scientific reasoning test, priming students by asking them to report their perceived importance and intended effort prior to the test may be useful to testing practitioners. The findings from the current study indicated that priming did not influence

average ratings of importance and effort measured after test completion or the interrelationships between importance, effort, and test performance. However, examinees who rated importance and effort before the test had, on average, two-point higher test scores, which is approximately equivalent to the learning gains found in students who have completed 45-70 credit hours. The two-point difference may result in a more accurate representation of examinee ability and could have implications for more accurate course placement decisions (i.e., math or foreign language placement testing). Thus, based on the current study, testing practitioners should consider simply measuring perceived test importance and intended effort prior to tests to obtain more accurate estimates of examinee ability.

The current study's findings suggest prospective and retrospective ratings of importance and effort may be interchangeable with respect to the unstandardized interrelationships between importance, effort, and performance. However, researchers should be wary when making decisions based on standardized indirect effects associated with prospective effort ratings. Moreover, effort ratings tended to decrease from before to after experiencing the quantitative and scientific reasoning test. Thus, testing practitioners basing decisions on examinee effort ratings should strongly consider other measures of examinee effort such as RTE.

Limitations and Future Studies

The current study has several limitations that could not be accounted for. First, this study was conducted using a single sample of incoming first-year college students who were administered a quantitative and scientific reasoning test. Given decreases in average reported importance and effort over time (Finney, Sundre et al., 2016; Sessoms

& Finney, 2015), it is reasonable to expect results different from those found in the current study. Moreover, previous research has shown that change in effort was different across different types of tests (Barry & Finney, 2016). Thus, it is strongly recommended that researchers replicate and extend the current study to samples of upper-class students, to different academic subjects, and to samples outside of the university.

An important limitation, and unanswered question, of the current study concerns the measurement of effort. Given retrospective effort ratings were found to be lower than prospective ratings via cross-sectional and longitudinal analyses, researchers and testing practitioners must make a decision with respect to when to measure effort. It may be that retrospective effort scores are more accurate given examinees have experienced the test (Freund & Holling, 2011). If researchers are solely interested in predicting test scores, retrospective effort scores may be the better option as they are more predictive of test performance. However, if researchers are interested in motivation filtering techniques, an alternative measure of examinee effort such as RTE may be a better option. Recall, RTE is a noninvasive measure of effort that does not rely on examinee self-reports. Future research should examine the validity of self-reported effort scores by collecting RTE, prospective, and retrospective effort data within a testing session. The relatively small correlation between RTE and retrospective effort scores ($r = .25$; Wise & Kong, 2005) suggests retrospective effort scores may not be representative of true expended effort. A larger correlation between RTE and prospective effort scores would suggest retrospective scores may be contaminated by construct-irrelevant variance and, therefore, prospective scores may better represent examinee effort.

Importantly, examinees reported, on average, lower effort after the test than before the test. Moreover, examinees did not change similarly over time, as indicated by the lack of stability in effort scores over time. Future research studies should focus on explaining these individual differences in change in effort over time. In the current study, test performance was the only predictor of change of effort examined, but only accounted for 3% of the variance in change in effort. Other predictors that may account for the variation in change in effort over time should be explored. One such predictor is conscientiousness. Previous research examining conscientiousness in low-stakes testing is inconclusive: conscientiousness has been related to effort at a single time point (e.g., Barry & Finney, 2016), not related to effort or performance (e.g., Myers et al., 2016), and negligibly related to performance (e.g., Finney, Sundre et al., 2016). It may be that examinees higher in conscientiousness provide accurate, and essentially equivalent, ratings of effort over time. Whereas, examinees lower in conscientiousness may decrease their ratings of effort over time to a greater degree than other examinees. In addition, Barry and Finney found agreeableness to be significantly related to change in effort across multiple different tests within a testing session.

Another limitation to the current study is the assumption that examinees accurately judge their own test performance. Recall, attributional bias may serve as a self-protective mechanism to a greater degree for examinees who perform poorly on the test than for examinees who perform better on the test. As mentioned above, if examinees perceptions of their performance is misaligned with their actual performance, the relationship between test performance and change in importance and effort may be nil. Although previous research has repeatedly shown that ability is not related to effort at a

given time point (Wise & Smith, 2016), future research should examine if change in effort is related to an independent measure of ability (e.g., SAT math) or examinees' perceptions of their performance.

Although the prospective and retrospective SOS performed well in this study, the SOS is not without faults. Prospective effort item seven was noninvariant over time. Error covariances were needed to account for shared variance likely due to negative wording effects and item wording redundancy. In addition, there may be other sources of construct-irrelevant variance associated with SOS scores. For example, it is questionable whether importance item three "I am not curious about how I did on this test relative to others" is a valid measure of perceived test importance. An examinee scoring high on this item seems to be more concerned with their performance in comparison to other examinees. Thus, item three appears to be representative of the performance approach achievement goal orientation (e.g., "It is important for me to do well compared to other students;" Finney, Pieper, & Barron, 2004). Additionally, consideration should be given to the phrasing of items on the SOS. Research suggests items in the form of questions (e.g., "How much effort will you put forth on this test?") tend to provide for more thoughtful responses than items in the form of statements (e.g., SOS item: "I will give my best effort on this test;" Petty, Rennie, & Cacioppo, 1987).

Given these issues, the SOS may not be the best measure of motivation in low-stakes testing contexts. A new measure of examinee motivation for low-stakes testing could address the aforementioned issues and incorporate the expectancy component of EV theory. Moreover, researchers should consider the recently revised version of EV theory, expectancy-value-cost (EVC) model which has emerged as a promising

framework used to explain examinee motivation. Previously, cost was considered a subcomponent of value. However, recent research suggests cost is a separate component negatively related to expectancy and value (Barron & Hulleman, 2015; Flake, Barron, Hulleman, McCoach, & Welsh, 2015). This revised conceptualization of cost is defined as the effort necessary to complete a task, affective demand associated with exerting effort on the task, and loss of ability to engage in other activities (Flake et al., 2015). For example, as examinees experience the test, their perceptions or realizations of the effort necessary to complete the test and corresponding affective demand may cause overall levels of cost to increase and, therefore, expended effort to decrease. Thus, cost may help explain why effort decreases over time. In sum, a new measure of examinee motivation may address current issues with the SOS and incorporate potentially relevant components of expectancy and cost.

An important limitation to the current study concerns the temporal ordering of importance, effort, and performance. Testing of mediation hypotheses with cross-sectional data (research question 2) only provides accurate parameter estimates when very strict assumptions are met (Cole & Maxwell, 2003). Moreover, using retrospective scores as proxies for variables of interest (e.g., importance and effort) theoretically occurring at an earlier time tends to result in biased estimates of the indirect effect (Cole & Maxwell, 2003). The mediated effect of prospective importance on performance via prospective effort (research question 4) was much closer to meeting the assumption of temporal ordering of the variables. However, importance and effort were collected simultaneously; thus, limiting the inferences that can be made from these results. The process of mediation is inherently longitudinal ($I \rightarrow E \rightarrow P$); thus, future research should

examine the indirect effect via a longitudinal mediation model in the latent growth curve modeling (LGM) framework. The LGM framework allows for modeling of individual differences in change over time and, importantly, allows for the modeling of predictors of these individual differences.

Conclusion

Although low-stakes tests used in this study and others may be low-stakes in nature to examinees, these tests are not low-stakes to educational institutions. Universities use these accountability test scores for accreditation purposes, to assess curriculum effectiveness, and to inform evidence-based changes to academic programming. Accordingly, it is of utmost importance that these test scores are accurate reflections of student ability. Unfortunately, the perplexing problem associated with biased estimates of ability due to low and amotivated students persists. I argue that one appealing option testing practitioners may utilize to obtain more accurate estimates of ability is via priming examinees by measuring examinee motivation prior to the test. Another option testing practitioners currently utilize to account for amotivated examinees is motivation filtering techniques. However, given average levels of reported effort tend to decrease after examinees experience the test, I strongly recommend testing practitioners obtain other measures of examinee effort prior to utilizing such techniques.

Table 1.

Third Variable Functions, Example Hypotheses, Path Diagrams, and Examples of Data Characteristics

Third Variable (Z)	Example Hypotheses	Model	Occurrence of Z	Relationships						
				<i>c</i>	<i>c'</i>	<i>ab</i>	<i>a</i>	<i>b</i>		
				β_{yx}	$\beta_{yx.z}$	<i>ab</i>	β_{zx}	β_{yz}	$\beta_{yz.x}$	
Complete Mediator	Does X affect Y indirectly via Z?	$X \xrightarrow{a} Z \xrightarrow{b} Y$	Between X and Y	.16	.00	.16	.40	.40	--	
Partial Mediator	Does X affect Y directly and indirectly via Z?	$X \xrightarrow{a} Z \xrightarrow{b} Y$ $X \xrightarrow{c'} Y$	Between X and Y	.36	.24	.12	.40	.40	.30	
Suppressor	Is X more predictive of Y when Z is modeled?	$X \xrightarrow{a} Z \xrightarrow{b} Y$ $X \xrightarrow{c'} Y$	Between X and Y	Classical	.00	-.16	.16	.40	--	.40
				Negative	.20	-.30	.50	.71	--	.71
Covariate	After controlling for Z, is X related to Y?	$Z \xrightarrow{b} Y$ $X \xrightarrow{c'} Y$	Before or coinciding with X	.40	.40	--	.00	--	.40	
Confounder	Is the relationship between X and Y spurious?	$X \xleftarrow{a} Z \xrightarrow{b} Y$	Before or coinciding with X	r_{yx}	$r_{yx.z}$	--	.40	.40	--	
Moderator	Does the effect of X on Y depend on Z?	$X \xrightarrow{c} Y$ $Z \downarrow$	Before or coinciding with X	+ 1 SD _Z	β_{yx}	--	--	.20	.20	--
				- 1 SD _Z	.40	.25	--	--	--	--

Note. Path coefficient c represents the total effect of X on Y and c' represents the direct effect of X on Y for the mediation and suppression examples. The product of a and b (ab) represents the indirect effect of X on Y via Z . Path coefficients a represent the relationship between X and Z and b represent the relationship between Z and Y . Path coefficient b can be represented by the regression equation β_{yz} for complete mediation or $\beta_{yz.x}$ for partial mediation. Path coefficient values are intended to provide a conceptual understanding of how interrelationships between X , Z , and Y might look depending on the function of the third variables. Mediation distinguishing characteristic is Z occurs between X and Y . With respect to partial mediation, although the b path coefficient would be represented by $\beta_{yz.x}$, the value for β_{yz} was included to illustrate that both X and Z contribute uniquely to the prediction of Y .

Suppression distinguishing characteristics are (1) the direct effects are stronger than the total effects (X - Y) and (2) Z occurs between X and Y . Covariates are typically not manipulable and used for statistical control, distinguishing characteristics of covariates are (1) Z occurs before or coincides with X and (2) the X - Z relationship is not substantial. Confounder presumes a completely spurious effect ($r_{yx.z} = .00$) where Z completely accounts for the bivariate X - Y relationship ($r_{yx} = .16$). Importantly, the relationships between the variables (represented by a , b , c , c' , and ab) for confounders and mediators are mathematically identical, the distinguishing characteristics are (1) the direction of the X - Z arrow and (2) Z occurs before or coincides with X . Moderation distinguishing characteristics are (1) Z occurs before or coincides with X and (2) the X - Z path coefficient changes at different levels of Z (e.g., $\pm 1 SD_z$). With respect to the values for the X - Z and Z - Y relationships, it is only necessary that X - Z and Z - Y are related.

			Measures					
			Prospective		Retrospective			
Research Questions	Analyses	Expected Results	Perceived Importance	Intended Effort	Perceived Importance	Expended Effort	Quantitative & Scientific Reasoning	Results can be found in Table X
1. On average, do examinees report different levels of retrospectively measured perceived test importance and examinee expended effort across measurement conditions?	Structural Means Modeling (SMM) using only retrospective scores	Priming examinees may invoke behavioral commitment (i.e., increase test scores) and importance and effort scores may be contaminated by attributional bias. Thus, I expect average levels of retrospective importance and effort will be higher in the prospective condition.						6
Retrospective Condition					X	X		
Combined Condition					X	X		
Prospective Condition					X	X		
2. Is the indirect effect of retrospectively measured perceived test importance on test performance via retrospectively measured examinee effort affected by asking examinees to rate their	Moderated Mediation Analysis using only retrospective scores	I expect the completely mediated model will fit the data across the three conditions. However, behavioral commitment may mitigate the effects of attributional bias in						8

perceived importance and intended effort before test completion?		the prospective condition. Thus, I expect the magnitude of the indirect effect estimated using retrospective scores will be smaller in the prospective condition.						
Retrospective Condition					X	X	X	
Combined Condition					X	X	X	
Prospective Condition					X	X	X	
3. Do examinees reporting retrospective perceived importance and examinee effort have lower scores, on average, than examinees reporting prospective importance and effort for the same test?	Structural Means Modeling (SMM) using retrospective and prospective scores	I expect average levels of prospectively measured perceived importance and examinee effort scores in the prospective condition will be higher than importance and effort scores from the retrospective condition.						10
Retrospective Condition					X	X		
Combined Condition			X			X		
Prospective Condition			X	X				
4. Does the theoretically and empirically suggested indirect effect of perceived importance on test performance via examinee effort differ across the three measurement conditions?	Moderated Mediation Analysis using retrospective and	Examinees make self-protective attributions to explain their test performance. Thus, I expect the standardized indirect effect in the prospective condition						12

	prospective scores	will be smaller in magnitude than the indirect effect in the retrospective condition.						
Retrospective Condition					X	X	X	
Combined Condition			X			X	X	
Prospective Condition			X	X			X	
5. Do examinees who complete both prospective and retrospective measures report different levels of prospective perceived importance and examinee effort than retrospective perceived importance and examinee effort?	Longitudinal Mean and Covariance Structure Analysis (LMACS)	Examinees may make post-test attributions to explain their test performance. Further, examinees may exhibit behavior in accordance with their behavioral intentions. As a result, prospective importance and effort scores, on average, may not differ from retrospective scores. I expect average levels of importance and effort scores will be essentially equivalent over time.						14
Prospective Condition			X	X	X	X		
6. Are perceived importance and examinee effort change scores related to performance on the quantitative and scientific reasoning test?	Latent Growth Model (LGM)	Attributional bias may serve as a self-protective mechanism to a greater degree for examinees who perform						Figure 10

		poorly on the test than for examinees who perform well. Thus, I expect larger differences in importance and effort scores for examinees who perform poorly than for the examinees who perform well.						
			X	X	X	X		
7. Does test performance differ, on average, as a function of measurement condition?	Analysis of Variance (ANOVA)	Given students tend to exhibit behavior in accordance with their behavioral intentions and assuming examinees report higher levels of prospective perceived importance and intended effort, I expect average test scores will be higher in the prospective condition.						
Retrospective Condition							X	
Combined Condition							X	
Prospective Condition							X	
<p><i>Note.</i> The information in this table aligns the current studies' research questions with the hypotheses and data analyses, and scores used with those analyses. The data used in the analyses is indicated by the X in the box corresponding with the measurement condition.</p>								

Table 3
Demographics by Measurement Condition

	Condition			Total
	Retrospective	Combined	Prospective	
Gender				
Female	61.25%	60.05%	63.96%	61.52%
Male	38.75%	39.95%	36.04%	38.48%
Ethnicity				
White	84.00%	88.22%	87.01%	86.42%
Asian	9.00%	6.70%	7.47%	7.71%
Hispanic	5.50%	5.77%	6.17%	5.78%
Black	5.50%	5.31%	5.52%	5.43%
American Indian	1.50%	0.69%	1.95%	1.31%
Pacific Islander	0.75%	0.69%	0.65%	0.70%
Not specified	3.00%	1.39%	1.95%	2.10%
Mean Age (years)	18.46	18.46	18.46	18.46
SAT Math	571.48	559.37	569.19	566.33
SAT Verb	567.59	560.47	564.31	564.03
<i>N</i>	400	437	308	1145

Note. Students could self-identify in more than one ethnicity category.

