Spring 2013

Students’ attitudes toward institutional accountability testing in higher education: Implications for the validity of test scores

Anna Zilberberg
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/diss201019

Part of the Psychology Commons

Recommended Citation
Zilberberg, Anna, 'Students' attitudes toward institutional accountability testing in higher education: Implications for the validity of test scores' (2013). Dissertations. 71.
https://commons.libjmu.edu/diss201019/71

This Dissertation 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 Dissertations by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
Students’ Attitudes toward Institutional Accountability Testing in Higher Education:

Implications for the Validity of Test Scores

Anna Zilberberg

A dissertation submitted to the Graduate Faculty of
JAMES MADISON UNIVERSITY

In
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy

Assessment and Measurement

May 2013
Acknowledgements

First and foremost, I would like to acknowledge the members of my dissertation committee, Dr. Donna Sundre, Dr. Keston Fulcher, and Dr. Pete Swerdzewski for their continuous support, thoughtful feedback, and unwavering encouragement throughout the process. The current work was inspired by their professional practice and scholarly research.

The chair of my dissertation committee and my academic advisor, Dr. Sara Finney, deserves a special recognition for her unmatched contribution to my professional and personal development over the years. She always set the same standard of uncompromising quality to my work as she did to her own, and did everything to help me achieve that standard. It has truly been an honor and a privilege to receive my academic training under Sara’s mentorship.

In addition, I would like to extend a very special thanks to all faculty and staff at the Center for Assessment and Research Studies. I sought advice from every single one of them at different points in my graduate career, and always received insightful opinions and fresh perspectives. I feel incredibly lucky to be affiliated with such a fine community of scholars.

Also, I would like to recognize Dr. Olga Pierrakos from School of Engineering who has been exceptionally supportive over the course of my dissertation work as well as throughout the past five years. I could not have asked for a better supervisor and a colleague than Olga.

Finally, I would like to thank my family and friends – from far and near – for their companionship, support, sense of humor, and relentless encouragement.
# Table of Contents

Acknowledgements ........................................................................................................ ii

List of Tables .................................................................................................................. vi

List of Figures ................................................................................................................ viii

Abstract ......................................................................................................................... ix

I. Introduction .................................................................................................................. 1

  Accountability Testing in Higher Education .............................................................. 1

  What Is “Large-scale Low-stakes Standardized Institutional Accountability Testing”? . 3

  Validity of Higher Education Accountability Test Scores: The Role of Students ....... 5

  Students’ Attitudes toward Institutional Accountability Testing in Higher Education . 6

Current Program of Research ......................................................................................... 8

  Stage I: Collecting Evidence for Validity of the SAIAT-HE ................................. 8

    Research Question #1 ............................................................................................. 10

    Research Question #2 .......................................................................................... 13

    Research Question #3 ......................................................................................... 14

  Stage II: Substantive Research Questions Related to Students’ Attitudes toward

    Institutional Accountability Testing ....................................................................... 17

    Research Question #4 .......................................................................................... 18

    Research Question #5 ......................................................................................... 19

      Model 1: Fully-mediated Model ........................................................................ 21

      Model 2: Partially-mediated Model .................................................................... 25

      Model 3: Partially-mediated Model Including SAIAT-K-12 ............................ 28

    Research Question #6 .......................................................................................... 29

II. Review of the Literature ............................................................................................ 31
Era of Accountability in Education ................................................................. 31
Proliferation of Accountability Testing in K-16 ........................................... 31
The Use of Accountability Test Scores ....................................................... 31
The Stakes of Accountability Tests ......................................................... 32
Implications of Low Stakes on Test Performance .................................... 43
Test-Taking Motivation in Low-Stakes Conditions .................................. 44
Students’ Attitudes towards Institutional Accountability Testing ............... 49
College Students’ Attitudes toward Assessment in General ....................... 51
What Can We Infer About Students’ Attitudes Toward Accountability Testing? 53
Measurement of Students’ Attitudes toward Testing ................................ 54
Students’ Perceptions of Standardized Tests ............................................ 55
Students’ Conceptions of Assessment Inventory (SCoA-VI) ..................... 56
Psychometric Properties of the SAIAT-K-12 and the SAIAT-HE ............... 57

III. Methods .................................................................................................. 63
Measures ..................................................................................................... 63
Participants and Procedures ...................................................................... 67
Data Analyses ............................................................................................. 72
Data Screening and Assumptions ............................................................... 72
  Research Question #1 ............................................................................. 76
  Research Question #2 ............................................................................. 77
  Research Question #3 ............................................................................. 79
  Research Question #4 ............................................................................. 79
  Research Question #5 ............................................................................. 80
  Research Question #6 ............................................................................. 81
IV. Results .......................................................................................................................... 83
  Research Question #1 ..................................................................................................... 83
  Research Question #2 ..................................................................................................... 93
  Research Question #3 ..................................................................................................... 96
  Research Question #4 ..................................................................................................... 96
  Research Question #5 .................................................................................................... 99
    Model 1: Fully-mediated Model .................................................................................. 100
    Model 2: Partially-mediated Model ............................................................................. 108
    Model 4: Modified Fully-mediated Model .................................................................. 109
    Model 3: Partially-mediated Model Including SAIAT-K-12 .................................... 113
    Model 5: Modified Fully-mediated Model with SAIAT-K-12 .................................... 114
  Research Question #6 .................................................................................................... 116
V. Discussion ..................................................................................................................... 119
  Stage I ............................................................................................................................. 119
  Stage II ........................................................................................................................... 122
  Limitations and Areas of Future Study ......................................................................... 128
  General Conclusion ....................................................................................................... 131
Appendix A ....................................................................................................................... 132
Appendix B ....................................................................................................................... 134
Appendix C ....................................................................................................................... 136
Appendix D ....................................................................................................................... 137
References ......................................................................................................................... 138
List of Tables

Table 1: Factor-Item Map for Two Models and Two Versions of the SAIAT-HE ........152
Table 2: Sample Sizes ................................................................................................................153
Table 3: Alignment Between Samples, Measures, and Research Questions ...........154
Table 4: Demographic Data for Samples .......................................................................................155
Table 5: Correlations and Descriptive Statistics for the SAIAT-HE Items for First-year Student Sample ........................................................................................................................156
Table 6: Correlations and Descriptive Statistics for the SAIAT-HE Items for Mid-career Calibration Student Sample ................................................................................................................157
Table 7: Correlations and Descriptive Statistics for the SAIAT-HE Items for Mid-career Validation Student Sample ................................................................................................................158
Table 8: Fit Indices from CFA for SAIAT-HE (22 items) and the SAIAT-HE-revised (10 items) for the First-year, Mid-career Calibration, and Mid-career Validation Student Samples.................................................................................................159
Table 9: Correlation Residuals for the Six-factor Model for the SAIAT-HE..........160
Table 10: Item-level Model Parameters from the Six-factor CFA model on the SAIAT-HE from the First-year and the Mid-career Student Samples..........161
Table 11: Correlation Residuals for the Three-factor Model for the SAIAT-HE-revised (10 items) .................................................................................................................................162
Table 12: Item-level Model Parameters from the Three-factor CFA model on the SAIAT-HE-revised from the First-year and the Mid-career Student Samples ..........163
Table 13: Tests of Invariance of SAIAT-HE-revised Across First-year and Mid-career Validation Student Samples ..................................................................................................................164
Table 14: Mean Differences of Observed Subscale Scores and Latent SAIAT-HE-revised Factor Scores across First-year and Mid-career Students.................................165

Table 15: Correlations among the SAIAT-K-12 and SAIAT-HE-revised Subscales for Matched Incoming First-year Student Sample ........................................166

Table 16: Descriptive Statistics and Correlations among the Variables Used in Path Analysis for the Matched Incoming and First-year Student Sample ......................167

Table 17: Descriptive Statistics and Correlations among the Variables Used in Path Analysis for the Mid-career A-Day 1 Student Sample .............................................168

Table 18: Fit Indices for Competing Single-Indicator Latent Variable Models..........169

Table 19: Effects Decomposition for Model 4: Modified Fully-mediated Model in Matched Incoming and First-year Student Sample.................................170

Table 20: Effects Decomposition for Model 4: Modified Fully-mediated Model in Mid-career Student Sample.........................................................................................172

Table 21: Effects Decomposition for Model 5: Modified Fully-mediated Model with SAIAT-K-12 in Matched Incoming and First-year Student Sample ...............175

Table 22: Descriptive Statistics and Correlations among the SAIAT-HE-revised Subscales and Conscientiousness in the Mid-career A-Day 2 Student Sample...176

Table 23: Descriptive Statistics and Correlations among the SAIAT-HE-revised Subscales and Conscientiousness in the Mid-career Make-up Student Sample107

Table 24: Mean Differences on the SAIAT-HE-revised Subscales across Compliant and Non-compliant Students Before and After Adjusting for Conscientiousness .....177
List of Figures

Figure 1: Schematic of Partially-mediated and Fully-mediated Path Models ..................178
Figure 2: Model 1: Fully-mediated Model ....................................................................179
Figure 3: Model 2: Partially-mediated Model ...............................................................180
Figure 4: Model 3: Partially-mediated Model with SAIAT-K-12 ..................................181
Figure 5: Model 4: Modified Fully-mediated Model ...................................................182
Figure 6: Model 5: Modified Fully-mediated Model with SAIAT-K-12 ......................183
Abstract

Recent calls for an increase in educational accountability in K-16 resulted in an uptick of low-stakes testing and, consequently, an increased need for ensuring that students’ test scores are reliable and valid representations of their true ability. Focusing on accountability testing in higher education, the current program of research was comprised of two stages: (1) collecting validity evidence for a self-report measure; (2) investigating the relationship between students’ attitudes and other related constructs.

The analyses subsumed under the first stage yielded a revised psychometrically sound self-report measure of students’ attitudes toward accountability testing in higher education (SAIAT-HE-revised) consisting of three interrelated, yet conceptually distinct, subscales. Moreover, invariance of the SAIAT-HE-revised was upheld across first-year and mid-career students, indicating that the measure can be used across these two populations. In addition, known-groups validity evidence was garnered given that mid-career students, as predicted, held more skeptical attitudes than first-year students.

Subsumed under the second stage, a series of structural models examined the effects of attitudes on test performance via the mediating variables of test-taking effort and perceived importance of the tests. First, it was revealed that students’ attitudes toward accountability testing in K-12 were related to, but distinct from, their attitudes toward such testing in college, thereby relieving higher education administrators from needing to address negative attitudes toward K-12 testing in an effort to improve performance on college accountability tests. Second, the extent to which first-year and mid-career students were disillusioned by college accountability testing indirectly affected their performance via perceived test importance and test-taking effort. Third,
students’ perceived understanding of the tests’ purpose indirectly affected performance via test-taking effort and perceived importance. Fourth, the extent to which students perceived such tests to be fair and valid did not influence their test-taking motivation or test performance.

In addition, the relationship between attitudes toward college accountability tests and compliance with testing (i.e., attendance) was examined. Non-compliant students had lower levels of perceived understanding of the tests’ purpose than compliant students, but did not differ with respect to other attitudes.

In tandem, these findings indicate that an intervention aimed at improving test-taking motivation and compliance with testing should occur in the middle of students’ academic careers and focus on clarifying the purpose of testing. More positive attitudes toward college accountability testing are likely to improve test-taking behavior, thereby leading to more valid test scores, and thus more accurate evaluation of academic programming.
Introduction

Accountability Testing in Higher Education

Over the past decade, educational accountability reforms have been sweeping the nation, prompting institutions of elementary, secondary, and higher learning to integrate systematic assessment of students’ learning outcomes into their practice. In 2006, shortly after the reauthorization of the Elementary and Secondary Education Act (the “No Child Left Behind [NCLB] Act of 2001”), the Commission on the Future of Higher Education released a report titled *A Test of Leadership: Charting the Future of U.S. Higher Education* (Spellings, 2006). The report called for higher education institutions to improve in the areas of access, affordability, quality, and accountability. With respect to improvements in the accountability realm, the report called for higher education accrediting bodies to not only *encourage* but also *request* institutions seeking accreditation to report the results and use of the assessment of student learning outcomes. Expectedly so, the Spellings report prompted an escalation of assessment activities in public universities nationwide, similar to the hortatory effect that NCLB had in K-12.

The Spellings report left the specifics pertaining to measuring institutional effectiveness at the postsecondary level (i.e., type, conditions, frequency of assessment) up to the states, which led to variability in assessment methods (Ewell, 2009). However, more and more postsecondary institutions face the need to efficiently and reliably evaluate the extent to which all students are meeting institution-wide, as opposed to program-specific, learning objectives (e.g., general education outcomes), signaling the need for large-scale common assessments. According to the report published in 2009 by the National Institute of Learning Outcomes Assessment (NILOA) titled *More Than You Think, Less Than We Need: Learning Outcomes Assessment in American Higher Education*. 
Education, the majority of the 2,809 surveyed institutions have each adopted a set of institution-wide student learning outcomes that applies to all undergraduate students at a given institution (Kuh & Ikenberry, 2009). Therefore, assessment of such learning outcomes allows each institution to make claims about educational effectiveness on the institutional level (i.e., making inferences about overall institutional effectiveness as opposed to the effectiveness of individual programs or majors). The vast majority of surveyed institutions (92%) used at least one assessment approach to measure institutional-wide learning outcomes. With respect to the types of assessments used at the institutional level, performance assessments (e.g., portfolios, expert judgment reviews) were used quite rarely by the surveyed institutions. One possible explanation of the low prevalence of performance assessments is that they are costly and require substantial resources, making their use impractical. In contrast, selected-response standardized measures administered on a large scale were much more common because they readily yield results representative of the student body’s progress on the institution-wide learning outcomes. Specifically, two-fifths (39%) of the surveyed colleges and universities reported the use of externally-developed standardized measures such as Collegiate Learning Assessment (CLA) and Collegiate Assessment of Academic Proficiency (CAAP). The popularity of these assessments may be spawned by the mandates of the Voluntary System of Accountability (VSA), in which 319 public 4-year universities are participating. The VSA requires participating institutions to use CLA, CAAP, or Measure of Academic Proficiency and Progress (MAPP) to assess institution-wide student learning outcomes. However, other institutions that are not members of the VSA use locally-developed standardized assessments of institution-wide learning outcomes for meeting
accountability mandates. The NILOA report does not, unfortunately, provide information on the prevalence of such “home-grown” tests. Nonetheless, large-scale low-stakes standardized institutional accountability testing appears to be a promising assessment method for gauging the degree of institutional effectiveness and is likely to be prominent in the future. Next, some terminology associated with such tests is defined, illustrating the reasons such testing is likely to be more widely used in the future.

What Is “Large-scale Low-stakes Standardized Institutional Accountability Testing”?  

Defining the term large-scale low-stakes institutional accountability testing is essential to avoid confusion with other types of testing environments. Large-scale, as the name implies, refers to tests administered to a large representative sample of students across classrooms, departments, colleges, and even institutions or districts, depending on the desired level of generalization. Popham (2001) offers the following explanation of large-scale testing:

It’s all about accountability. Large-scale assessment programs, the bulk of which are of the high-stakes variety, are in place chiefly because someone believes that the annual collection of students’ achievement scores will allow the public and educational policymakers to see if educators are performing satisfactorily. (p. 34)

As emphasized above, the primary unit of analysis of such large-scale assessments is the institution, not the student, which is true for both higher education and K-12. The CLA is an example of a large-scale test used in higher education for institutional accountability, whereas Standards of Learning (SOL) is an example of a
large-scale test used in K-12 in the state of VA for institutional accountability. In addition, postsecondary institutions have the option to develop their own large-scale tests.

*Low-stakes* tests, in this manuscript, refers to tests that have no or negligible consequences directly for students. That is, these tests may be highly consequential for various stakeholders (e.g., institutions, districts, teachers), but do not carry high personal stakes for students. High-stakes tests, such as the SAT, stand in contrast to low-stakes tests due to highly consequential personal outcomes that hinge on test scores, such as admission to college.

*Standardized* tests are defined as tests containing content that is “equivalent across administrations” and given under “conditions that are the same for all test-takers” (Sireci, 2005, p. 113). In the world of educational testing, standardization is a way to ensure fairness and consistency of evaluation. There are several reasons standardized tests are particularly useful as measures of institutional effectiveness. First, standardized test scores facilitate generalizability of inferences about the population. For example, upon completion of a required general education program, scores from a random sample of college sophomores on a standardized measure of quantitative reasoning can be used to infer the performance of other students on this test under the same testing conditions who had completed the same coursework. In this manner, standardized tests differ from classroom-specific tests that are unique to each classroom and instructor. With classroom-specific tests, instructors decide on the format of the tests, the relative weight assigned to each item, the testing conditions, and other features. Ultimately, the purpose of classroom tests is to accurately ascertain whether individual students mastered the specific content covered in that particular class, whereas standardized test scores can be
aggregated across students for the purposes of making inferences about the performance of a larger student population. Second, and related to the first point, standardized tests allow for comparisons among student groups due to the equivalence in test format and conditions. This feature is important as it allows for a more rigorous assessment design, leading to a more refined investigation of program effectiveness, and ultimately more targeted programmatic improvement. For example, if a standardized test of quantitative reasoning is administered to a random sample of students who completed the required general education program and to a random sample of students who had not, the higher performance of the former group can be more strongly attributed to the program they completed as opposed to a myriad of other factors. This type of comparison could not be accomplished using differing classroom tests.