Table 4
Correlations and Descriptive Statistics for Quantitative and Scientific Reasoning (NW9) and Retrospective and Prospective Perceived Test Importance and Examinee Effort Scores by Measurement Condition

	NW9	Retrospective		Prospective	
		Importance	Effort	Importance	Effort
Retrospective Condition (n = 400)					
NW9	.761	--	--	--	--
Retrospective Importance	.185	.732	--	--	--
Retrospective Effort	.376	.491	.846	--	--
Prospective Importance	--	--	--	--	--
Prospective Effort	--	--	--	--	--
Mean	44.150	17.605	19.200	--	--
SD	7.011	3.226	3.607	--	--
Skew	-0.266	0.155	-0.542	--	--
Kurtosis	0.089	-0.026	0.076	--	--
Combined Condition (n =439)					
NW9	.799	--	--	--	--
Retrospective Importance	.200	.763	--	--	--
Retrospective Effort	.347	.544	.854	--	--
Prospective Importance	.184	.697	.422	.741	--
Prospective Effort	--	--	--	--	--
Mean	45.366	17.739	19.002	17.670	--
SD	7.642	3.287	3.578	3.332	--
Skew	-0.457	-0.291	-0.496	-0.508	--
Kurtosis	0.586	0.516	0.175	0.845	--
Prospective Condition (n = 308)					
NW9	.771	--	--	--	--
Retrospective Importance	.127	.813	--	--	--
Retrospective Effort	.314	.480	.836	--	--
Prospective Importance	.161	.786	.381	.762	--
Prospective Effort	.196	.402	.580	.473	.804
Mean	46.065	17.945	19.604	17.877	20.494
SD	7.091	3.456	3.235	3.302	2.872
Skew	-0.303	-0.178	-0.478	-0.257	-0.760
Kurtosis	-0.239	-0.259	0.364	-0.066	1.159

Note. Values on the diagonal represent coefficient α . NW9 scores range from 0 to 66. Perceived importance and examinee effort scores range from 5 to 25, with higher scores reflecting higher levels of importance and effort.

Table 5
Correlations and Descriptive Statistics for Retrospective SOS by Measurement Condition

Retrospective Condition (<i>n</i> = 400)										
Item	Item									
	1 _I	3 _I *	4 _I *	5 _I	8 _I	2 _E	6 _E	7 _E *	9 _E *	10 _E
3 _I *	.204	--								
4 _I *	.334	.424	--							
5 _I	.590	.214	.336	--						
8 _I	.446	.414	.373	.298	--					
2 _E	.436	.208	.227	.255	.445	--				
6 _E	.445	.227	.199	.368	.455	.719	--			
7 _E *	.240	.165	.114	.231	.266	.494	.577	--		
9 _E *	.401	.219	.237	.308	.344	.592	.603	.554	--	
10 _E	.311	.211	.170	.160	.384	.516	.474	.295	.488	--
Mean	3.828	3.373	3.348	3.175	3.883	4.175	4.000	3.178	3.810	4.038
SD	.812	1.082	.990	.895	.837	.822	.929	1.072	.986	.740
Skew	-.267	-.366	-.242	.239	-.395	-.987	-.868	-.150	-.669	-.657
Kurt	-.164	-.568	-.434	.057	-.271	1.122	.537	-.864	-.131	.789
Combined Condition (<i>n</i> = 437)										
3 _I *	.322	--								
4 _I *	.466	.400	--							
5 _I	.615	.132	.342	--						
8 _I	.489	.370	.458	.355	--					
2 _E	.493	.246	.272	.367	.388	--				
6 _E	.534	.230	.264	.387	.415	.714	--			
7 _E *	.383	.185	.212	.279	.258	.533	.618	--		
9 _E *	.448	.247	.207	.326	.321	.616	.652	.605	--	
10 _E	.399	.206	.192	.282	.376	.477	.424	.349	.424	--
Mean	3.659	3.584	3.384	3.229	3.883	4.108	3.982	3.153	3.792	3.968
SD	.878	.948	.955	.930	.874	.821	.911	1.057	.951	.729
Skew	-.401	-.768	-.469	-.126	-.889	-1.001	-.841	.019	-.764	-.880
Kurt	-.025	.442	-.131	-.084	1.185	1.389	.399	-.800	.260	1.890
Prospective Condition (<i>n</i> = 308)										
3 _I *	.339	--								
4 _I *	.483	.485	--							
5 _I	.683	.298	.467	--						
8 _I	.514	.527	.514	.385	--					
2 _E	.546	.250	.263	.347	.430	--				
6 _E	.533	.254	.264	.402	.464	.759	--			
7 _E *	.279	.189	.128	.256	.217	.419	.479	--		
9 _E *	.371	.225	.193	.198	.318	.606	.601	.511	--	
10 _E	.306	.231	.191	.191	.365	.525	.538	.286	.478	--
Mean	3.802	3.565	3.513	3.231	3.834	4.156	4.123	3.273	3.964	4.097
SD	.864	.995	.970	.889	.844	.771	.768	.997	.863	.751
Skew	-.370	-.521	-.382	.063	-.497	-.962	-.735	-.072	-.788	-.952
Kurt	-.322	-.185	-.411	-.156	.107	1.548	.661	-.816	.500	2.091

Note. *Denotes items that are reversed prior to scoring. Respondents rate their agreement with the 10 items on a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) with higher scores indicating higher levels of reported effort and importance. ^I Denotes items from importance subscale. ^E Denotes items from effort subscale. Kurt = kurtosis.

Table 6

Model Fit Indices for One-Factor, Two-Factor, and Multiple-Condition Invariance Testing of the Retrospective SOS Scores across Measurement Conditions

Model	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	ΔCFI	CFI	RMSEA	SRMR	Correlation residuals > .15	Mean residuals > .15
Retrospective										
1-factor	339.302	35	--	--	--	.797	.147	.093	5	--
2-factor	190.716	34	148.586*	1	.099	.896	.107	.063	1	--
2-factor with I3, I4, & I8	131.591	32	59.125*	2	.017	.934	.088	.056	1	--
Combined										
1-factor	344.360	35	--	--	--	.825	.142	.087	5	--
2-factor	137.077	34	202.451*	1	.116	.942	.083	.053	1	--
2-factor with I3, I4, & I8	100.285	32	36.792*	2	.006	.961	.070	.046	1	--
Prospective										
1-factor	385.774	35	--	--	--	.741	.180	.111	8	--
2-factor	167.890	34	194.574*	1	.161	.901	.113	.064	2	--
2-factor with I3, I4, & I8	111.642	32	56.248*	1	.023	.941	.089	.056	1	--
Invariance Testing										
Configural	342.897	96	--	--	--	.970	.083	--	3	--
Metric	369.718	116	26.281	20	.001	.969	.076	--	2	--
Scalar	422.320	132	52.602*	16	.004	.965	.076	--	2	1
Partial Scalar	410.612	130	11.563*	2	.000	.965	.076	--	2	0
Measurement Error	473.414	152	103.696*	36	.008	.961	.075	--	2	1

Note. * $p < .01$. χ^2 = maximum likelihood chi-square; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Local misfit was assessed via examination of correlation and mean residuals $> |.15|$ in magnitude. The 1-factor model estimates 20 parameters (1 factor variance, 9 factor pattern coefficients, and 10 error variances) from 55 observations (10 variances and 45 covariances). The 2-factor model estimates 21 parameters (1 factor covariance, 2 factor variances, 8 factor pattern coefficients, and 10 error variances) from 55 observations. The 2-factor with I3, I4, & I8 model adds 2 error covariances (between importance items 3 and 4 and items 3 and 8; 23 parameters are estimated). In the configural model, 99 parameters (3 factor covariances, 6 factor variances, 24 factor pattern coefficients, 30 error variances, 30 intercepts, and 6 error covariances) are estimated from 195 observations. In the metric invariance model, 20 fewer parameters are estimated than the configural model by constraining 8 factor pattern coefficients and 2 error covariances to be equal across conditions (79 total). In the scalar invariance model, 16 fewer parameters are estimated by constraining 8 intercepts to be equal across conditions (63 total). In the partial scalar invariance model, 2 additional parameters are estimated by releasing the equality constraint on importance item three's intercept across groups. In the measurement error invariance model, 20 fewer parameters than the scalar model are estimated by constraining 10 error variances to be equal across conditions (43 total).

Table 7

Unstandardized and (Standardized) Factor Pattern Coefficients for the Retrospective SOS Scores across Measurement Conditions

Item	Retrospective		Combined		Prospective	
	Importance	Effort	Importance	Effort	Importance	Effort
1. Doing well on this test was important to me.	1.000 (.809)		1.000 (.879)		1.000 (.887)	
3. I am not curious about how I did on this test relative to others.*	0.518 (.317)		0.439 (.359)		0.536 (.420)	
4. I am not concerned about the score I receive on this test.*	0.694 (.460)		0.668 (.540)		0.737 (.582)	
5. This was an important test to me.	0.907 (.665)		0.812 (.674)		0.861 (.742)	
8. I would like to know how well I did on this test.	0.752 (.590)		0.669 (.590)		0.675 (.613)	
2. I engaged in good effort throughout this test.		1.000 (.824)		1.000 (.814)		1.000 (.858)
6. I gave my best effort on this test.		1.173 (.857)		1.175 (.863)		1.024 (.881)
7. While taking this test, I could have worked harder on it.*		1.011 (.640)		1.109 (.702)		0.810 (.538)
9. I did not give this test my full attention while completing it.*		1.076 (.740)		1.096 (.771)		0.921 (.706)
10. While taking this test I was able to persist to completion of the task.		0.640 (.587)		0.590 (.541)		0.693 (.610)

Note. *Items were reversed prior to scoring. All parameter estimates were statistically significant at $p < .01$. The metric of factors were established by fixing importance item 1 and effort item 2 factor loadings to one. The parameters are estimated from fitting a two-factor model with error covariances between importance items three and eight and importance items three and four to scores from each of the three conditions with no invariance constraints. The error covariances between importance item three and eight were statistically significant and moderate in size across conditions (residual correlations = .38, .23, .49) as were the residual correlations for item three and four (.30, .23, .26).

Table 8
Fit Indices for Completely Mediated Model using Retrospective SOS Scores by Measurement Condition and Invariance Testing

Model	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	ΔCFI	CFI	RMSEA	SRMR
Single-Condition								
Retrospective Condition	0.657	1	--	--	--	1.000	< .001	.012
Combined Condition	0.558	1	--	--	--	1.000	< .001	.011
Prospective Condition	0.460	1	--	--	--	1.000	< .001	.010
Multiple-Condition								
Unconstrained	2.177	3	--	--	--	1.000	< .001	--
Constrained Direct Paths	9.907	7	7.730	4	.006	.994	.033	--
Completely Constrained	17.275	13	7.368	6	.003	.991	.029	--

Note. χ^2 = maximum likelihood chi-square; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. χ^2 and degrees of freedom from the unconstrained multiple-condition model are the sum from the completely mediated models fit to each of the three conditions. The unconstrained model was estimated to obtain baseline CFI, and RMSEA values. The completely constrained model fixed the unstandardized direct paths, variance of importance, and disturbance variances (effort and performance) to be equivalent across conditions. * $p < .01$.

Table 9
Correlations and Descriptive Statistics for Combined SOS by Measurement Condition
 Retrospective Condition ($n = 400$)

Item	Item									
	1 _I	3 _I *	4 _I *	5 _I	8 _I	2 _E	6 _E	7 _E *	9 _E *	10 _E
3 _I *	.204	--								
4 _I *	.334	.424	--							
5 _I	.590	.214	.336	--						
8 _I	.446	.414	.373	.298	--					
2 _E	.436	.208	.227	.255	.445	--				
6 _E	.445	.227	.199	.368	.455	.719	--			
7 _E *	.240	.165	.114	.231	.266	.494	.577	--		
9 _E *	.401	.219	.237	.308	.344	.592	.603	.554	--	
10 _E	.311	.211	.170	.160	.384	.516	.474	.295	.488	--
Mean	3.828	3.373	3.348	3.175	3.883	4.175	4.000	3.178	3.810	4.038
SD	.812	1.082	.990	.895	.837	.822	.929	1.072	.986	.740
Skew	-.267	-.366	-.242	.239	-.395	-.987	-.868	-.150	-.669	-.657
Kurt	-.164	-.568	-.434	.057	-.271	1.122	.537	-.864	-.131	.789
Combined Condition ($n = 437$)										
3 _I *	.229	--								
4 _I *	.353	.382	--							
5 _I	.558	.228	.383	--						
8 _I	.494	.284	.350	.416	--					
2 _E	.308	.209	.206	.314	.286	--				
6 _E	.306	.197	.175	.314	.302	.714	--			
7 _E *	.216	.200	.149	.254	.239	.533	.618	--		
9 _E *	.261	.199	.180	.292	.239	.616	.652	.605	--	
10 _E	.270	.168	.190	.247	.276	.477	.424	.349	.424	--
Mean	3.703	3.462	3.359	3.213	3.934	4.108	3.982	3.153	3.792	3.968
SD	.943	1.019	.963	.879	.947	.821	.911	1.057	.951	.729
Skew	-.679	-.629	-.339	-.042	-1.218	-1.001	-.841	.019	-.764	-.880
Kurt	.607	-.121	-.447	.148	1.799	1.389	.399	-.800	.260	1.890
Prospective Condition ($n = 308$)										
3 _I *	.265	--								
4 _I *	.437	.444	--							
5 _I	.542	.167	.416	--						
8 _I	.415	.425	.454	.372	--					
2 _E	.486	.203	.247	.347	.465	--				
6 _E	.407	.160	.219	.373	.511	.677	--			
7 _E *	.039	.059	.021	.128	.168	.272	.346	--		
9 _E *	.312	.154	.223	.336	.392	.549	.632	.319	--	
10 _E	.329	.178	.184	.319	.423	.485	.596	.287	.476	--
Mean	3.912	3.403	3.403	3.269	3.890	4.360	4.334	3.357	4.227	4.224
SD	.824	.995	1.011	.855	.914	.746	.728	.871	.762	.721
Skew	-.503	-.377	-.340	.081	-.863	-1.596	-1.069	-.315	-1.165	-.995
Kurt	.360	-.251	-.492	.165	.673	4.338	1.521	.019	2.229	2.137

Note. *Denotes items that are reversed prior to scoring. Respondents rate their agreement with the 10 items on a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) with higher scores indicating higher levels of reported effort and importance. ^I Denotes items from importance subscale. ^E Denotes items from effort subscale. Kurt = kurtosis.