In summary, large-scale low-stakes standardized institutional accountability testing carries promise as a preferred assessment method for accountability purposes. From here on, the term “accountability tests” will be used to refer to large-scale low-stakes standardized institutional accountability testing used mainly for the purposes of evaluating institutional effectiveness. Despite the promise of such tests, responsible parties are concerned that scores from accountability tests are reliable and accurate representations of students’ knowledge. Specifically, the low-stakes context of such tests has drawn attention to students’ test-taking behavior and potential implications of such behavior on the validity of the resulting test scores.

**Validity of Higher Education Accountability Test Scores: The Role of Students**

A substantial body of research suggests that students’ performance on low-stakes tests (including accountability tests) is lower than their performance on high-stakes tests
(e.g., Cole & Osterlind, 2008; Napoli & Raymond, 2004). Nonetheless, there is considerable variability in students’ performance on low-stakes tests, indicating that other factors, besides the stakes of the test, affect performance (e.g., Thelk, Sundre, Horst, & Finney, 2009; Wise & DeMars, 2005). More specifically, variability in low-stakes test scores is partially explained by variability in test-taking motivation, which is defined as the perceived importance of the test and test-taking effort (Sundre & Kitsantas, 2004; Thelk et al., 2009; Zerpa, Hachey, Barneveld, & Simon, 2011). That is, students are not uniformly motivated in low-stakes testing contexts; instead, they tend to vary in their perceptions of test importance, as well as in their test-taking effort (Cole & Bergin, 2005; Thelk et al., 2009).

This variability in test-taking motivation begs the question “What influences test-taking motivation in a low-stakes testing session?” Answering this question is not an easy task. In fact, when discussing low-stakes testing for institutional accountability purposes, Erwin and Wise noted: “The challenge to motivate our students to give their best effort when there are few or no personal consequences is probably the most vexing assessment problem we face” (2002, p. 71). One possible influence of test-taking motivation is students’ attitudes toward accountability tests.

**Students’ Attitudes toward Institutional Accountability Testing in Higher Education**

According to Birenbaum and Feldman (1998), "In light of the effect assessment has on students, both as performers and as the objects (and often victims) of the decisions based on the assessment results, it is surprising to witness the paucity of research
regarding students’ assessment attitudes and preferences” (p. 91). So, what is the expected nature of such attitudes and their effects on test performance?

Evidence suggests that students’ attitudes toward *accountability testing in higher education* are likely to be negative and are also likely to undermine student performance. First, the low-stakes nature of such tests is reason for concern. As Huffman, Adamopoulos, Murdock, Cole, and McDermid (2011) noted, “low-stakes or non-consequential tests, appraised as having little value, can elicit negative affect in the test-taker” (p. 92). That is, an activity that is mentally strenuous but perceived as valueless is likely to engender negative emotion. Thus, the lack of personal consequences associated with low-stakes tests may result in negative affect, which may be the culprit of students’ low test-taking motivation, which in turn negatively influences test performance (Wise & DeMars, 2005). Second, upon examining the characteristics of accountability tests, it becomes evident that such tests are precisely the types of tests that students tend to dislike. More specifically, research indicates that students do not like assessments that result in no individual feedback, are mandatory in nature, and are often given as single-shot evaluations (see Struyven, Dochy, & Janssen, 2005 for a review). Third, older students (high school and beyond) experience more disillusionment with *all* standardized achievement testing (Paris, Roth, & Turner, 2000) and accountability measures tend to be standardized. Fourth and perhaps most importantly, today’s college students are yesterday’s middle- and high-school students who grew up under the repeated and publically disputed testing associated with the NCLB. According to the theory of generational development, events shared during the formative years (ages 10-18) impact the generational value system (Howe & Strauss, 2003). The experience of repeatedly
engaging in low-stakes standardized testing is a unique formative experience shared by
the millennial college students. Thus, it is plausible that students have developed attitudes
toward testing in high school and carried them over to college. Collectively, evidence
strongly suggests that today’s college students have fairly well-formed skeptical attitudes
towards accountability testing. Given that these tests have negligible, if any,
consequences to students directly, such attitudes are likely to have a negative impact on
how important students view the assessments, how much effort they invest in them, and,
ultimately, how well they perform.

As measurement professionals, we are concerned with bolstering the validity of
the inferences made from accountability test scores. Furthermore, as educators and
researchers, we are concerned with students’ attitudes toward any educational activity
and strive to discover ways of improving such attitudes. Therefore, disentangling the
relationship between students’ attitudes, test-taking motivation, and test performance is
paramount. Guided by the goal to understand the influence of students’ attitudes on the
validity of test scores, this program of research consists of two stages: (1) collecting
validity evidence for a recently developed measure of students’ attitudes toward
institutional accountability testing (SAIAT-HE); (2) conducting a number of studies to
address substantive research questions related to student attitudes toward institutional
accountability tests. These two stages are described next.

**Current Program of Research**

**Stage I: Collecting evidence for validity of the SAIAT-HE scores.** An
instrument was designed to measure students’ attitudes toward institutional accountability
tests administered in higher education (SAIAT-HE) by a team of assessment specialists
(Zilberberg, Anderson, Finney, & Marsh, in press). The following six inter-related dimensions were specified: (1) Validity, (2) Purpose, (3) Disillusionment, (4) Parents, (5) Professor, and (6) Students. The first three dimensions were meant to reflect students’ perception of accuracy of accountability test scores, usage of these test scores for identifying areas for improvement, and dissatisfaction with the tests, respectively. The last three dimensions were meant to reflect the notion that college students may have external influences affecting their attitudes; namely, their parents, professors and other students, respectively. Twenty-two items were written to represent six dimensions (see Appendix A). Note that the instructions prompted students to think specifically of accountability assessments completed throughout the course of college, with examples of these types of tests. As detailed in Chapter 2, the initial psychometric analyses of the Student Attitudes towards Institutional Accountability Tests in Higher Education (SAIAT-HE), based on a sample of mid-career students, revealed low reliability of the “influence” factors and identified several poorly functioning items. A single psychometric analysis notwithstanding, further investigation of the scale functionality is necessary for building a validity argument.

Building a validity argument for the SAIAT-HE involves the following steps. First, a set of psychometric analyses will be performed on the SAIAT-HE using independent samples of first-year students and mid-career students, with necessary revisions completed. Second, the extent to which the SAIAT-HE functions invariantly across first-year and mid-career students will be evaluated. Third, known-groups validity evidence will be examined by investigating the difference in attitudes toward higher education accountability tests between first-year and mid-career students.
**Research question #1.** What revisions to the SAIAT-HE are necessary based on results from first-year and mid-career student samples? As detailed in Chapter 2, the initial psychometric investigation of the SAIAT-HE using a mid-career student sample generally supported the fit of the correlated six-factor structure (Model 1 in Table 1). However, the results also suggested areas the SAIAT-HE can be improved, yielding a revised three-factor structure (Model 2 in Table 1). The current study of the SAIAT-HE will examine if the previous factor analytic findings replicate on two independent mid-career student samples as well as on a first-year students sample. Scale revisions will follow this set of psychometric analyses, with the SAIAT-HE being revised independently for first-year and mid-career students. Next, revisions are outlined in more detail.