Table 10

Model Fit Indices for One-Factor, Two-Factor, and Multiple-Condition Invariance Testing of the Prospective and Retrospective SOS Scores across Measurement Conditions

Model	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	ΔCFI	CFI	RMSEA	SRMR	Correlation residuals > .15	Mean residuals > .15
Retrospective										
1-factor	339.302	35	--	--	--	.797	.147	.093	5	--
2-factor	190.716	34	148.586*	1	.099	.896	.107	.063	1	--
2-factor with I3, I4, & I8	131.591	32	59.125*	2	.017	.934	.088	.056	1	--
Combined										
1-factor	344.360	35	--	--	--	.825	.142	.087	5	--
2-factor	137.077	34	202.451*	1	.116	.942	.083	.053	1	--
2-factor with I3, I4, & I8	48.829	32	88.248*	2	.047	.989	.035	.036	0	--
Prospective										
1-factor	385.774	35	--	--	--	.741	.180	.111	8	--
2-factor	167.890	34	194.574*	1	.161	.901	.113	.064	2	--
2-factor with I3, I4, & I8	88.819	32	79.071*	1	.023	.947	.076	.055	1	--
Invariance Testing										
Configural	268.746	96	--	--	--	.975	.069	--	2	--
Metric	326.055	116	57.309*	20	.005	.970	.069	--	3	--
Partial Metric	304.153	115	35.407	19	.002	.973	.066	--	2	--
Partial Scalar	339.717	130	35.564*	15	.003	.970	.065	--	2	0
Measurement Error	426.465	152	100.410*	36	.009	.961	.069	--	24	0

Note. * $p < .01$. χ^2 = maximum likelihood chi-square; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Local misfit was assessed via examination of correlation and mean residuals $> |.15|$ in magnitude. The 1-factor model estimates 20 parameters (1 factor variance, 9 factor pattern coefficients, and 10 error variances) from 55 observations (10 variances and 45 covariances). The 2-factor model estimates 21 parameters (1 factor covariance, 2 factor variances, 8 factor pattern coefficients, and 10 error variances) from 55 observations. The 2-factor with I3, I4, & I8 model adds 2 error covariances (between importance items 3 and 4 and items 3 and 8; 23 parameters are estimated). In the configural model, 99 parameters (3 factor covariances, 6 factor variances, 24 factor pattern coefficients, 30 error variances, 30 intercepts, and 6 error covariances) are estimated from 195 observations. In the metric invariance model, 20 fewer parameters are estimated than the configural model by constraining 8 factor pattern coefficients and 2 error covariances to be equal across conditions (79 total). In the partial metric invariance model, 1 additional parameter was estimated by releasing the equality constraint on effort item 7's factor pattern coefficient in the prospective condition. In the partial scalar invariance model, 15 fewer parameters are estimated by constraining 8 intercepts (except for effort item 7's intercept was freely estimated in the prospective condition) to be equal across conditions. In the measurement error invariance model, 36 fewer parameters than the metric model are estimated by constraining 8 intercepts and 10 error variances to be equal across conditions (43 total).

Table 11
Unstandardized and (Standardized) Factor Pattern Coefficients for the Prospective and Retrospective SOS Scores across Measurement Conditions

Item	Retrospective		Combined		Prospective	
	Importance	Effort	Importance	Effort	Importance	Effort
1. Doing well on this test was important to me.	1.000 (.809)		1.000 (.756)		1.000 (.719)	
3. I am not curious about how I did on this test relative to others.*	0.518 (.317)		0.478 (.335)		0.558 (.335)	
4. I am not concerned about the score I receive on this test.*	0.694 (.460)		0.681 (.504)		0.996 (.584)	
5. This was an important test to me.	0.907 (.665)		0.888 (.720)		0.970 (.672)	
8. I would like to know how well I did on this test.	0.752 (.590)		0.844 (.635)		1.025 (.665)	
2. I engaged in good effort throughout this test.		1.000 (.824)		1.000 (.814)		1.000 (.771)
6. I gave my best effort on this test.		1.173 (.857)		1.167 (.857)		1.110 (.878)
7. While taking this test, I could have worked harder on it.*		1.011 (.640)		1.120 (.708)		0.577 (.382)
9. I did not give this test my full attention while completing it.*		1.076 (.740)		1.103 (.776)		0.950 (.718)
10. While taking this test I was able to persist to completion of the task.		0.640 (.587)		0.585 (.537)		0.838 (.669)

Note. *Items were reversed prior to scoring. All parameter estimates were statistically significant at $p < .01$. The metric of factors were established by fixing importance item 1 and effort item 2 factor loadings to one. The parameters are estimated from fitting a two-factor model with error covariances between importance items three and eight and importance items three and four to scores from each of the three conditions with no invariance constraints. The error covariances between importance item three and eight were statistically significant and moderate in size across conditions (residual correlations = .38, .22, .47) as were the residual correlations for item three and four (.30, .21, .24).

Table 12
Fit Indices for Completely Mediated Model using Prospective and Retrospective SOS Scores by Measurement Condition and Invariance Testing

Model	χ^2	df	$\Delta\chi^2$	Δdf	ΔCFI	CFI	RMSEA	SRMR
Single-Condition								
Retrospective Condition	0.097	1	--	--	--	1.000	< .001	.012
Combined Condition	0.097	1	--	--	--	1.000	< .001	.005
Prospective Condition	0.846	1	--	--	--	1.000	< .001	.017
Multiple-Condition								
Unconstrained	1.597	3	--	--	--	1.000	< .001	--
Constrained Direct Paths	8.456	7	6.859	4	.004	.996	.023	--
Completely Constrained	33.344	13	24.888*	6	.048	.948	.064	--
Constrained ¹ Paths and Disturbance	13.074	12	4.618	5	.001	.997	.015	--

Note. Retrospective condition = retrospective importance and effort scores. Combined condition = prospective importance and retrospective effort scores. Prospective condition = prospective importance and effort scores. χ^2 = maximum likelihood chi-square; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. χ^2 and degrees of freedom from the unconstrained multiple-condition model are the sum from the completely mediated models fit to each of the three conditions. The unconstrained model was estimated to obtain baseline CFI, and RMSEA values. The completely constrained model fixed the unstandardized direct paths, variance of importance, and disturbance variances (effort and performance) to be equivalent across conditions. ¹ Efforts' disturbance variance in the prospective condition was allowed to freely vary and this model was compared to the constrained direct paths model. * $p < .01$.

Table 13
Correlations and Descriptive Statistics for Retrospective and Prospective Importance and Effort Scores from the Prospective Measurement Condition

Item	Item																			
	Prospective										Retrospective									
	1 ₁₁	3 ₁₁ *	4 ₁₁ *	5 ₁₁	8 ₁₁	2 _{E1}	6 _{E1}	7 _{E1} *	9 _{E1} *	10 _{E1}	1 ₁₂	3 ₁₂ *	4 ₁₂ *	5 ₁₂	8 ₁₂	2 _{E2}	6 _{E2}	7 _{E2} *	9 _{E2} *	10 _{E2}
1 ₁₁	--																			
3 ₁₁ *	.27	--																		
4 ₁₁ *	.44	.44	--																	
5 ₁₁	.54	.17	.42	--																
8 ₁₁	.42	.43	.45	.37	--															
2 _{E1}	.49	.20	.25	.35	.46	--														
6 _{E1}	.41	.16	.22	.37	.51	.68	--													
7 _{E1} *	.04	.06	.02	.13	.17	.27	.35	--												
9 _{E1} *	.31	.15	.22	.34	.39	.55	.63	.32	--											
10 _{E1}	.33	.18	.18	.32	.42	.49	.60	.29	.48	--										
1 ₁₂	.61	.25	.45	.62	.48	.36	.43	.14	.35	.27	--									
3 ₁₂ *	.26	.50	.44	.23	.44	.16	.17	.09	.07	.20	.34	--								
4 ₁₂ *	.39	.35	.58	.43	.46	.25	.26	.03	.17	.27	.48	.49	--							
5 ₁₂	.43	.23	.39	.68	.38	.32	.40	.15	.32	.26	.68	.30	.47	--						
8 ₁₂	.29	.42	.43	.37	.64	.31	.36	.13	.28	.30	.51	.53	.51	.39	--					
2 _{E2}	.36	.17	.23	.31	.39	.49	.52	.26	.43	.41	.55	.25	.26	.35	.43	--				
6 _{E2}	.32	.16	.18	.34	.36	.46	.53	.29	.39	.41	.53	.25	.26	.40	.46	.76	--			
7 _{E2} *	.17	.13	.09	.20	.12	.27	.25	.36	.25	.15	.28	.19	.13	.26	.22	.42	.48	--		
9 _{E2} *	.27	.19	.18	.17	.24	.33	.35	.24	.35	.29	.37	.23	.19	.20	.32	.61	.60	.51	--	
10 _{E2}	.22	.09	.16	.23	.28	.34	.38	.23	.33	.42	.31	.23	.19	.19	.37	.53	.54	.29	.48	--
Mean	3.91	3.40	3.40	3.27	3.89	4.36	4.33	3.36	4.23	4.22	3.80	3.57	3.51	3.23	3.83	4.16	4.12	3.27	3.96	4.10
SD	0.82	1.00	1.01	0.86	0.91	0.75	0.73	0.87	0.76	0.72	0.86	1.00	0.97	0.89	0.84	0.77	0.77	1.00	0.86	0.75
Skew	-0.50	-0.38	-0.34	0.08	-0.86	-1.60	-1.07	-0.32	-1.17	-1.00	-0.37	-0.52	-0.38	0.06	-0.50	-0.96	-0.74	-0.07	-0.79	-0.95
Kurt	0.36	-0.25	-0.49	0.17	0.67	4.34	1.52	0.02	2.23	2.14	-0.32	-0.19	-0.41	-0.16	0.11	1.55	0.66	-0.82	0.50	2.09

Note. *Denotes items that are reversed prior to scoring. Respondents rate their agreement with the 10 items on a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) with higher scores indicating higher levels of reported effort and importance. Subscripts denote importance (I) and effort (E) subscales and prospective (1) and retrospective (2) items. Prospective items were completed prior to the test, whereas retrospective items were completed by the same examinees after the test.

Table 14

Longitudinal Invariance Testing of Retrospective and Prospective Importance and Effort Scores from the Prospective Condition

Model	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	ΔCFI	CFI	RMSEA	SRMR	Correlation residuals > .15	Mean residuals > .15
Configural - error covariances	303.130	146	--	--	--	.949	.059	.051	1	--
Metric	330.354	154	27.224*	8	.006	.943	.061	.061	5	--
Partial Metric – I8 free	319.482	153	10.872	1	.003	.946	.059	.057	2	--
Partial Scalar – I8 free	343.753	160	24.271	7	.006	.940	.061	.058	2	0
LGM	343.753	160	24.271	7	.006	.940	.061	.058	2	0
Conditional LGM	373.750	176	29.997	16	.004	.936	.060	.058	--	--

Note. * $p < .01$. χ^2 = maximum likelihood chi-square; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Local misfit was assessed via correlation and mean residuals > |.15| in magnitude. In the configural model with error variances, 84 parameters (4 factor variances, 6 factor covariances, 16 factor pattern coefficients, 20 error variances, 18 error covariances, and 20 means) were estimated from 230 observations (20 variances, 190 covariances, and 20 means). In addition to the 10 error covariances across time points (e.g., prospective importance item 1 and retrospective importance item 1), 8 error covariances between prospective and retrospective importance items 3 and 4 and prospective and retrospective items 3 and 8 were estimated to account for shared variance after controlling for the importance factors. In the metric invariance model, 8 factor pattern coefficients were constrained to be equal across time points (76 estimated parameters). The partial metric invariance model frees the equality constraint on importance item 8's factor loading (77 estimated parameters). In the partial scalar invariance model with no equality constraint on importance item 8's factor loading and intercept, 7 fewer parameters were estimated after constraining 7 intercepts. As noted in the text, the partial scalar model with importance item eight freely estimated is an equivalent model to the LGM model with the equality constraints.

Table 15

Unstandardized and (Standardized) Factor Pattern Coefficients from Longitudinal Analysis of Prospective and Retrospective SOS Scores from the Prospective Measurement Condition

Item	Prospective		Retrospective	
	Importance	Effort	Importance	Effort
1. Doing well on this test is important to me.	1.000 (.697)			
1. Doing well on this test was important to me.			1.000 (.889)	
3. I am not curious about how I will do on this test relative to others.*	0.579 (.339)			
3. I am not curious about how I did on this test relative to others.*			0.537 (.424)	
4. I am not concerned about the score I will receive on this test.*	0.986 (.564)			
4. I am not concerned about the score I receive on this test.*			0.734 (.585)	
5. This is an important test to me.	1.048 (.710)			
5. This was an important test to me.			0.834 (.729)	
8. I would like to know how well I do on this test.	1.052 (.658)			
8. I would like to know how well I did on this test.			0.687 (.622)	
2. I will engage in good effort throughout this test.		1.000 (.774)		
2. I engaged in good effort throughout this test.				1.000 (.861)
6. I will give my best effort on this test.		1.100 (.872)		
6. I gave my best effort on this test.				1.020 (.881)
7. After taking this test, I expect I could have worked harder on it.*		0.607 (.400)		
7. While taking this test, I could have worked harder on it.*				0.783 (.524)
9. I will not give this test my full attention while completing it.*		0.955 (.721)		
9. I did not give this test my full attention while completing it.*				0.912 (.701)
10. While taking this test, I will persist to completion of the task.		0.831 (.667)		
10. While taking this test I was able to persist to completion of the task.				0.691 (.612)

Note. *Denotes items that are reversed prior to scoring. The prospective item wording is presented first followed by the corresponding retrospective item. All parameter estimates were statistically significant at $p < .01$. A two-factor model with error covariances between importance items 3 and 4 and items 3 and 8 was fit to longitudinal data collected before and after test completion.

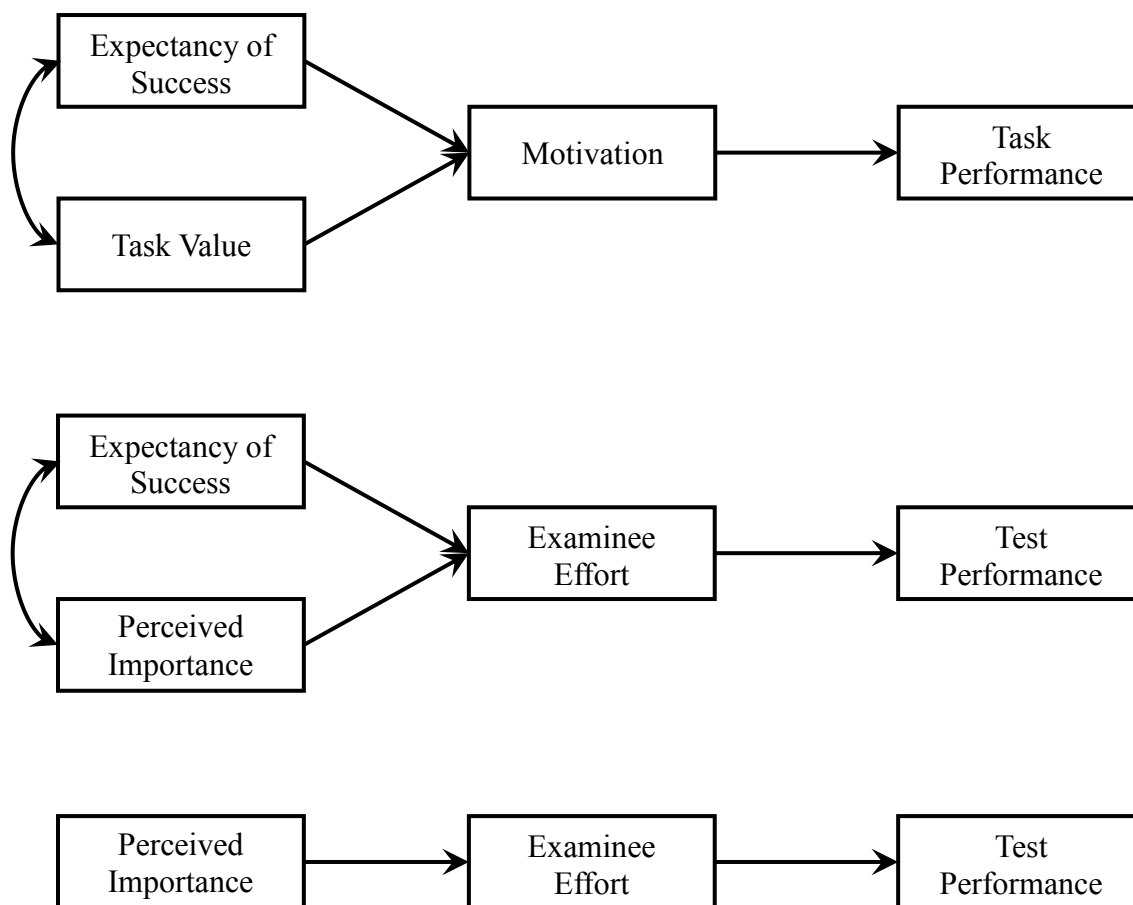


Figure 1. Illustration of EV theory (top), EV theory applied to low-stakes testing (middle) and EV theory as conceptualized in the current study (bottom). In the top model, notice the general conceptualization of EV theory suggests expectancy of success and task value indirectly affect task performance via motivation. EV theory applied to low-stakes testing specifies expectancy of success and perceived importance indirectly affect test performance via examinee effort. Because expectancy of success may not be relevant in low-stakes testing contexts, the bottom model specifies perceived test importance indirectly affects test performance via examinee effort.

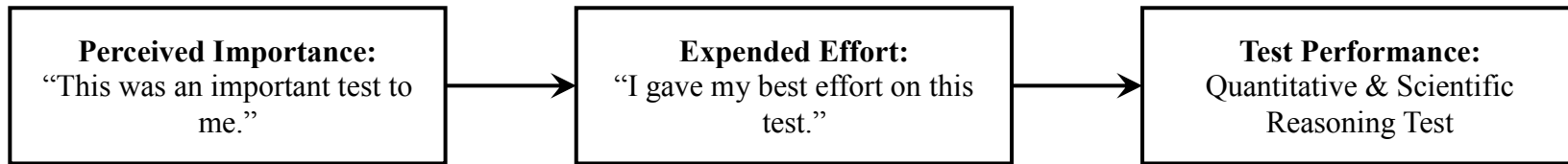
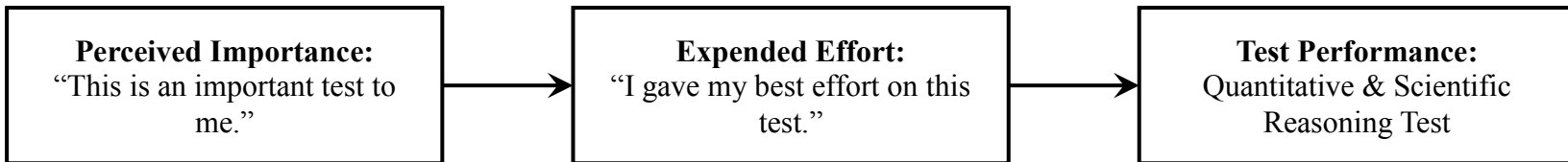
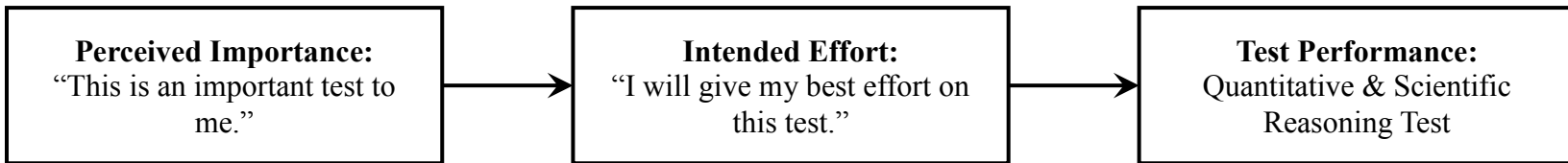
Retrospective Condition**Combined Condition****Prospective Condition**

Figure 2. Illustration of measurement conditions with example items. The model in the retrospective condition (top) illustrates modeling retrospectively measured perceived test importance and expended effort. The model in the combined condition (middle) illustrates modeling prospectively measured perceived test importance and retrospectively measured expended effort. The model in the prospective condition (bottom) illustrates prospective measurement of perceived test importance and intended effort. (i.e., aligning with the theoretically suggested temporal relationship between the variables).

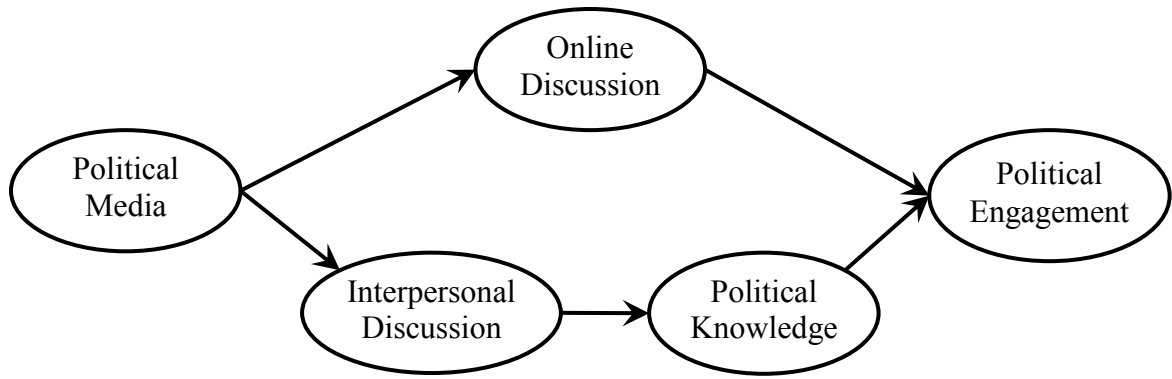


Figure 3. Illustration of the relationship between political media and political engagement. This model suggests political media indirectly affects political engagement (1) via online discussion and (2) via interpersonal discussion and political knowledge.

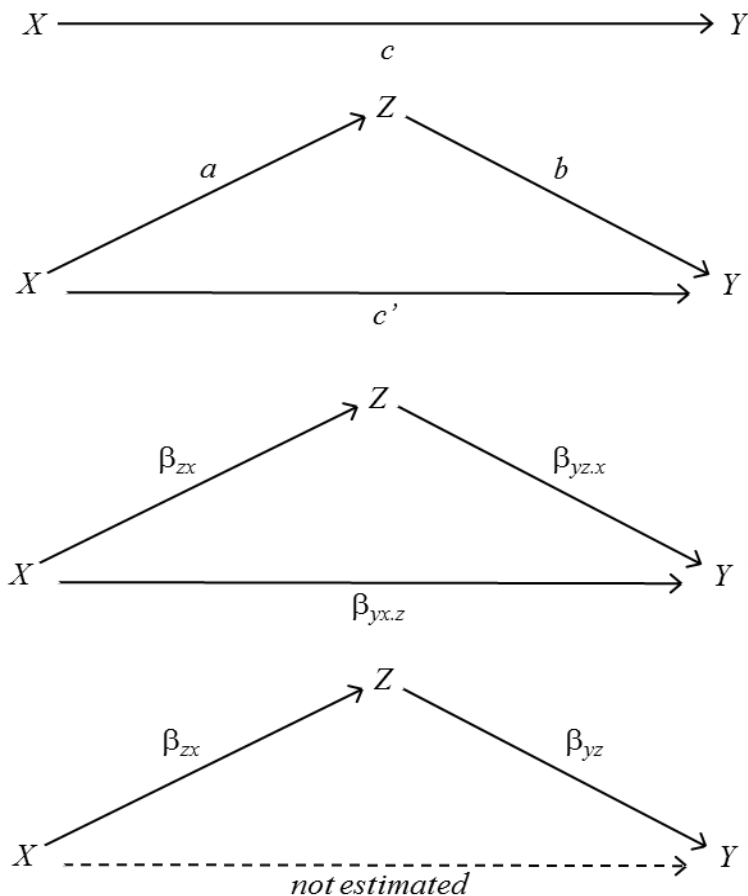
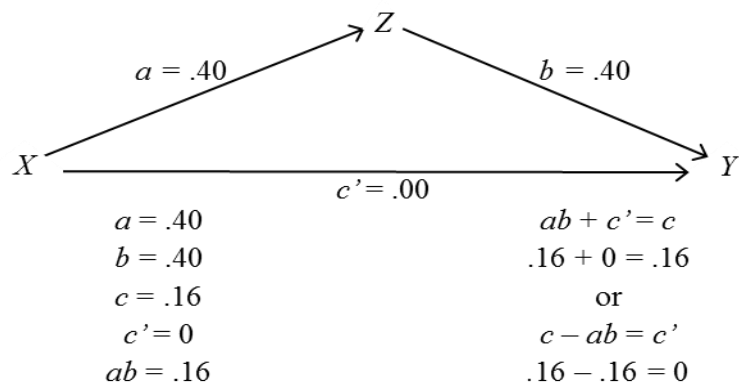


Figure 4. Illustration of parameter symbols and regression formulas for partial and complete mediation. The top two path models illustrate a bivariate relationship (top) with total effect of X on Y (i.e., c) and the second model illustrates the addition of a third variable functioning as a mediator where the X - Z , Z - Y , and X - Y controlling for Z relationships are represented by a , b , and c' , respectively. The bottom two path models illustrate regression formulas representing symbols a , b , and c' from the second path model. The regression formulas used in the third model assume partial mediation where the c' parameter (direct effect; $\beta_{yx.z}$) is estimated and the b parameter is estimated controlling for X ($\beta_{yz.x}$). In contrast, if complete mediation is assumed (e.g., SEM model-fit approach) the estimation of the b parameter does not control for X (β_{yz}) and the c' parameter (direct effect) is not estimated.

Complete Mediation



Partial Mediation

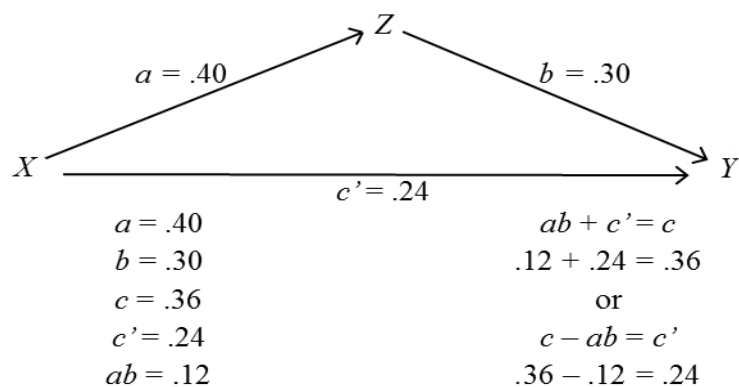


Figure 5. Illustration of complete and partial mediation. The top model illustrates complete mediation where the sum of the indirect effect ($ab = .16$) and the direct effect ($c' = 0$) equal the total effect ($c = .16$). Equivalently, the difference between the total effect ($c = .16$) and the indirect effect ($ab = .16$) equals the direct effect ($c' = .00$). Notice, in complete mediation, the direct effect ($c' = .00$) is not significantly different from zero. Thus, all of the effect of X on Y is through Z . In contrast, the bottom model illustrates partial mediation where the total effect ($c = .36$) minus the indirect effect ($ab = .12$) equals the direct effect ($c' = .24$). Thus, in addition to the indirect effect, a direct path from X to Y is necessary to account for the bivariate relationship between X and Y ($c = .36$).