Assuming the findings regarding the internal structure of the SAIAT-HE are replicated, the following scale revisions will take place. First, if the items representing the three “influence” factors (Parents, Professors, Students) continue to function poorly, then these three factors will be omitted from the scale. Why is there an expectation that “influence” factors will consistently function poorly for mid-career students? The items representing the Parents factor are likely to function poorly because parents are more distal from their children’s higher education experience than they are from the K-12 experience, especially when their children are sophomores or juniors. Thus, the majority of students should agree with statements about parental lack of involvement, yielding items with floor and ceiling effects and low variability. Given this finding was observed in the first study of the SAIAT-HE using mid-career students (Zilberberg et al., in press) and there is no reason to believe that parental involvement would be dramatically
different for an independent cohort of mid-career students, the Parents items are expected
to function similarly as they did in the previous study, with the majority of the Parents
items yielding floor/ceiling effects and low variability. The Student items are also
expected to elicit extreme responses (similarly to what was observed in the previous
study), with the majority of mid-career students strongly agreeing with a statement such
as “Fellow students speak negatively about the assessments tests at the University” and
strongly disagreeing with a statement such as “Fellow students urged me to try my best
on the assessment tests”. Such an extreme and consistent response pattern will result in
floor and ceiling effects. Finally, responses to the Professor items are expected to be
extreme or cluster around the middle of the scale, as they did in the original study of the
measure. That is, assuming that professors did not alter their communication to students
over the past few years, statements such as “My professor(s), an academic advisor, or
resident advisor encouraged me to prepare for the assessment tests” are likely to elicit
extreme responses with the majority of students disagreeing and little variability among
the responses. Items such as “My professors don’t value the assessment tests I complete
at the University” are expected to result in many students selecting “Neither agree nor
disagree”, reflecting uncertainty about professors’ values.

How are the “influence” factors expected to function for first-year students? With
regards to the Parents factor, these items are expected to function better for first-year
students than for mid-career students (i.e., greater variability, more salient item-factor
loadings, and fewer floor/ceiling response patterns). Given that incoming first-year
students have just left their parents’ home and have not yet established a social network
in college, it is likely that parents still have an influential role on these students’ attitudes
towards all academic activities. Furthermore, the Parents factor functioned well on a measure created to assess attitudes toward K-12 accountability tests, with students reporting that their parents exhibited some concern regarding performance on K-12 accountability tests (Zilberberg et al., in press). Quite possibly, parents express similar concern regarding their children’s performance on college accountability tests – at least at first. Thus, it is reasonable to expect that the SAIAT-HE Parents items are likely to be more salient to their factor for first-year students than for mid-career students. The Parents factor will be retained in the revised SAIAT-HE if it functions well for first-year students. With regards to the Professors factor, the expectation is that first-year students will consistently gravitate towards the middle of the scale (“Neither agree nor disagree”), reflecting their lack of experience interacting with the professors and lack of knowledge about professors’ values. Similarly, Student items are also likely to elicit extreme response patterns resulting in floor/ceiling effects and low variability, although the responses are likely to be less extreme than predicted for the mid-career students because first-year students will have been on campus for just a few days when assessments are administered, thus they will not have had a chance to learn what other students think about assessments. If the predictions about the functioning of the three “influence” factors hold for first-year students, the Parents factor will be retained, and the Professors and the Students factors will be omitted.

Pertaining to the other factors of the SAIAT-HE, minor revisions are expected for the Purpose and Validity factors for both first-year and mid-career student populations. In the first psychometric study of the SAIAT-HE, the Purpose item “Someone explained to me why I take standardized tests” correlated weakly with other Purpose items when
administered to mid-career students, indicating that being told the purpose of assessment is unrelated to actual understanding of the purpose of assessment. Therefore, it is expected that the responses to this Purpose item from both student samples will confirm low utility of this item and thus support for its deletion. Only 15% of the Validity item’s “Assessment tests are unfair to some students” variance was explained by the Validity factor. One explanation of this finding is that this Validity item taps into a conceptually distinct dimension of “test fairness”, thus resulting in poor saliency to the Validity factor and low correlations with other items. If the empirical results replicate for this item, it will be removed from the SAIAT-HE.

It is important to note that the same sample of first-year students will be used for all psychometric analyses and revisions of the SAIAT-HE, highlighting the need for replicating the results on an independent first-year student sample. Two independent mid-career student samples (validation and calibration samples) are available, allowing for replicating the results in the current study. See Chapter 3 for a complete description of the samples used to address this Research Question.

Research question #2. Does the SAIAT-HE function invariantly for first-year students and mid-career students? In order for researchers to interpret the SAIAT-HE scores the same way for both first-year students and mid-career students, it is necessary to establish measurement invariance of the SAIAT-HE across these two student groups (Brown, 2006).

Measurement invariance of the SAIAT-HE will be assessed in the following three steps: configural, metric, and scalar invariance. First, configural invariance, or the extent to which the same factor structure holds across student groups, will be investigated.
Fortunately, when answering the first research question above, it will become apparent if the same factor structure emerges for the two groups, thus indicating whether the two groups of students conceptualize the construct in the same way (Vandenberg & Lance, 2000). Assuming configural invariance holds, metric invariance, or the strength of item-factor relationships across these two groups will be tested. If factor pattern coefficients are found to be equivalent, thus supporting metric invariance, the salience of the SAIAT-HE items to their respective factors is identical across groups. Third, given configural and at least partial metric invariance, scalar invariance, or the equality of item intercepts across the two groups, will be tested. If established, scalar invariance will indicate that observed mean differences on the SAIAT-HE items reflect latent mean differences between the two student groups on SAIAT-HE factors.

If measurement invariance is established, the analysis will then proceed to group comparisons of latent means (which aligns with the next research question). In the event that measurement invariance does not hold, group mean comparisons (observed or latent) should not be conducted because such analyses are “meaningful only if the factor loadings and indicator intercepts have been found to be invariant” (Brown, 2006, p. 269). Nonetheless, if measurement invariance does not hold, this finding will be of interest in and of itself because it will imply that first-year and mid-career students either conceptualize the construct or use the measure differently.

**Research question #3.** Do mid-career students have worse attitudes about accountability tests than first-year students? Given that the SAIAT-HE functions equivalently for first-year students and mid-career students (i.e., measurement invariance is supported), the difference between first-year students and mid-career students’
attitudes will be assessed. This difference in SAIAT-HE scores serves as known-groups validity evidence if the difference aligns with theoretical expectations. Thus, theoretically expected differences between first-year students and mid-career students are discussed below.

Some evidence indicates that older students in K-12 (i.e., those approaching high school graduations) hold more skeptical views of assessment in general and of accountability tests in particular than younger students (i.e., those in middle school). For example, Paris and his colleagues (2000) examined differences in K-12 students’ (grades 4, 7, and 10) attitudes toward Michigan Educational Assessment Program (MEAP, a state-mandated low-stakes reading test) and found that older students harbor more negative attitudes toward MEAP and report investing less test-taking effort than younger students. Further, older students perceived MEAP scores to be invalid, as exemplified by the low percentage of 10th graders (8%) agreeing with the statement “This test is a good measure of my reading ability”, compared with 32% and 22% of 4th and 7th graders, respectively. Also, older students were more disillusioned about the MEAP, as exemplified by the low percentage of 10th graders (10%) agreeing with the statement “It is important for me to do well on this test.”, compared with 67% and 62% of 4th and 7th graders, respectively. Moreover, older students thought their parents did not value MEAP tests, as exemplified by the low percentage of 10th graders (16%) agreeing with the statement “My parents think my score on this test is important”, compared with 72% and 62% of 4th and 7th graders, respectively. In addition, the majority of 4th graders (77%) agreed with the statement “I tried to do my best on this test”, compared to 68% of 7th graders and 36% of tenth graders. This decrease in test-taking effort is likely indicative of
the decrease in positive attitudes towards these tests. In tandem, these findings suggest that students’ attitudes toward low-stakes accountability assessments follow a downward spiral.

Although Paris et al. (2000) only examined K-12 students’ attitudes, it is reasonable to assume that this downward spiral persists throughout the college years, resulting in mid-career students having worse attitudes than first-year students. In fact, a study comparing first-year and mid-career college students found that mid-career students perceived K-12 institutional accountability tests as less valid and tended to be more disillusioned (i.e., perceived assessments as a waste of time; Zilberberg et al., in press). Notably, these results apply to accountability tests in the K-12 context, not in college. Although this study did not compare first-year and mid-career students’ attitudes toward accountability tests in higher education, the descriptive analyses of mid-career students’ attitudes toward accountability tests in higher education showed that these older students hold neutral to negative views about the validity of such tests, are uncertain about the purpose of these tests, are slightly disillusioned about them, and have parents who exhibit little concern about students’ performance on such tests. In another study conducted in the college settings about half of the sophomore students identified as having “suspect” scores on a low-stakes accountability exam said that they did not understand the purpose of the test; about 47% said that their attitude toward taking the exam was not serious, and all students agreed that they did not approach the accountability test the same way they would a final course exam (Olsen & Wilson, 1997).

Collectively, the few empirical studies reviewed above suggest the following pattern of mean differences pertaining to first-year and mid-career college students’
attitudes about institutional accountability assessments in higher education. First, mid-career students are expected to perceive such tests as less valid than first-year students (i.e., mid-career students’ latent mean score on the SAIAT-HE Validity factor is predicted to be significantly lower than first-year students, with a moderate effect size). Second, mid-career students are expected to be slightly more informed about the purpose of accountability testing than first-year students, given their experience in college (i.e., mid-career students’ latent mean score on the SAIAT-HE Purpose factor is predicted to be significantly higher than first-year students, with a small effect size). Third, mid-career students are expected to be more disillusioned about such tests (i.e., mid-career students’ latent mean score on the SAIAT-HE Disillusionment factor is predicted to be significantly higher than first-year students, with a moderate effect size). Finally, whether or not the comparison of parental influence between first-year students and mid-career students can be made will depend on the results of the psychometric scale revisions (see research questions #1 and #2). If the Parents factor is retained for both student samples, it is expected that mid-career students will report less parental involvement than first-year students (i.e., their latent mean score on the SAIAT-HE Parents factor is predicted to be significantly lower than first-year students, with a moderate effect size).

**Stage II: Substantive Research Questions Related to Student Attitudes toward Testing**

The following substantive research questions involving the use of the SAIAT-HE are based on the assumption that adequate validity evidence for the SAIAT-HE has been garnered in the previous stage. Of the six factors originally proposed, it is expected that at least three of the factors will be supported and will be measured adequately across the
two student populations. That is, Validity, Purpose, and Disillusionment should be supported for both first-year and mid-career students and invariance of factor loadings and intercepts should be found. Moreover, known-groups validity evidence is expected to support the interpretation of the scores.

Research question #4. How strongly are first-year students’ attitudes toward K-12 accountability tests related to their attitudes toward higher education accountability tests? The strength of the relationship between first-year students’ attitudes toward K-12 accountability tests and their attitudes toward such tests in college has important implications for understanding the development and consequences of such attitudes.

Examining incoming first-year college students’ attitudes is of particular interest in this regard. Essentially, incoming students are in the transition stage – they are recent high school graduates and newly minted college students who have not yet experienced the college classroom. They likely have well-established attitudes toward K-12 accountability tests due to their recent exposure to such tests. Despite the fact that first-year students are encountering college accountability testing for the first time, they are likely to have had developed a cognitive schema of “accountability testing”, and are thus likely to view this novel experience through the lens of their prior experience. If this hypothesis is true and students’ attitudes engendered in K-12 do carry over to college, then important implications arise for higher education assessment practice. That is, if negative attitudes develop in K-12 and carry over to college, then students’ early experiences need to be taken into account and efforts aimed at improving such attitudes need to have a broad focus.
Theoretically, it is expected that first-year college students will have very similar attitudes toward accountability tests completed in K-12 as they do toward accountability tests completed in college. Although there are indisputable differences between K-12 and higher education accountability tests (e.g., different contexts, frequency of administration, differences in stakes), the similarities between the types of assessments are likely to elicit the same affective responses from students. These similarities include, but are not limited to, mandatory participation, low personal stakes attached to the performance, the use of test scores to measure school effectiveness, large-scale nature of test administration, predominantly multiple-choice test format, and standardized conditions.

In the proposed study, students’ attitudes toward K-12 accountability tests (measured using SAIAT-K12, described in detail in Chapter 2) will be assessed the summer prior to the fall semester and the same students’ attitudes toward higher education accountability tests (SAIAT-HE) will be measured at the beginning of the fall semester. Bivariate correlations among corresponding SAIAT-K12 and SAIAT-HE factors are expected to be positive and range from moderate to strong for all three factors expected to be retained after scale revisions.

**Research question #5.** What is the nature of the relationship between students’ attitudes towards accountability testing, test-taking motivation, and performance on higher education accountability test? The ultimate goal of assessing students’ attitudes toward institutional accountability testing in higher education is untangling the relationship between such attitudes, test-taking motivation, and performance on accountability tests. The relationships between these variables can be conveyed via
different path models (Figures 1, 2, 3). Prior to detailing such models specific to the
current program of research, it is important to discuss the differences between fully- and
partially-mediated models in general and their respective roles in theory building and
hypotheses testing.

Fully- and partially-mediated path models. Figure 4 depicts two simple path
models of the relationship between predictor (X) and an outcome (Y). In the fully-
mediated model (bottom) the mechanism through which the predictor influences the
outcome is fully explained via an intervening variable (M) called a mediator. That is, a
causal relationship is “transmitted” from X to Y via M; this indirect effect is represented
by a causal link from a to b. In the partially-mediated model (top), a direct effect between
X and Y is represented by a causal link c. This direct effect implies that X and Y are
related above and beyond what is explained by M.

Fully-mediated models are highly restrictive and thus rarely fit the data (Rucker,
Preacher, Tormala, & Petty, 2011). Partially-mediated models, on the other hand, are
more flexible and allow for testing more hypotheses. According to Shrout and Boulger
(2002), direct effects in the partially-mediated models can be interpreted in one of the
following ways. One explanation is that X has a theoretically-meaningful direct effect on
Y, in addition to the indirect effect via M. Alternatively, the direct effect can represent
the effects of unmeasured mediating variables that are omitted from the model,
essentially capturing misspecification of the fully-mediated model. That is, in the top part
of Figure 4, in addition to X directly affecting the mediator M, which in turn affects Y,
predictor X may also have a direct effect on Y or may directly affect other mediators not
included in the model. As such, a direct effect of X on Y is needed in order to fully reproduce the observed relationship between X and Y.

In sum, when predictors have both direct and indirect effects on an outcome, it may be due to a direct causal link or it may represent a presence of an unmodeled mediator. Fully- and partially-mediated models specified to represent theoretically expected relationships between variables of interest in the current program of research are described next.

Model 1: Fully mediated model. The relationship between attitudes towards higher education accountability tests and test performance may be fully mediated via test-taking effort and perceived importance (two components of test-taking motivation), after controlling for individual differences in verbal and math ability (see Figure 1).

A direct positive effect is specified between effort and performance. Theoretically, the amount of effort expended on a task is expected to predict performance on that task (Wigfield & Eccles, 2000). Moreover, several empirical studies found a moderate positive correlation (~.30) between effort and performance (Thelk, 2006; Wise and Kong, 2005; Wise, Wise, & Bhola, 2006). Furthermore, a direct path is specified between importance and effort to reflect their empirically supported positive relationship (e.g., Sundre & Wise, 2003; Sundre & Moore, 2002) and to represent the hypothesis that perceived importance of a test has an indirect effect on test performance, mediated through test-taking effort (Sundre & Kitsantas, 2004; Wise & DeMars, 2005).

With respect to the effect of attitudes on performance, from one theoretical standpoint, students’ attitudes toward assessment (i.e., Validity, Purpose, Disillusionment) directly affect perceived importance of the tests, which in turn directly
affect effort, which in turn affects performance. To reflect the theoretical expectation that students who perceive the tests to be fair and valid also view them as more important, a direct positive effect is specified between Validity and Importance. This hypothesis is based on a qualitative study that revealed that students refer to the validity of the tests as “test fairness” and repeatedly bring it up as a determining feature of how important the tests are (Sambell, McDowell, & Brown, 1997). Although a more recent study with college students did not find a significant direct effect of student attitudes toward K-12 accountability tests on perceived importance of higher education accountability tests (Zilberberg, Finney, Marsh, & Anderson, in progress), this effect is likely to emerge in the current study, given a closer proximity between the constructs.

Also, a direct positive effect is specified between Purpose and Importance to reflect the notion that educating students about the purpose of accountability tests should positively affect the extent to which they perceive such tests as important. Empirical evidence provides support for this direct relationship. For example, Brown and Gaxiola (2010) found that explaining to students that testing was used by the university to improve instruction positively affected test-taking motivation. Moreover, Swing (2001) reported that implementation of a new assessment plan that focused on clearly communicating the purpose of accountability testing likely elicited voluntary and whole-hearted participation from students. In addition, Zilberberg et al. (in progress) found a significant positive effect of students’ perceptions of purpose of K-12 accountability tests on their perceived importance of higher education accountability tests ($\beta = .15$), further justifying a direct path between Purpose and Importance specified in the current study.
A direct negative effect is specified between Disillusionment and Importance, reflecting the notion that students who are more disillusioned about higher education accountability assessment tend to perceive these tests as less important. According to Paris and his colleagues, older students tend to view standardized state-mandated tests as less important than younger students, presumably because familiarity with tests over time “breeds contempt and distrust” (2000, p. 23). It follows that disillusionment (i.e., negative affect that may subsume contempt and distrust) directly and negatively affects importance.