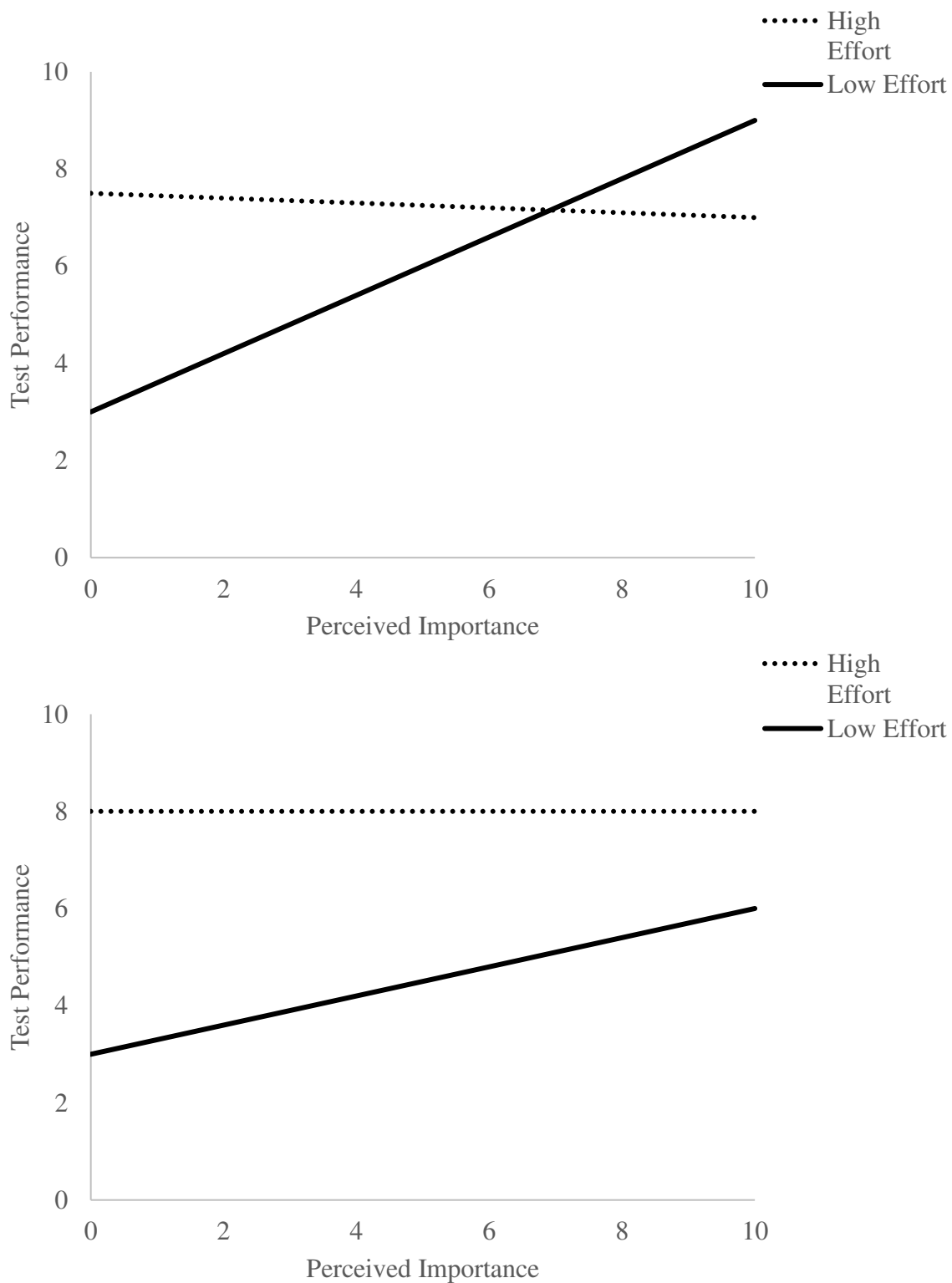
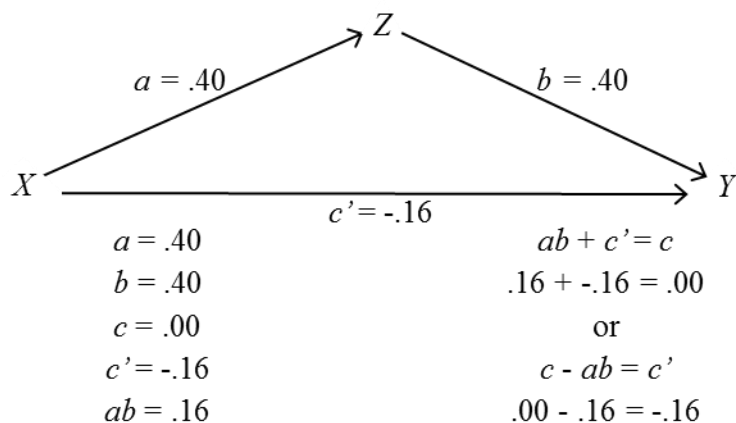


Figure 6. Illustration of ordinal (top) and disordinal (bottom) moderation effects.

Classical Suppression



Negative Suppression

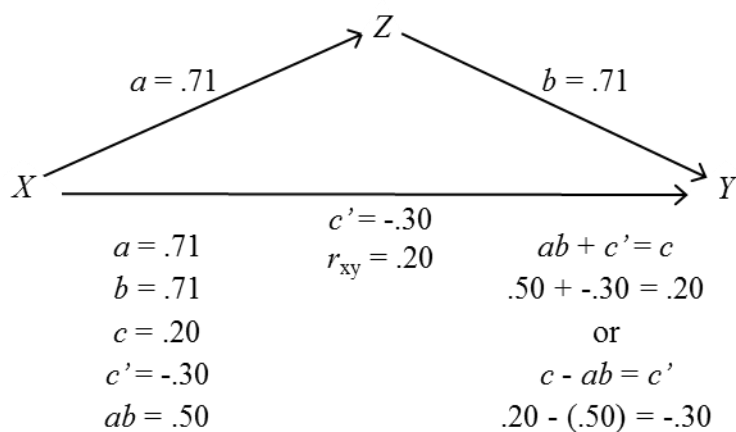
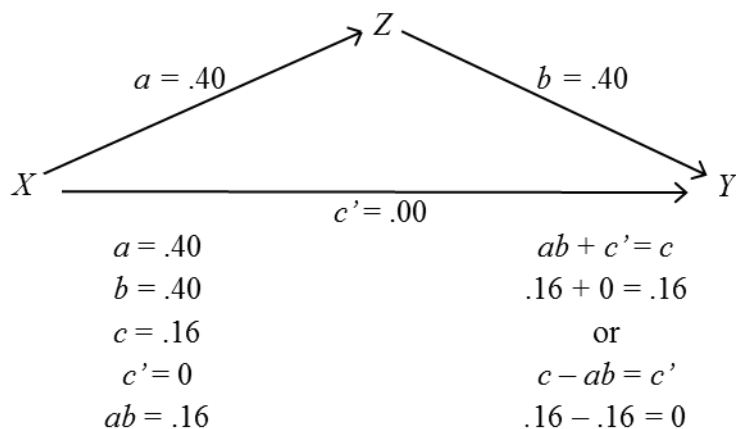


Figure 7. Illustration of classical and negative suppression models. The classical suppression model (top) illustrates how a nil total effect (i.e., $c = .00$) can be decomposed into a nonzero indirect effect ($ab = .16$) and direct effect ($c' = -.16$). The negative suppression model (bottom) illustrates how the observed bivariate relationship ($r_{xy} = .20$) is in the opposite direction as the direct effect beta weight ($c' = .52$).

Mediation



Confounding

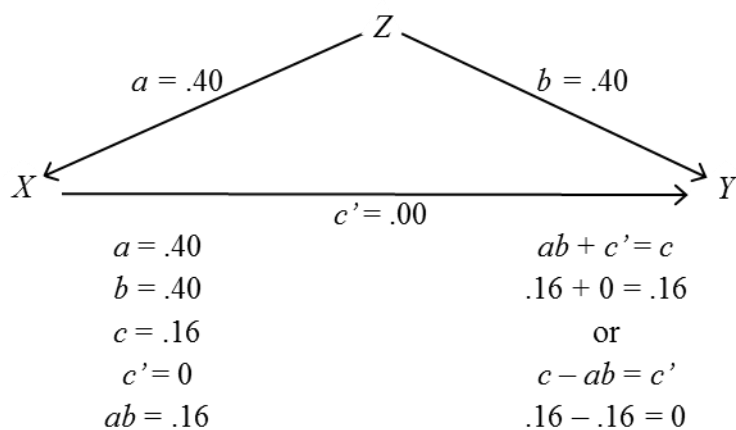
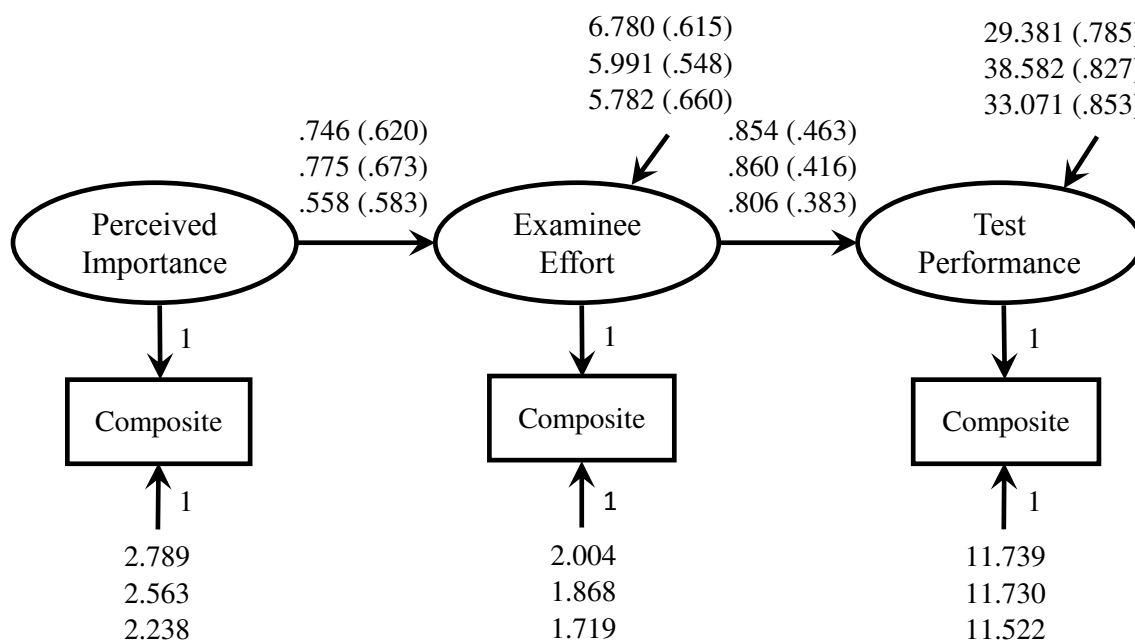


Figure 8. Comparison of complete mediation and complete confounding models. Notice the complete mediation (top) and confounding (bottom) models are equivalent (e.g., identical parameter estimates and nil direct effect), with the exception of the direction of the arrow between X and Z . Despite the identical model-data fit, the top model (complete mediation) suggests X indirectly causes Y via Z , whereas the confounding model suggests that Z is a cause of both X and Y .

Retrospective SOS



Prospective/Retrospective SOS

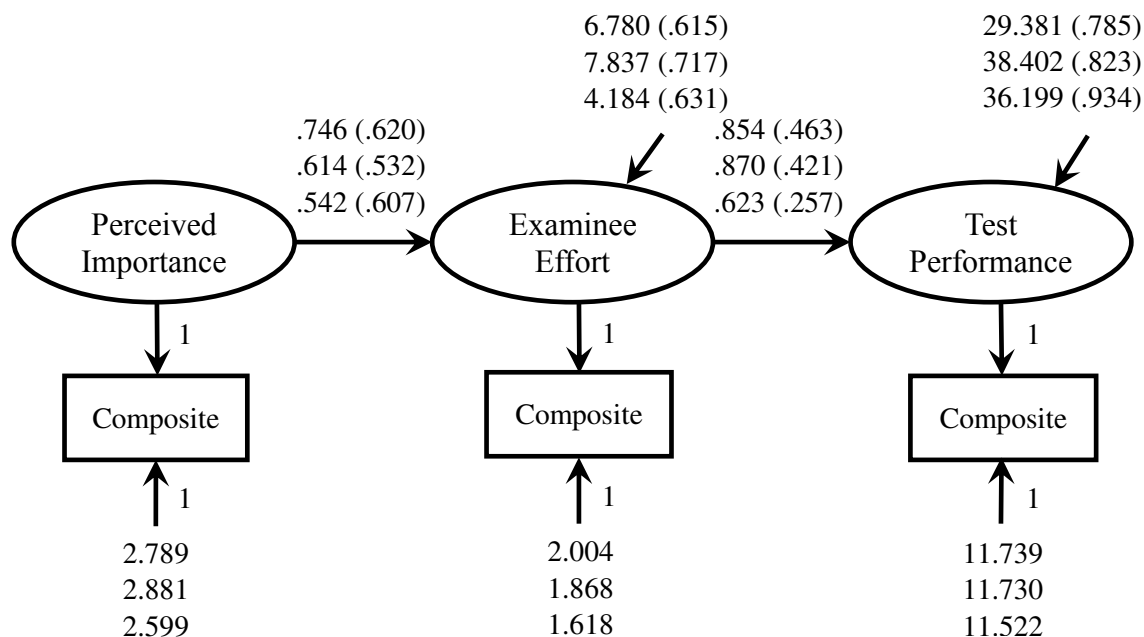


Figure 9. Completed mediated models modeling retrospective SOS scores and combined prospective/retrospective SOS scores. Unstandardized and (standardized) parameter estimates for retrospective (top), combined (middle), and prospective (bottom) conditions. All parameter estimates were statistically significant at $p < .01$. Single-indicator latent variables were modeled by setting the composite variable's error variance

to the proportion of the composite's variance due to measurement error ($[1-r_{xx}] * s_x^2$). Thus, leaving the essentially measurement error free variance to be modeled. The proportion of variance explained in the latent variables can be interpreted as (1 – standardized disturbance term). For example, the proportion of variance explained in test performance in the retrospective condition is .215 (1-.785). Thus, 21.5% of the variance in test performance can be explained by examinee effort in the retrospective condition.

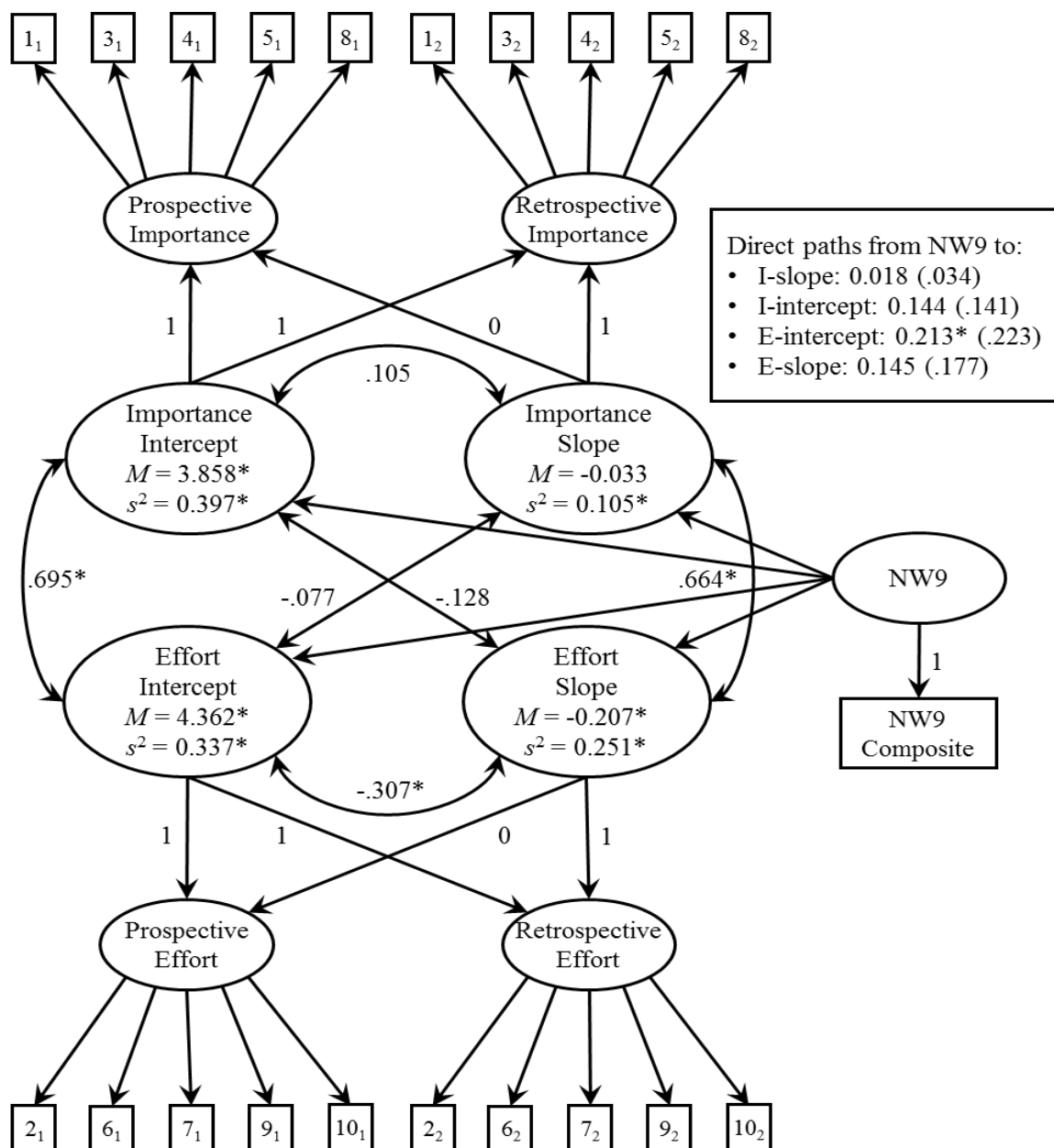


Figure 10. Conditional LGM: Predicting change in importance and effort from test scores. Model-data fit was adequate: $\chi^2 = 373.75$, $p < .01$, CFI = .94, RMSEA = .06, and SRMR = .06. Values in intercept factors represent mean (M) and unexplained variance (s^2) of baseline (prospective) scores. Values in slope factors represent average rate of change (M) and unexplained variance (s^2) of change scores. Values in box = unstandardized and (standardized) direct paths. Curved and double-headed paths = correlations. Unstandardized pattern coefficients from intercept factors to prospective (time 1) and retrospective (time 2) factors were fixed to 1. Unstandardized pattern coefficients from slope factors to prospective