Finally, direct effects of both SAT-verbal and SAT-math on test performance are included in the model to control for potential pre-existing differences both in mathematical and verbal ability because the performance measure is an assessment of global experience and requires both text comprehension (i.e., verbal ability) and interpretation of data and graphs (i.e., quantitative ability). Ability (for which SAT scores serve as proxies) has been shown to be an important predictor of academic performance in college in general (Robbins, Lauver, Le, Davis, Langley, & Carlstrom, 2004) and of performance on low-stakes accountability tests in particular (Zilberberg et al., in progress; Steedle, 2010; Sundre & Wise, 2003). No direct effects are specified between SAT scores and Effort because studies suggest that the relationship between ability and effort is negligible (e.g., Wise, 2009; Wise & DeMars, 2005). With respect to perceived importance, ability is expected to have a weak negative relationship, thus a direct negative effect is specified between SAT-math, SAT-verbal and Importance. That is, although Snyder (2012) found that middle school students of higher ability tended to ascribe more importance to non-consequential tests than students of lower ability, a study
using a college student sample suggests the opposite. Zilberberg et al. (in progress) found a slight negative significant relationship between SAT-math and perceived Importance ($r = -0.10$). This slight negative relationship is quite puzzling because it suggests that higher math ability is associated with lower perceived importance of accountability tests. One explanation is that students with higher SAT-math scores have increased their score by taking the test multiple times, which also increased their frustration with the whole experience and led them to perceive all standardized tests (including those used for accountability purposes) as unimportant. Another explanation is that the non-zero direct effect reflects the presence of another, unmeasured variable that either causes or mediates the relationship between SAT-math and Importance (Shrout & Bolger, 2002). Although the reasons for the direct negative relationship between SAT-math scores and Importance are not immediately clear, it is of interest to test whether the weak direct effects will replicate in the current study.

Model 1 will be fit to two samples: first-year students and mid-career students. The directions of the relationships are hypothesized to be the same across the two samples. Although mid-career students are expected to differ in attitudes toward accountability testing in higher education from first-year students (i.e., perceive tests as less valid, be more disillusioned by them, but recognize greater purpose to them), the relationship between these attitudes, the mediating variables (importance, effort), and performance is expected to be similar. That is, no large differences are expected to emerge with respect to model fit, variance explained, and significance of effects across first-year and mid-career student samples.
Fully-mediated Model 1 is quite stringent. According to this model, the effect of students’ attitudes toward higher education accountability tests on test performance is *fully* mediated (i.e., completely transmitted) by perceived importance of the tests and test-taking effort. Zilberberg et al. (in progress) tested a similar fully-mediated model (although these authors investigated the relationship between college student attitudes toward *K-12 tests* and performance on a *college-level* quantitative/scientific reasoning measure). Although the fully mediated model tested by Zilberberg and her colleagues yielded good global and localized model-data fit, it left a substantial proportion of unexplained variance in Importance (.95), Effort (.81), and Performance (.74), prompting the authors to reject the fully-mediated model and champion an alternative partially-mediated model. The fully-mediated Model 1 specified in this study may not fit the data well or may leave a substantial amount of variance in perceived Importance, Effort, or Performance unexplained. To explore alternative explanations of the mechanisms underlying the relationship between attitudes and performance, partially-mediated models are specified. Next, theoretical rationales for these partially-mediated models are outlined.

**Model 2: Partially mediated model.** In the current study, a partially mediated model with added direct effects is specified in order to investigate whether there is a direct causal link between attitudes and performance or whether another variable, not currently modeled, mediates this relationship. Although a review of the literature did not reveal too many obvious variables mediating this relationship, there is evidence to indicate that some important variables may be missing from the current model.
One such variable is “academic citizenship” – a term coined by Wise (2009) and defined as students’ willingness to help improve the performance of their institution. Students’ attitudes towards the methods their institution employs to measure its effectiveness (i.e., accountability tests) are likely to influence their sense of academic citizenship, which in turn may affect students’ performance. Research suggests that students have strong attitudes regarding test validity (referred to by students as “test fairness”) which is viewed by students as a fundamental aspect of assessment (Sambell et al., 1997). However, if and how students’ perception of accountability test validity (i.e., directly or indirectly via mediating variables) is related to their performance is unclear. It is plausible that the extent to which students perceive tests to be fair or valid affects perceived importance of the test, which in turn affects effort, which in turn affects performance (as explicated in Model 1). However, it is also plausible that the extent to which students perceive tests to be fair or valid increases students’ sense of academic citizenship, which in turn affects their performance. A direct path specified between Validity and performance allows us to investigate the possibility of either a direct relationship between these two variables or the potential presence of a mediator (e.g., academic citizenship).

With respect to the other test attitudes (i.e., Purpose and Disillusionment), similar logic applies. The extent to which students recognize the purpose of accountability assessment (i.e., to improve one’s university) may affect one’s sense of academic citizenship, which may in turn influence performance. Furthermore, a previous study by Zilberberg et al. (in progress) found a significant, although small, direct effect of SAIAT-K-12 Purpose on performance in their partially mediated model, which may suggest
either that recognizing the purpose of K-12 assessments directly affects performance or that another variable(s) mediates this relationship. Given the slight direct relationship between students’ recognition of the purpose of the K-12 tests and their performance on the college accountability test, it is likely that students’ recognition of the purpose of higher education accountability testing and their performance on such a test is likely to be stronger, due to a more proximate connection between these variables. As such, a direct path is specified between Purpose and performance. In the similar vein with the arguments made above, students’ disillusionment with accountability testing is hypothesized to negatively affect their sense of academic citizenship. In turn, students who feel less of an obligation to improve the performance of their institution may perform worse on such a test. Alternatively, there may be a direct relationship between Disillusionment and performance, not mediated by academic citizenship. Importantly, specifying direct paths between attitudes and performance will not provide an empirical test in support of academic citizenship because this variable was not measured. Nonetheless, testing Model 2 will reveal whether partial mediation explains the relationship between attitudes and performance better than full mediation, thus leading to the next step of exploring alternative mediators (such as academic citizenship) or direct effects.

In summary, specifying direct paths from Purpose, Validity, and Disillusionment to performance, in addition to the indirect paths (through importance and effort) represents two very different interpretations that can’t be distinguished empirically: 1) other variables missing from the model (e.g., academic citizenship) mediate this relationship; or 2) attitudes and performance are related to one another directly. In either
case, testing this partially mediated model is informative as it assess whether test-taking motivation sufficiently mediates the relationship between attitudes and performance or whether alternative explanations need to be investigated.

Model 3: Partially mediated model including SAIAT-K12. This partially-mediated model is very similar to Model 2 in that the relationship between attitudes towards higher education accountability tests and test performance is specified to be partially mediated via test-taking effort and perceived importance, after controlling for the effects of verbal and math ability on importance and performance. In addition, the three inter-related attitudes towards higher education accountability tests (Validity, Purpose, Disillusionment) are expected to be directly influenced by three corresponding attitudes towards K-12 accountability tests (see Figure 3). The rationale for this model is outlined below.

Today’s college students have been acclimated to taking accountability tests since K-12 and their past experience is likely to matter in the present (Snyder, 2001). Quite possibly, attitudes formed over the course of K-12 carry over to college and affect student attitudes toward college accountability tests. Recall that RQ4 focused on the bivariate relationships between students’ attitudes toward institutional accountability tests in K-12 (SAIAT-K-12) and their attitudes toward such tests in higher education (SAIAT-HE). In addition to examining the simple bivariate relationship between these attitudes, it is also informative to assess the effects of attitudes toward K-12 accountability tests on performance on higher education accountability tests via their effects on attitudes toward higher education tests. The SAIAT-K-12 scores are likely to be strong predictors of the SAIAT-HE scores, which may then transmit the effect of attitudes toward K-12
accountably tests on higher education tests performance. The SAIAT-HE factors are expected to be inter-related, or share variance above and beyond what is explained by the SAIAT-K-12 factors – that is depicted in Figure 3 by the correlated errors among the factors.

Zilberberg and colleagues (in progress) found very weak direct effects of attitudes toward K-12 accountability tests on perceived importance of and performance on higher education accountability tests, concluding that attitudes toward K-12 accountability tests do not have an influence on test-taking behavior on higher education accountability tests. However, the weak effects observed by Zilberberg and her colleagues may be due to model misspecification. That is, student attitudes toward higher education accountability tests (excluded in Zilberberg et al. study) may be important mediators of the relationship between attitudes toward K-12 accountability tests and performance on higher education accountability tests. This hypothesis is reflected in Model 3.