and retrospective factors were fixed to 0 and 1, respectively. Covariances between prospective and retrospective importance and effort factors and disturbance terms were omitted from the figure for clarity. $*p < .01$, indicates value is statistically significantly different from zero. The latent importance ($M = 3.858$) and effort ($M = 4.362$) intercept factors equal the prospective importance and effort latent means for examinees with average test performance, as estimated in the LMACS model. Moreover, latent importance ($M = -0.033$) and effort ($M = -0.207$) slope factors equal the latent change in importance and effort for examinees with average performance, as estimated in the LMACS model. Note, NW9 composite scores were scaled by a factor of 10; thus, the unstandardized direct paths are interpreted as follows. For every 10 unit increase in test performance, examinees tended to increase effort ratings by 0.213 units. The intercept and slope factor correlations can be interpreted as follows. After controlling for performance, the relationship between prospective importance and effort was significant (partial $r = .695$) and the relationship between prospective effort and change in effort was significant (partial $r = -.307$). Given the negative effort slope ($M = -0.207$), examinees who reported higher intended effort before the test tended to decrease their effort ratings over time and examinees who reported lower effort before the test tended to increase their effort ratings over time. Interestingly, after controlling for performance, prospective importance is not related to change in importance (partial $r = .105$). Moreover, the relationships between prospective importance and change in effort (partial $r = -.128$) and prospective effort and change in importance (partial $r = -.077$) were nonsignificant after controlling for performance.

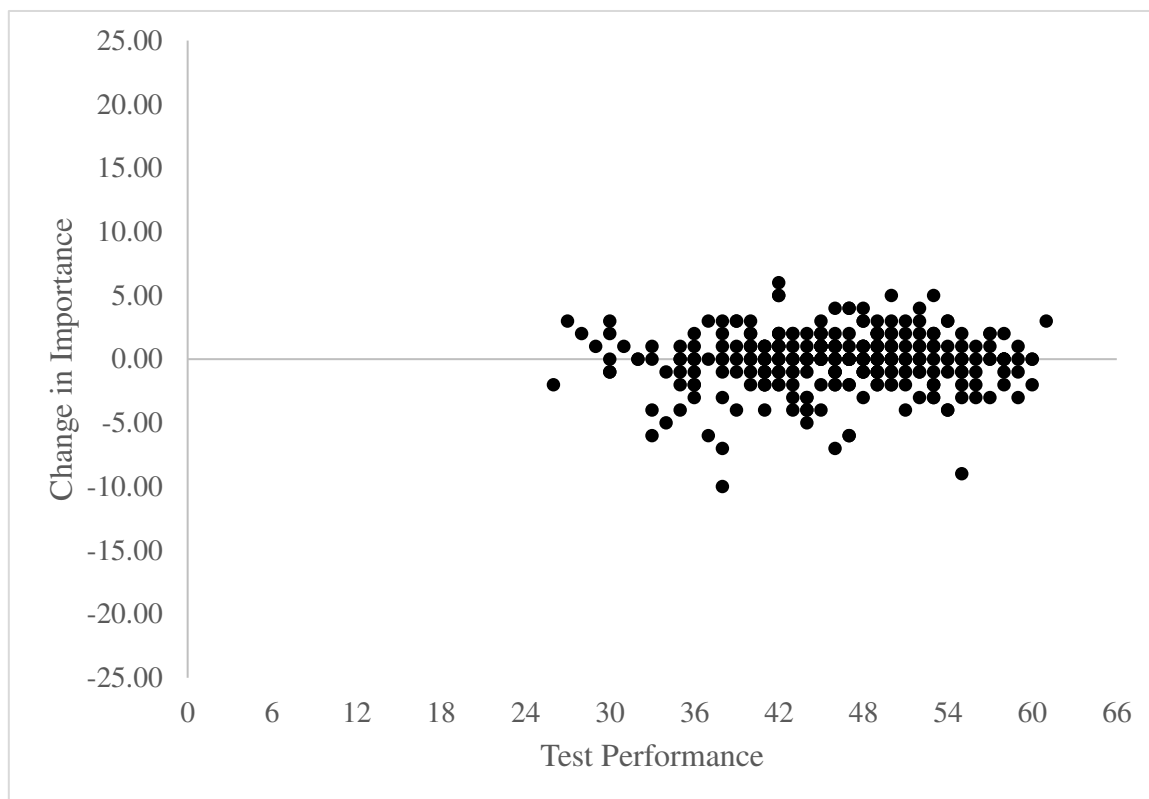


Figure 11. Plot of importance change scores as a function of test performance. Change scores were calculated by subtracting retrospective importance scores from prospective importance scores. Thus, importance change scores greater than zero reflect a decrease in ratings of importance from before to after test completion.

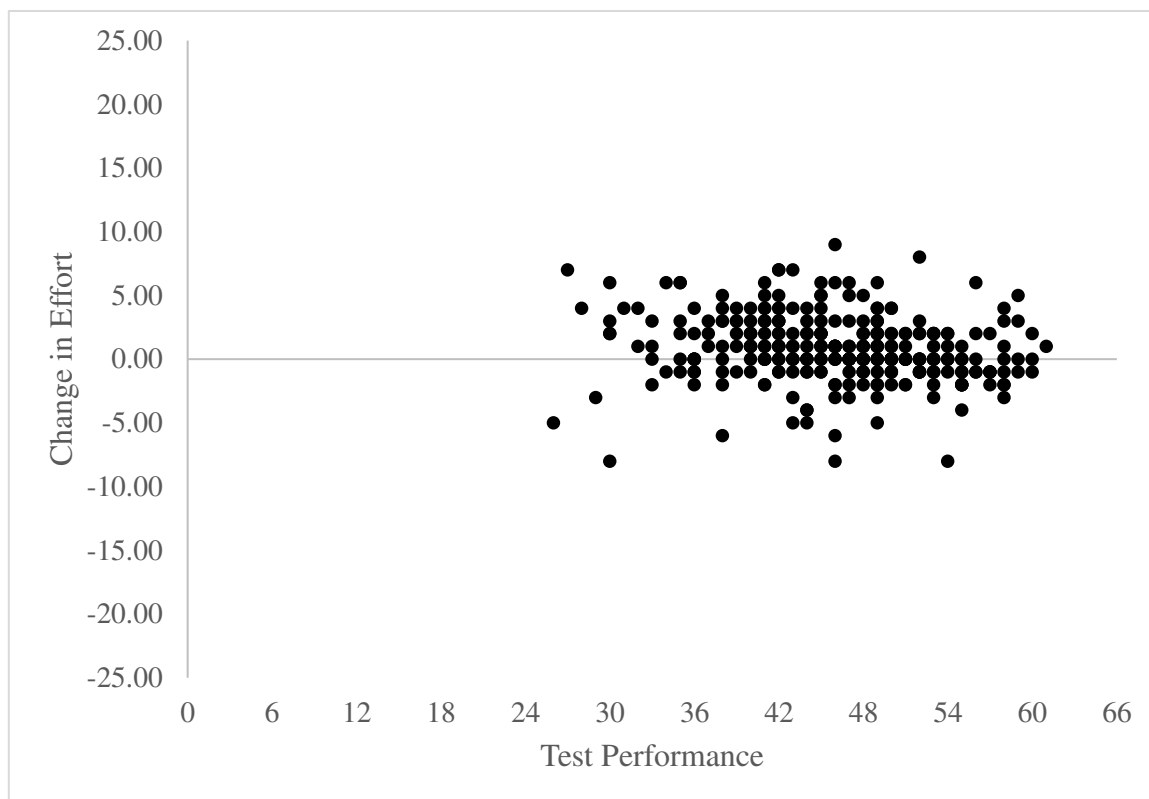


Figure 12. Plot of effort change scores as a function of test performance. Change scores were calculated by subtracting retrospective effort scores from prospective effort scores. Thus, effort change scores greater than zero reflect a decrease in effort ratings from before to after test completion. Notice, the plot suggests that examinees who performed poorer on the test tended to decrease their effort ratings from before to after the test to a greater degree than examinees who performed better on the test.

Appendix A

Natural World-Student Opinion Scale

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

A = Strongly Disagree

B = Disagree

C = Neutral

D = Agree

E = Strongly Agree

1. Doing well on this test was important to me.
2. I engaged in good effort throughout this test.
3. I am not curious about how I did on this test relative to others.
4. I am not concerned about the score I receive on this test.
5. This was an important test to me.
6. I gave my best effort on this test.
7. While taking this test, I could have worked harder on it.
8. I would like to know how well I did on this test.
9. I did not give this test my full attention while completing it.
10. While taking this test, I was able to persist to completion of the task.
11. While taking this test, I thought about how poorly I was doing.
12. While taking this test, I thought about items on other parts of this test I could not answer.
13. While taking this test, I had an uneasy upset feeling.
14. While taking this test, I thought of the consequences of performing poorly.
15. While taking this test, I felt my heart beating fast.
16. While taking this test, I felt so tense that my stomach was upset.
17. While taking this test, I felt that I did not do as well as I could have.
18. While taking this test, I felt nervous.
19. I do not feel very confident about my performance on this test.
20. While taking this test, I felt panicky.

Appendix B

ONWA Opinions of the Natural World Scale

Before we begin, here is an example item from the Natural World Test.

Example Item:

Goldstar Inc. claims that its SAT preparation course is superior to the course offered by Premiere Inc. A study conducted by Goldstar compared SAT scores from 500 students who took Goldstar's course and 500 students who took Premiere's course. Their study concluded that students who took Goldstar's course scored significantly higher than students who took Premiere's course. Is Goldstar justified in its claim that its SAT preparation course is superior to Premiere's course?

- x. The evidence strongly supports this claim.
- y. The evidence contradicts this claim.
- z. The evidence is not sufficient to support or contradict this claim.

Please think about the **quantitative and scientific reasoning test that you are about to complete**. Mark the answer that best represents how you feel about each of the statements below.

A = Strongly Disagree

B = Disagree

C = Neutral

D = Agree

E = Strongly Agree

1. Doing well on this test is important to me.
2. I am not curious about how I will do on this test relative to others.
3. I am not concerned about the score I will receive on this test.
4. This is an important test to me.
5. I would like to know how well I do on this test.

Appendix C

ONWB

Opinions of the Natural World Scale

Before we begin, here is an example item from the Natural World Test.

Example Item:

Goldstar Inc. claims that its SAT preparation course is superior to the course offered by Premiere Inc. A study conducted by Goldstar compared SAT scores from 500 students who took Goldstar's course and 500 students who took Premiere's course. Their study concluded that students who took Goldstar's course scored significantly higher than students who took Premiere's course. Is Goldstar justified in its claim that its SAT preparation course is superior to Premiere's course?

- x. The evidence strongly supports this claim.
- y. The evidence contradicts this claim.
- z. The evidence is not sufficient to support or contradict this claim.

Please think about the **quantitative and scientific reasoning test that you are about to complete**. Mark the answer that best represents how you feel about each of the statements below.

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

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

Appendix D

Assessment Day Fall, 2015 NW9 and Retrospective SOS Test Instructions

[Proctors – students can use scrap paper for this test. Please pass out scrap paper if you have not already done so.]

Please make sure you have correctly filled in your name and JACard number on the scantron form and that **NW-9+SOS** is written in the top right corner of the scantron form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At JMU we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed at JMU with faculty who teach in Cluster Three of JMU's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete this test. Please do not write on the test; answer all questions on the scantron provided. You will have scrap paper to help you; if you need more, just raise your hand. Read all test directions carefully, and answer the items to the best of your ability.

We are pleased to let you know that you will be able to find out, via MyMadison, how you scored on the quantitative and scientific reasoning measures and what your scores can tell you about these reasoning skills. Later in the semester, you will receive an email providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of JMU faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

You will be told when there are 10 and 5 minutes remaining. Take your time. When you have completed all of the items, please wait quietly until the other students have finished.

Thank you in advance for your effort and concentration on this important test.

You may begin.

[Remind students when they have 10 and 5 minutes remaining. After the students have completed the NW9, collect the test forms, scantrons, and scrap paper and put them in the designated envelopes within your bin.]

Appendix E

Assessment Day Fall, 2016 Prospective Importance Instructions

[Proctors – students can use scrap paper for this test. Please pass out scrap paper if you have not already done so.]

You will use two scantrons for this portion of the testing session. Please make sure you have correctly filled in your name and JACard number on both scantron forms and that **ONWA** is written in the top right corner of one scantron form and **NW9SOS** is written in the top right corner of the other scantron form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At JMU we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed at JMU with faculty who teach in Cluster Three of JMU's award winning General Education program. The results are used to inform and improve our General Education program.

We are pleased to let you know that you will be able to find out, via MyMadison, how you scored on the quantitative and scientific reasoning measures and what your scores can tell you about these reasoning skills. Later in the semester, you will receive an email providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of JMU faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

The Natural World test is a 60-minute, 66-item multiple-choice test designed to assess your knowledge of natural sciences, specifically your quantitative and scientific reasoning skills. These skills include use of mathematical analyses, formulation and evaluation of scientific hypotheses, knowledge of basic and applied research, interpretation of graphical data, and evaluation of the credibility of scientific information.

Please take a moment to review an example item from this test displayed on the screen at the front of the room and on the top of the blue **Opinions of the Natural World Scale** in front of you:

[Please allow students 30 seconds to read the following item then read the question aloud.]