Research question #6. Do students who skip versus attend A-Day have different SAIAT-HE score profiles? Students who choose not to attend scheduled assessments and instead attend a mandatory make-up session (i.e., non-compliant students) tend to differ from students who complete their assessments when they are scheduled (i.e., compliant students; Brown & Finney, 2011; Swerdzewski, Harmes, & Finney, 2009). A recent study conducted by Swerdzewski and his colleagues found that non-compliant examinees tend to be male, slightly older, with a somewhat lower average GPA than compliant examinees. Furthermore, non-compliant examinees scored much lower than compliant examinees both on test-taking effort ($d = 0.42$) and perceived importance of the test ($d = 0.25$) (Swerdzewski et al., 2009). Also, research shows that non-compliant examinees
tend to be more academically entitled (Kopp & Finney, in press), and more psychologically reactant (Brown & Finney, 2011) than compliant examinees. As such, it is of interest to also examine whether compliant and non-compliant examinees also differ in their attitudes toward higher education accountability testing.

Prior meta-analytic research indicates that conscientiousness is a strong positive predictor of academic success as measured by various proxies (GPA, exam grade, essay grade, etc.) (O’Connor & Paunonen, 2007; Poropat, 2009) as well as the quality of effort expended with academic and personal/social activities (Bauer & Liang, 2003). Therefore, it is likely that academically compliant students who attend their scheduled assessment session are higher in conscientiousness than their non-compliant peers. As such, it is of interest to further examine the differences on the SAIAT-HE subscale means across compliant and non-compliant examinees after controlling for conscientiousness.
Review of the Literature

Era of Accountability in Education

**Proliferation of institutional accountability testing in K-16.** Historically, assessment of student learning outcomes for the purposes of evaluating performance of individual students has been an integral part of American education at the primary, secondary, and tertiary levels. Over the past two decades, in response to national and international calls for institutional accountability, assessment of student learning outcomes has also been used to evaluate the effectiveness of institutions (Hamilton, Stecher, & Klein, 2002). In contrast to measuring educational *inputs* (e.g., teachers’ salary, class size, resources), the focus on the measurement of desired educational *outputs* (i.e., students’ learning outcomes) has become the cornerstone of the recent accountability movement both in K-12 and higher education contexts. Federal mandates for educational accountability, such as No Child Left Behind (NCLB, 2002) and Race to the Top (Obama, 2009), spurred an increase in the output-focused accountability initiatives in K-12, whereas the Spellings report of 2006 had a similar effect in higher education.

**The use of accountability test scores.** Used primarily for evaluating institutional effectiveness (although in some cases accountability tests are used for teacher and/or student evaluations), these test scores are meant to draw public attention to achievement disparities, assist in allocation of resources and remedial assistance and reporting of state-imposed accountability demands (Haertel & Herman, 2005; Linn, 2000). Given the importance of such accountability tests, it is essential to determine that the scores yielded by them are *reliable* representations of students’ true ability upon which decision makers...
can make valid inferences. Furthermore, several professional organizations call for efforts to build strong validity arguments for the inferences made based on the test scores (Standards for Educational and Psychological Testing; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Given the issue of validity of test scores is closely tied to the use of test scores, it becomes especially important to consider how accountability test scores are used in different educational contexts (i.e., K-12 and higher education). In other words, whether the tests are used for assessing institutional effectiveness only (i.e., high-stakes for institutions and low-stakes for students) or for determining students’ promotion or graduation (i.e., high-stakes for students) has important implications for the validity of such test scores. Given the explicit purpose of accountability testing is to evaluate institutional effectiveness, it follows that such tests have at least moderate, and sometimes high stakes (depending on the sanctions) for schools, districts, and states. However, whether accountability tests are high-stakes for students is a question that warrants further discussion, because the consequences attached to test scores could play an important role in students’ performance on the tests.

The stakes of accountability tests. According to Cole and Osterlind (2008) “understanding the consequences of a test is the critical element in deciding whether the test is low- or high-stakes” (p. 120). Tests that are universally recognized as low-stakes serve informational purposes only and are often used for national or international comparisons. For example, Trends in International Mathematics and Science Study (TIMSS) is a test of mathematics and science achievement that has been administered to
4th and 8th grade students in over 60 countries, with no consequences associated with performance for students, schools, or districts. Such tests stand in contrast to tests that are used for making high-stakes individual decisions, such as college acceptance (e.g., SAT). When it comes to institutional accountability testing, the consequences associated with test scores are not as clear-cut as they are for tests such as TIMSS (no stakes for any entity) or the SAT (personal high stakes). That is, accountability tests often have high consequences for institutions and teachers, but may have moderate to low consequences for students. In the current work, accountability tests with no to negligible consequences for students (whether K-12 or higher education) are referred to as low-stakes tests, regardless of how high the stakes of these tests may be for districts, institutions, and/or teachers.

It is important to carefully consider what the stakes of institutional accountability testing in K-16 are for students. In some cases, consideration of the test difficulty is necessary in order to judge whether students are likely to exert their optimal test-taking effort. For example, passing a test may be required for a high-stakes individual outcome, such as graduation, but if the passing score is set very low, then students only have to exert minimal effort to achieve the needed passing score. Similarly, if non-attendance of the testing session results in a hold placed on a student’s academic record (i.e., a high-stake individual outcome), then students are likely to attend and even give some minimal effort to the test. However, it is doubtful that students will exert the same optimal amount of effort on such tests as they would on the tests with high personal consequences (e.g., classroom exams). In short, optimal performance is dictated not only by the direct consequences of the test, but also by the difficulty of the task at hand.
What are the stakes for K-12 students? The NCLB Act places accountability mandates on K-12 public schools in all 50 states (U.S. Department of Education; NCLB, 2001). The NCLB act is based on four main cornerstones: (1) stronger accountability for results, (2) increased flexibility and local control, (3) expanded options for parents, and (4) an emphasis on teaching methods that have been proven to work (Hamilton et al., 2002, p. 6). The basic premise at the heart of this legislative action is that states, schools, and districts are held accountable to the federal government for students’ achievement by annually assessing students’ progress in math and reading in grades 3-8 and at least once in science during elementary, middle, and high school. The measure by which states, schools, and districts are held accountable is called adequate yearly progress (AYP). AYP is a set of the measurable test-based objectives annually specified by each state every year for each demographic subgroup (e.g., English language learners, ethnic minorities and majorities). Section 1111 (b)(F) of the NCLB requires that "each state shall establish a timeline for adequate yearly progress. The timeline shall ensure that not later than 12 years after the 2001-2002 school year, all students in each group will meet or exceed the State's standards". That is, the states must determine their own AYP proficiency standards in reading and math for each demographic subgroup, test all students annually in grades 3 through 8, and publically report AYP towards the proficiency goal (NCLB, 2001). The federal parameters stipulate that in order to make AYP, at least 95 percent of students in each of the subgroups must participate in state assessments, and each subgroup of students must meet or exceed the state-determined annual benchmarks for that year (NCLB, 2001). The baseline or starting point is based on the lowest-achieving demographic group or the lowest-achieving school (whichever one is higher) and
subsequent thresholds for the expected increase in students’ performance must be raised every three years so that state-determined proficiency levels are reached in 12 years. The yearly assessments must be reported in the form of annual report cards and made available to educators, administrators, policymakers, parents, and the public at large. Whether the school makes their statewide AYP is determined by annual peer review (NCLB, 1111(b)(2)).

The stakes associated with accountability tests in K-12 for schools and teachers are quite high. Failure to meet AYP leads to an escalated set of corrective actions. After two consecutive years of failing to meet AYP, a school is identified as “in need for improvement” and a plan is developed in consultation with parents, school staff, districts, and outside experts. The plan must include technical assistance offered to schools as well as supplemental services and an option to transfer to a different school offered to students. If a school continues to fail to make AYP for the next two years after identification, the district must take more drastic corrective actions, including restructuring the internal organization of the school (i.e., replacing teachers and administrators), diminishing management authority of the school (i.e., state takeover), implementing new curriculum, and appointing outside experts (U.S. Department of Education, 2001). Such underperforming schools are labeled as “priority” or “focus” schools. For schools that do meet or exceed their AYP, NCLB requires that states provide these schools with state academic achievement awards and allows the use of Title I (federal aid program for disadvantaged) funds for financial rewards to teachers. These schools are labeled as “reward” schools.
As outlined above, all states must conform to the federal expectations tied to meeting AYP\(^1\), making accountability tests in all states high-stakes for schools and teachers. Beyond meeting the NCLB federal mandates, the use of test scores is largely up to the individual states and depends on the state political culture that acts as “a filter for implementation of national reforms and policies” (Marshall, Douglas, & Frederick, 1989, p. 4). Namely, the stakes associated with accountability tests in K-12 for students, which could impact student effort on these tests, vary from state to state, making it necessary to examine the NCLB policy at the state level.