Goldstar Inc. claims that its SAT preparation course is superior to the course offered by Premiere Inc. A study conducted by Goldstar compared SAT scores from 500 students who took Goldstar's course and 500 students who took Premiere's course.

Their study concluded that students who took Goldstar’s course scored significantly higher than students who took Premiere’s course. Is Goldstar justified in its claim that its SAT preparation course is superior to Premiere’s course?

[x. The evidence strongly supports this claim.]

[y. The evidence contradicts this claim.]

[z. The evidence is not sufficient to support or contradict this claim.]

This example item is representative of the types of items you will encounter on the Natural World test. In case you are curious, the answer to the example item is z, but again, this is just a sample of the items you are about to encounter.

Before you take the quantitative and scientific reasoning test, we ask that you complete the blue **Opinions of the Natural World Scale**. It is a 5-item assessment used to measure your opinions with respect to the quantitative and scientific reasoning test you are about to complete. Please do not write on the blue sheet; answer all items as honestly as possible on the scantron labeled **ONWA**. You will have 2 minutes to complete this scale.

When you have completed all of the items, please wait quietly until the other students have finished. Thank you in advance for your effort and concentration on this important test.

You may begin.

[After ONWA time has expired, ask them to pass their ONWA test forms and completed scantrons to the aisle to collect. After assuring the students completed the correct scantron, please place them in the designated envelopes within your bin.]

Appendix F

Assessment Day Fall, 2016 NW9 and Retrospective SOS Test Instructions

[Proctors – students can use scrap paper for this test. Please pass out scrap paper if you have not already done so.]

Please make sure you have correctly filled in your name and JACard number on the scantron form and that **NW9SOS** is written in the top right corner of the scantron form.

You will now have 60 minutes to complete the Natural World test. Please do not write on the test; answer all questions on the scantron labeled **NW9SOS**. You will have scrap paper to help you; if you need more, just raise your hand. Read all test directions carefully, and answer the items to the best of your ability.

You will be told when there are 10 and 5 minutes remaining. Take your time. When you have completed all of the items, please wait quietly until the other students have finished.

Thank you in advance for your effort and concentration on this important test.

You may begin.

[Remind students when they have 10 and 5 minutes remaining. After NW9 time has expired, collect the test forms, scantrons, and scrap paper. After assuring the students completed the correct scantron, please place them in the designated envelopes within your bin.]

Appendix G

Assessment Day Fall, 2016 Prospective Importance and Effort Instructions

[Proctors – students can use scrap paper for this test. Please pass out scrap paper if you have not already done so.]

You will use two scantrons for this portion of the testing session. Please make sure you have correctly filled in your name and JACard number on both scantron forms and that **ONWB** is written in the top right corner of one scantron form and **NW9SOS** is written in the top right corner of the other scantron form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At JMU we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed at JMU with faculty who teach in Cluster Three of JMU's award winning General Education program. The results are used to inform and improve our General Education program.

We are pleased to let you know that you will be able to find out, via MyMadison, how you scored on the quantitative and scientific reasoning measures and what your scores can tell you about these reasoning skills. Later in the semester, you will receive an email providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of JMU faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

The Natural World test is a 60-minute, 66-item multiple-choice test designed to assess your knowledge of natural sciences, specifically your quantitative and scientific reasoning skills. These skills include use of mathematical analyses, formulation and evaluation of scientific hypotheses, knowledge of basic and applied research, interpretation of graphical data, and evaluation of the credibility of scientific information.

Please take a moment to review an example item from this test displayed on the screen at the front of the room and on the top of the peach **Opinions of the Natural World Scale** in front of you:

[Please allow students 30 seconds to read the following item then read the question aloud.]

Goldstar Inc. claims that its SAT preparation course is superior to the course offered by Premiere Inc. A study conducted by Goldstar compared SAT scores from 500 students who took Goldstar's course and 500 students who took Premiere's course. Their study concluded that students who took Goldstar's course scored significantly

higher than students who took Premiere's course. Is Goldstar justified in its claim that its SAT preparation course is superior to Premiere's course?

[x. The evidence strongly supports this claim.]

[y. The evidence contradicts this claim.]

[z. The evidence is not sufficient to support or contradict this claim.]

This example item is representative of the types of items you will encounter on the Natural World test. In case you are curious, the answer to the example item is z, but again, this is just a sample of the items you are about to encounter.

Before you take the quantitative and scientific reasoning test, we ask that you complete the peach **Opinions of the Natural World Scale**. It is a 10-item assessment used to measure your opinions with respect to the quantitative and scientific reasoning test you are about to complete. Please do not write on the peach sheet; answer all items as honestly as possible on the scantron labeled **ONWB**. You will have 2 minutes to complete this scale.

When you have completed all of the items, please wait quietly until the other students have finished. Thank you in advance for your effort and concentration on this important test.

You may begin.

[After ONWB time has expired, ask them to pass their ONWB test forms and completed scantrons to the aisle to collect. After assuring the students completed the correct scantron, please place them in the designated envelopes within your bin.]

References

- Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied, 13*, 57-58.
- AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. *Evaluation Review, 14*, 374-390. DOI: 10.1177/0193841X9001400403
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.
- Arkin, R. M., Appelman, A. J., & Burger, J. M. (1980). Social anxiety, self-presentation, and the self-serving bias in causal attribution. *Journal of Personality and Social Psychology, 38*, 23-35. DOI: 10.1037/0022-3514.38.1.23
- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*, 359-372. DOI: 10.1037/h0043445
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. DOI: 10.1037/0022-3514.51.6.1173
- Barron, K. E., & Hulleman, C. S. (2015). Expectancy-Value-Cost model of motivation. In J.D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences, 2nd edition* (Vol. 8, pp. 503-509). Oxford: Elsevier. DOI: 10.1016/B978-0-08-097086-8.26099-6

- Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*, 46-64. DOI: 10.1080/08957347.2015.1102914
- Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441-462.
- Bensley, D. A., Rainey, C., Murtagh, M. P., Flinn, J. A., Maschiocchi, C., Bernhardt, P. C., & Kuehne, S. (2016). Closing the assessment loop on critical thinking: The challenges of multidimensional testing and low test-taking motivation. *Thinking Skills and Creativity, 21*, 158-168.
- Bentler, P. M. (1998, March 10). Kurtosis, residuals, fit indices. Message posted to <http://listserv.ua.edu>
- Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research, 45*, 661-701. DOI: 10.1080/00273171.2010.498292
- Boekaerts, M. (2002). The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. *Advances in Motivation and Achievement: A Research Annual, 12*, 77-120.
- Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115-140. DOI: 10.2307/271084

- Carver, C. S., & Scheier, M. F. (2000). On the structure of behavioral self-regulation. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 41-84). London, UK: Elsevier Academic Press.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance, *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 464-504. DOI: 10.1080/10705510701301834
- Cheng, P. Y., & Chiou, W. B. (2010). Achievement, attributions, self-efficacy, and goal setting by accounting undergraduates. *Psychological Reports*, 106, 54-64.
- Cheong, J., MacKinnon, D. P., & Khoo, S. T. (2003). Investigation of mediational processes using parallel process latent growth curve modeling. *Structural Equation Modeling*, 10, 238-262. DOI: 10.1207/S15328007SEM1002_5
- Cheung, G. W., & Lau, R. S. (2008). Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models. *Organizational Research Methods*, 11, 296-325. DOI: 10.1177/1094428107300343
- Cheung, G. W., & Lau, R. S. (2015). Accuracy of parameter estimates and confidence intervals in moderated mediation models: A comparison of regression and latent moderated structural equations. *Organizational Research Methods*, 1, 1-24. DOI: 10.1177/1094428115595869
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- Cheung, M. W. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling*, 14, 227-246. DOI: 10.1080/10705510709336745

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 4*, 558-577. DOI: 10.1037/0021-843X.112.4.558
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*, 300-315.
- Cole, J. S. (2007). *Motivation to do well on low-stakes tests* (Doctoral dissertation). Retrieved from University of Missouri-Columbia Doctoral Dissertation Archives: <http://hdl.handle.net/10355/4676>
- Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education, 57*, 119-130.
- Cole, J. S., Bergin, D. A., & Whittaker T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624. DOI: 10.1016/j.cedpsych.2007.10.002
- Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34*, 35-46.
- Cook, T., & Campbell, D. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton Mifflin.

- Crespo, C., Carona, C., Silva, N., Canavarro, M. C., & Dattilio, F. (2011). Understanding the quality of life for parents and their children who have asthma: Family resources and challenges. *Contemporary Family Therapy, 33*, 179-196. DOI: 10.1007/s10591-011-9155-5
- DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment, 8*, 69-82.
- DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412-423.
- Dong, Y., Stupnisky, R. H., & Berry, J. C. (2013). Multiple causal attributions: An investigation of college students learning a foreign language. *European Journal of Psychology of Education, 28*, 1587-1602.
- Eccles, J. S., Adler, T. E., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75-146). San Francisco, CA: W. H. Freeman.
- Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311-326. DOI: 10.1080/15305050701438074
- Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education, 27*, 31-45.

- Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science, 10*, 87-99. DOI: 10.1007/s11121-008-0109-6
- Falk, C. F., & Biesanz, J. C. (2015). Inference and interval estimation methods for indirect effects with latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 24-38. DOI: 10.1080/10705511.2014.935266
- Falk, C. F., & Biesanz, J. C. (2016). Two cross-platform programs for inferences and interval estimation about indirect effects in mediational models. *SAGE Open, 6*, 1-13. DOI: 10.1177/2158244015625445
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In Hancock, G. R. & Mueller, R. O. (Eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching*. (pp. 439-492), Charlotte, NC, US: Information Age Publishing.
- Finney, S. J., DiStefano, C., & Kopp, J. P. (2016). Overview of estimation methods and preconditions for their application with structural equation modeling. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods for test construction: Standards and recent advancements*. (pp. 135-165). Boston, MA: Hogrefe.
- Finney, S. J., Mathers, C. E., & Myers, A. J. (2016). Investigating the dimensionality of test-taking motivation across test instruction conditions in low-stakes testing contexts. *Research & Practice in Assessment, 11*, 5-17.
- Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 64*, 365-383.

- Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low- stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment, 21*, 60-87. DOI: 10.1080/10627197.2015.1127753
- Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology, 41*, 232-244.
- Folkman, S., & Lazarus, R. S. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology, 48*, 150-170. DOI: 10.1037/0022-3514.48.1.150
- Folkman, S., & Lazarus, R. S. (1988). Coping as a mediator of emotion. *Journal of Personality and Social Psychology, 54*, 466-475. DOI: 10.1037/0022-3514.54.3.466
- Frazier, P. A., Tix, A. P., & Baron, K. E. (2004). Testing moderator and mediator effects in counselling psychology. *Journal of Counseling Psychology, 51*, 115-134. DOI: 10.1037/0022-0167.51.1.115
- French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*, 378-402.
- Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence, 39*, 233-243.

- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*, 233-239. DOI: 10.1111/j.1467-9280.2007.01882.x
- Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research, 47*, 61-87. DOI: 10.1080/00273171.2012.640596
- Garland, H. (1985). A cognitive mediation theory of task goals and human performance. *Motivation and Emotion, 9*, 345-367.
- Goetz, T., Preckel, F., Pekrun, R., & Hall, N. C. (2007). Emotional experiences during test taking: Does cognitive ability make a difference? *Learning and Individual Differences, 17*, 3-16.
- Gollwitzer, P. M. (1993). Goal achievement: The role of intentions. *European Review of Social Psychology, 4*, 141-185.
- Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist, 54*, 493-503.
- Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69-119.
- Greenberg, J., Pyszczynski, T., & Solomon, S. (1982). The self-serving attributional bias: Beyond self-presentation. *Journal of Experimental Social Psychology, 18*, 56-67.
- Gresky, D. M., Eyck, L. L. T., Lord, C. G., & McIntyre, R. B. (2005). Effects of salient multiple identities on women's performance under mathematics stereotype threat. *Sex Roles, 53*, 703-716. DOI: 10.1007/s11199-005-7735-2

- Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition, 12*, 695-707. DOI: 10.1016/S1053-8100(03)00052-7
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17-27.
- Hathcoat, J. D., Sundre, D. L., & Johnston, M. M. (2015). Assessing college students' quantitative and scientific reasoning: The James Madison University story. *Numeracy, 8*(1), Article 2. DOI: 10.5038/1936-4660.8.1.2
- Hawthorne, K. A., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test. *Research & Practice in Assessment, 10*, 30-39.
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*, 408-420. DOI: 10.1080/03637750903310360
- Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science, 24*, 1918-1927. DOI: 10.1177/0956797613480187
- Hershberger, S. L. (2006). The problem of equivalent structural models. In G. R. Hancock & R. O. Muller (Eds.), *Structural equation modeling: A second course* (pp. 13-41). Greenwich, CT: Information Age Publishing.
- Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*, 1280-1300. DOI: 10.1037/0003-066X.52.12.1280

- Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26*, 501-517.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology, 18*, 449-484. DOI: 10.2307/271055
- Hong, E., & Peng, Y. (2008). Do Chinese students' perceptions of test value affect test performance? Mediating role of motivational and metacognitive regulation in test preparation. *Learning and Instruction, 18*, 499-512. DOI: 10.1016/j.learninstruc.2007.10.002
- Howard, G. S., Dailey, P. R., & Gulanick, N. A. (1979). The feasibility of informed pretests in attenuating response-shift bias. *Applied Psychological Measurement, 3*, 481-494.
- Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research*, (pp. 195-222). Thousand Oaks, California: Sage Publications.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.
- Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology, 17*, 140-154.
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*, 309-334. DOI: 10.1037/a0020761

- James, L. R., & Brett, J. M. (1984). Mediators, moderators, and test for mediation. *Journal of Applied Psychology, 69*, 307-321. DOI: 10.1037/0021-9010.69.2.307
- James, L. R., Mulaik, S. A., & Brett, J. (1982). Causal analysis: Models, assumptions, and data. *Beverly Hills: Sage*.
- James, L. R., Mulaik, S. A., & Brett, J. M. (2006). A tale of two methods. *Organizational Research Methods, 9*, 233-244. DOI: 10.1177/1094428105285144
- Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*, 116-119. DOI: 10.1111/j.1533-8525.1973.tb00868.x
- Johnson, M. P. (1973). Commitment: A conceptual structure and empirical application. *The Sociological Quarterly, 14*, 395-406.
- Jöreskog, K. G., & Sörbom, D. (2015). *LISREL 9.20 for Windows* [Computer Software]. Lincolnwood, IL: Scientific Software International.
- Judd, C. M., & Kenny, D. A. (1981a). *Estimating the effects of social interventions*. Cambridge; New York: Cambridge University Press.
- Judd, C. M., & Kenny, D. A. (1981b). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*, 602-619. DOI: 10.1177/0193841X8100500502
- Jung, N., Kim, Y., & de Zúñiga, H. G. (2011). The mediating role of knowledge and efficacy in the effects of communication on political participation. *Mass Communication and Society, 14*, 407-430. DOI: 10.1080/15205436.2010.496135
- Kenny, D. A. (2016, September 24). Mediation. Retrieved from <http://davidakenny.net/cm/mediate.htm#IE>

- Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science, 25*, 334-339. DOI: 10.1177/0956797613502676
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (4th ed.) (Vol. 1) (pp. 233-265). New York: McGraw-Hill.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling*. New York: Guilford Press.
- Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014, April). *A strategy for increasing student motivation on low-stakes assessments*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology, 27*, S101-S108. DOI: 10.1037/0278-6133.27.2(Suppl.).S101
- Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry, 158*, 848-856. DOI: 10.1176/appi.ajp.158.6.848
- Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59*, 877-883. DOI: 10.1001/archpsyc.59.10.877

- Lee, W., Lee, M. J., & Bong, M. (2014). Testing interest and self-efficacy as predictors of academic self-regulation and achievement. *Contemporary Educational Psychology, 39*, 86-99.
- Little, T. D., Card, N. A., Preacher, K. J., & McConnell, E. (2009). Modeling longitudinal data from research on adolescence. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (3rd ed.) (pp. 15-54). Hoboken, NJ: Wiley.
- Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 4*, 352-362.
- Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*, 79-94. DOI: 10.1080/10627197.2015.1028618
- Locke, E. A., & Latham, G. P. (1990). Work motivation and satisfaction: Light at the end of the tunnel. *Psychological Science, 1*, 240-246.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*, 705-717.
- Lockwood, C. M., & MacKinnon, D. P. (1998, March). Bootstrapping the standard error of the mediated effect. In *Proceedings of the 23rd annual meeting of SAS Users Group International* (pp. 997-1002). Cary, NC: SAS Institute.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185-199. DOI: 10.1037/0033-2909.114.1.185

- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*, 95-107. DOI: 10.1037/h0056029
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Lawrence Erlbaum.
- MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*, 144-158. DOI: 10.1177/0193841X9301700202
- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the m to y relation in mediation analysis. *Personality and Social Psychology Review*, *19*, 30-43. DOI: 10.1177/1088868314542878
- MacKinnon, D. P., Coxé, S., & Baraldi, A. N. (2012). Guidelines for the investigation of mediating variables in business research. *Journal of Business and Psychology*, *27*, 1-14. DOI: 10.1007/s10869-011-9248-z
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593-614. DOI: 10.1146/annurev.psych.58.110405.085542
- MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*, 384-389. DOI: 10.3758/BF03193007
- MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., & Wang, E. Y. I. (1991). Mediating mechanisms in a school-based drug prevention program: first-year effects of the Midwestern Prevention Project. *Health Psychology*, *10*, 164-172. DOI: 10.1037/0278-6133.10.3.164

- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of mediation, confounding, and suppression effect. *Prevention Sciences, 1*, 173-181. DOI: 10.1023/A:1026595011371
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*, 99-128. DOI: 10.1207/s15327906mbr3901_4
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104. DOI: 10.1037/1082-989X.7.1.83
- MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30*, 41-62. DOI: 10.1207/s15327906mbr3001_3
- Manstead, A. S., & Eekelen, S. A. (1998). Distinguishing between perceived behavioral control and self-efficacy in the domain of academic achievement intentions and behaviors. *Journal of Applied Social Psychology, 28*, 1375-1392.
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*, 320-341. DOI: 10.1207/s15328007sem1103_2
- Marsh, H. W., Wen, Z., Hau, K. T., & Nagengast. (2013). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.),

Structural equation modeling: A second course (2nd ed.) (pp. 267-308). Charlotte, NC: Information Age.

Mathers, C. E., Finney, S. J., & Myers, A. J. (2016, July). "While taking this test I felt my heart beating fast:" *The unintended effect of manipulating test instructions to increase examinee motivation in low-stakes testing*. Poster presented at the annual International Meeting of the Psychometric Society, Asheville, NC.

Matthews, K. A., Gallo, L. C., & Taylor, S. E. (2010). Are psychosocial factors mediators of socioeconomic status and health connections? *Annals of the New York Academy of Sciences*, 1186, 146-173. DOI: 10.1111/j.1749-6632.2009.05332.x

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12, 23-44. DOI: 10.1037/1082-989X.12.1.23

McDonald, R. P. (1997). Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research*, 32, 1-38. DOI: 10.1207/s15327906mbr3201_1

McFatter, R. M. (1979). The use of structural equation models in interpreting regression equations including suppressor and enhancer variables. *Applied Psychological Measurement*, 3, 123-135.

McLeod, D. M., Kosicki, G. M., & McLeod, J. M. (2002). Resurveying the boundaries of political communication effects. In J. Bryant & D. Zillmann (Eds.), *Media effects: Advances in theory and research* (2nd ed., pp. 215-267). Mahwah, NJ: Lawrence Erlbaum Associates.

Meeker, W. Q., Jr., Cornwell, L. W., & Aroian, L. A. (1981). *Selected tables in mathematical statistics, volume VII: The product of two normally distributed random variables*. Providence, RI: American Mathematical Society.

- Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin, 130*, 711-747.
- Millsap, R. E. & Meredith, JW. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100: Historical developments and future directions* (pp. 131- 152). Mahwah, NJ: Lawrence Erlbaum.
- Morgan, C. L. (1932). C. Lloyd Morgan. In C. Murchison (Ed.), *A history of psychology in autobiography* (Vol. 2, pp. 237-264). Clark University Press, Worcester, MA.
- Myers, A. J., Finney, S. J., & Mathers, C. E. (2016, July). *A moderated mediation model of test importance, examinee effort, and test performance across test instruction conditions*. Paper presented at the biannual meeting of the International Test Commission, Vancouver, Canada.
- O'Neil, Jr. H. F., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*, 135-157.
- O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208. DOI: 10.1207/s15326977ea1003_3
- Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology, 26*, 291-310.

- Owens, M., Stevenson, J., Norgate, R., & Hadwin, J. A. (2008). Processing efficiency theory in children: Working memory as a mediator between trait anxiety and academic performance. *Anxiety, Stress, & Coping, 21*, 417-430. DOI: 10.1080/10615800701847823
- Pandey, S., & Elliott, W. (2010). Suppressor variables in social work research: Ways to identify in multiple regression models. *Journal of the Society for Social Work and Research, 1*, 28-40.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers
- Pek, J., & Hoyle, R. H. (2016). On the (in) validity of tests of simple mediation: Threats and solutions. *Social and Personality Psychology Compass, 10*, 150-163. DOI: 10.1111/spc3.12237
- Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*, 36-48.
- Pekrun, R., Goetz, T., Perry, R. P., Kramer, K., Hochstadt, M., & Molfenter, S. (2004). Beyond test anxiety: Development and validation of the Test Emotions Questionnaire (TEQ). *Anxiety, Stress and Coping, 17*, 287-316.
- Penk, C., & Richter, D. (2016). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability, 29*, 55-79. DOI: 10.1007/s11092-016-9248-7

- Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27-35. DOI: 10.1016/j.lindif.2015.08.002
- Perry, R. P., Stupnisky, R. H., Daniels, L. M., & Haynes, T. L. (2008). Attributional (explanatory) thinking about failure in new achievement settings. *European Journal of Psychology of Education, 23*, 459-475.
- Petty, R. E., Rennie, G. A., & Cacioppo, J. T. (1987). Assertion versus interrogation format in opinion surveys: Questions enhance thoughtful responding. *Public Opinion Quarterly, 51*, 481-494. DOI: 10.1086/269053
- Pituch, K. A., & Stapleton, L. M. (2011). Hierarchical linear and structural equation modeling approaches to mediation analysis in randomized field experiments, multilevel analysis and tests of mediation. In P. Vogt & M. Williams (Eds.), *Handbook on methodological innovations in the social sciences* (pp. 590-619). London, UK: Sage.
- Plante, I., O'Keefe, P. A., & Théorêt, M. (2013). The relation between achievement goal and expectancy-value theories in predicting achievement-related outcomes: A test of four theoretical conceptions. *Motivation and Emotion, 37*, 65-78. DOI: 10.1007/s11031-012-9282-9
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569.

- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66*, 825-852. DOI: 10.1146/annurev-psych-010814-015258
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers, 36*, 717-731. DOI: 10.3758/BF03206553
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*, 77-98. DOI: 10.1080/19312458.2012.679848
- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*, 185-227.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*, 209-233. DOI: 10.1037/a0020141
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. [Computer software]. Vienna, Austria. Available from <https://www.R-project.org/>.
- Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling, 8*, 142-149. DOI: 10.1207/S15328007SEM0801_8
- Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research, 34*, 199-244. DOI: 10.1207/S15327906Mb340204

- Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Vol. 1. Equivalence in measurement* (pp. 21-50). Greenwich, CT: Information Age.
- Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying unmotivated examinees on Student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 161*, 69-82. DOI: 10.1002/ir.20068
- Rose, B. M., Holmbeck, G. N., Coakley, R. M., & Franks, E. A. (2004). Mediator and moderator effects in developmental and behavioral pediatric research. *Journal of Developmental and Behavioral Pediatrics, 25*, 58-67.
- Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General, 132*, 566-594.
- Schunk, D. H. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment* (pp. 281-303). Springer US.
- Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River, NJ: Pearson Education.
- Schwartz, S. H., & Tessler, R. C. (1972). A test of a model for reducing measured attitude-behavior discrepancies. *Journal of Personality and Social Psychology, 24*, 225-236. DOI: 10.1037/h0033365
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.

- Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*, 356-388. DOI: 10.1080/15305058.2015.1034866
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Sheeran, P., Orbell, S., & Trafimow, D. (1999). Does the temporal stability of behavioral intentions moderate intention-behavior and past behavior-future behavior relations? *Personality and Social Psychology Bulletin, 25*, 724-734.
- Shrout, P., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*, 422-445. DOI: 10.1037/1082-989X.7.4.422
- Smith, E. R. (1982). Beliefs, attributions, and evaluations: Nonhierarchical models of mediation in social cognition. *Journal of Personality and Social Psychology, 43*, 248-259.
- Smith, T. W., Snyder, C. R., & Handelsman, M. M. (1982). On the self-serving function of an academic wooden leg: Test anxiety as a self-handicapping strategy. *Journal of Personality and Social Psychology, 42*, 314-321. DOI: 10.1037/0022-3514.42.2.314
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290-312. DOI: 10.2307/270723
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining

psychological processes. *Journal of Personality and Social Psychology*, 89, 845-851. DOI: 10.1037/0022-3514.89.6.845

Spoth, R., Trudeau, L., Gyll, M., Shin, C., & Redmond, C. (2009). Universal intervention effects on substance use among young adults mediated by delayed adolescent substance initiation. *Journal of Consulting and Clinical Psychology*, 77, 620-632. DOI: 10.1037/a0016029

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797-811. DOI: 10.1037/0022-3514.69.5.797

Stone, C. A., & Sobel, M. E. (1990). The robustness of estimates of total indirect effects in covariance structure models estimated by maximum. *Psychometrika*, 55, 337-352.

Stone-Romero, E. F., & Rosopa, P. J. (2008). The relative validity of inferences about mediation as a function of research design characteristics. *Organizational Research Methods*, 11, 326-352. DOI: 10.1177/1094428107300342

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

Sundre, D. L., & Thelk, A. D. (2010). Advancing assessment of quantitative and scientific reasoning. *Numeracy*, 3(2), Article 2. DOI: 10.5038/1936-4660.3.2.2

- Sundre, D. L., Thelk, A. D., & Wigtil, C. (2008). *The Natural World Test, Version 9: A measure of quantitative and scientific reasoning, test manual*. Retrieved from <http://www.madisonassessment.com/uploads/NW-9%20Manual%202008.pdf>
- Sungur, S. (2007). Contribution of motivational beliefs and metacognition to students' performance under consequential and nonconsequential test conditions. *Educational Research and Evaluation, 13*, 127-142. DOI: 10.1080/13803610701234898
- Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188.
- Tanaka, A., & Yamauchi, H. (2001). A model for achievement motives, goal orientations, intrinsic interest, and academic achievement. *Psychological Reports, 88*, 123-135. DOI: 10.2466/pr0.2001.88.1.123
- Taylor, A. B., & MacKinnon, D. P. (2012). Four applications of permutation methods to testing a single-mediator model. *Behavior Research Methods, 44*, 806-844. DOI: 10.3758/s13428-011-0181-x
- Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education, 58*, 129-151. DOI: 10.1353/jge.0.0047
- Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology, 37*, 226-234. DOI: 10.1080/01973533.2015.1049351

- Thompson, T. (1996). Self-worth protection in achievement behaviour: A review and implications for counselling. *Australian Psychologist, 31*, 41-47.
- Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43*, 692-700. DOI: 10.3758/s13428-011-0076-x
- Tofighi, D., & MacKinnon, D. P. (2016). Monte Carlo confidence intervals for complex functions of indirect effects. *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 194-205. DOI: 10.1080/10705511.2015.1057284
- Valente, M. J., Gonzalez, O., Miočević, M., & MacKinnon, D. P. (2015). A note on testing mediated effects in structural equation models: Reconciling past and current research on the performance of the test of joint significance. *Educational and Psychological Measurement, 1*, 1-23. DOI: 10.1177/0013164415618992
- Vassiou, A., Mouratidis, A., Andreou, E., & Kafetsios, K. (2016). Students' achievement goals, emotion perception ability and affect and performance in the classroom: A multilevel examination. *Educational Psychology, 36*, 879-897.
- Weary, G. (1979). Self-serving attributional biases: Perceptual or response distortions? *Journal of Personality and Social Psychology, 8*, 1418-1420. DOI: 10.1037/0022-3514.37.8.1418
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*, 548-573.
- Weiner, B. (2000). Attributional thoughts about consumer behavior. *Journal of Consumer Research, 27*, 382-387.

- Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12*, 265-310. DOI: 10.1016/0273-2297(92)90011-P
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81. DOI: 10.1006/ceps.1999.1015
- Wigfield, A., & Eccles, J. S. (2002). *Educational psychology: Development of achievement motivation*. San Diego: Academic Press.
- Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test, *Applied Measurement in Education, 19*, 95-114, DOI: 10.1207/s15324818ame1902_2
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17. DOI: 10.1207/s15326977ea1001_1
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183.
- Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204-220). New York: Routledge.
- Wise, V. L. (2004). *The effects of the promise of test feedback on examinee performance and motivation under low-stakes testing conditions* (Doctoral Dissertation). Retrieved from <http://digitalcommons.unl.edu/dissertations/AAI3131570>

- Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*, 65-83.
- Wolf, L. F., & Smith, J. K. (1995). The consequences of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.
- Woodworth, R. S. (1928). Dynamic psychology. In C. Murchison (Ed.), *Psychologies of 1925*. Worcester, MA: Clark University Press.
- Worden, J. K., & Flynn, B. S. (2000). Effective use of mass media to prevent cigarette smoking. *Journal of Public Health Management & Practice, 6*, vii-viii.
- Wright, S. (1923). The theory of path coefficients a reply to Niles's criticism. *Genetics, 8*, 239-255.
- Wu, A. D., & Zumbo, B. D. (2008). Understanding and using mediators and moderators. *Social Indicators Research, 87*, 367-392.
- Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*, 301-322. DOI: 10.1037/a0016972
- Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*, 360-384. DOI: 10.1080/15305058.2014.928301