The testing policies implemented in different states tend to vary considerably (Greene, Winters, & Forster, 2003; Heinecke, Curry-Corcoran, & Moon, 2003). More specifically, whether or not accountability test scores are used for making student-level decisions, as well as what tests are used, varies by state. Greene et al. (2003) reported differences among the states in the types of tests used for K-12 accountability purposes, the grades in which these tests are administered, the exact consequences for students attached to the test results, and whether the test scores are used for promotion and/or graduation. For example, Ohio and Massachusetts require that students pass the state exam in order to receive a high school diploma. Chicago public schools require students to pass a certain benchmark score on the Iowa Test of Basic Skills in order to move on to the next level (notably, this requirement does not apply to the whole state of Illinois, further highlighting variability in policy even within states). The state of Florida is especially aggressive in their use of K-12 accountability tests. Florida students are required to pass the state exam in order to move from the 3\(^{rd}\) grade to the 4\(^{th}\), and are also required to pass all sections of the 10\(^{th}\) grade exam in order to graduate. In states where
assessment results lead to the labeling of students, the detrimental effects on the academic self-esteem of lower-achieving students have been documented. According to Broadfoot and his colleagues (1992), “a competitive climate produced by labeling” results in “the already successful thriving and the less successful (always the majority) underachieving more and more as they are repeatedly discouraged by the way their efforts are judged.” (p. 6) More recent empirical research supports this claim. In their paper titled “The High-Stakes Effects of ‘Low-Stakes’ Testing”, Papay, Murnane, and Willett (2011) document the effects of labels accompanying individual students’ performance on state-mandated tests (e.g., failing, in need of improvement, proficient) on students’ college plans in Massachusetts. They found that for urban, low-income students in 10th grade who did not originally plan on attending a four-year college, earning a more positive performance label had a substantial positive impact on their decision to attend college, whereas earning a more negative performance label had the opposite effect. This finding suggests that, for these students, state-mandated tests with the labeling system have relatively high-stakes.

What about Virginia, the state that is considered to be at the forefront of the accountability movement (Department of Education, 2001)? Margaret Grogan in her review of Educational Leadership in an Age of Accountability: The Virginia Experience by Daniel Linden Duke (2003) says Virginia’s response to accountability initiatives was a complex one, but one “overarching effect stands out clearly – the importance of high-stakes tests.” Standards of Learning (SOL) tests implemented in VA are unarguably high-stakes for schools due to important accountability decisions hinging on the results of SOLs. For example, under Standards of Accreditation (SOA), any Virginia school with
more than 30% of students failing the SOLs in 2007 was subject to the loss of accreditation (Schrag, 2000). However, SOL tests do not have such high consequences for students directly. At the elementary and middle school levels, SOLs do not impact a student’s promotion to the next grade level, and repeated failing of SOLs is followed only by optional remediation (Virginia Department of Education). At the high school level, however, passing SOLs is required for advancing to the next grade level and, ultimately, graduation (66% correct). Further, whether a student receives a modified, standard, or advanced studies diploma depends on a number of “verified unit credits” (SOLs) passed. In addition, the requirements differ based on a county; some counties within VA raise their achievement bar (US Department of State). Nonetheless, all VA students are allowed unlimited retakes of the SOLs until they get a passing score (Virginia Department of Education). So, how likely are the VA students to exert optimal effort on such SOLs? Logically, if achieving the required graduation passing score is not a major cause of concern, multiple attempts are allowed, and tests themselves are fairly easy, then students are likely to exert only a minimal amount of effort on SOLs. This begs the question, “What is the difficulty of the SOL tests?”

Examination of the VA report card provides a snapshot of students’ overall performance on SOLs, thus allowing us to crudely gauge how difficult the SOLs are for VA students. In 2008-2009, 89% passed the English part of SOL, and 86% passed the Math portion of SOL. In 2009-2010, 89% passed the English part of SOL, and 88% passed the Math portion of SOL. In 2010-2011, 88% passed the English part of SOL, and 87% passed the Math portion of SOL. In addition, all student groups must meet Adequate Yearly Progress (AYP) objectives. In 2008-2009, all student groups met the objectives.
In 2009-2010, the following groups did not meet the objectives related to English performance: black, economically disadvantaged, students with disabilities. In 2009-2010, students with disabilities were the only group that did not meet the objectives related to Math performance. In 2010-2011, the following groups did not meet the objectives related to English performance: black, economically disadvantaged, Hispanic, limited English proficient, and students with disabilities. In 2010-2011, the same demographics groups did not meet the objectives related to Math performance. In summary, the SOL passing rates needed for progressing to the next grade level have been consistently high over the past several years and if a student fails to pass the SOL test they are offered additional help as well as unlimited retakes. Thus, it follows that the SOL tests can be considered fairly low-stakes for students, at least for the demographic majority. In summary, the stakes associated with state-mandated institutional accountability tests on the K-12 level tend to differ depending on the state and the demographic characteristics of the students. Next, the stakes associated with institutional accountability testing in the higher education context are considered.

**What are the stakes for university students?** In 2005, a few years after the NCLB act of 2002, the U.S. Secretary of Education Margaret Spellings formed the Commission on the Future of Higher Education, also known as the Spellings Commission. This commission released a report in 2006 titled A Test of Leadership: Charting the Future of U.S. Higher Education, more commonly known as the Spellings report of 2006. The report focused on four main areas pertinent to post-secondary and higher education in the U.S.: (1) access, (2) affordability, (3) quality, and (4) accountability. The latter focus area – accountability of public higher education institutions to their constituencies (students,
families, taxpayers) – is of particular interest here as it pertains directly to the increase in accountability testing in higher education. The Spellings report called for the creation of a public database containing institutional profiles with such information as cost, price, admissions data, college completion rates, as well as the results of actual students’ learning outcomes. The latter part, students’ learning outcomes, received special attention. Namely, the Spellings report (2006) strongly encouraged higher education institutions to measure educational value-added (i.e., growth over time) using quality assessment data, urged faculty to take an active role in defining educational objectives and creating instruments to assess student progress towards these goals, and make aggregate assessment results (as well as graduation rates, costs, etc.) publically available to all interested parties in a consumer-friendly form. The report also said the “the federal government should provide incentives for states, higher education associations, university systems, and institutions to develop interoperable outcomes-focused accountability systems” (p. 24). Furthermore, the role of accrediting bodies was redefined in the report, moving from the historically private relationship between an accrediting agency and an institution to a more public relationship with a strong emphasis on transparency and well-articulated conditions for accreditation. More specifically, the report called for accrediting bodies to prioritize assessment of the performance outcomes to enable national and international comparisons.

Not surprisingly, the release of the Spellings report in 2006 prompted a profound increase in systematic assessment of student learning and development outcomes in higher education, similar to the hortatory effect NCLB had in K-12. A notable similarity of these two legislative actions is the focus on objective measurement of student learning
outcomes. Furthermore, the emphasis on educational value-added in higher education is similar to the emphasis on adequate yearly progress in K-12, with both notions implying the need to measure student progress. However, in contrast to NCLB, the Spellings report neither requested states or institutions to specify annual achievement benchmarks, nor specified corrective actions that would follow the failure of meeting these goals. In addition, the nature of the accountability mandates, and thus the nature of stakes attached to the test scores, is different in higher education than it is in the K-12. Namely, the role of the federal government in higher education accountability movement is both limited and indirect, with no mention of federal responsibility for postsecondary education in the U.S. Constitution (Kuh & Ikenberry, 2009). The State governments, as units of organization and funding for higher education, provide direct and indirect (via financial support to students) funding to institutions, as well as registration and licensure (Ewell, 2009). However, most states tie financial support to graduation rates, as opposed to learning outcomes measures, with rare exceptions (Tennessee has an established performance funding scheme based on learning outcomes) (Ewell, 2009). Although NCLB listed specific federal and state-level accountability mandates for K-12 (e.g., use of statewide assessment data to meet AYP, corrective actions following failure to meet AYP), the Spellings report emphasized the role of accreditation as the primary vehicle of appraising the quality of education. As a result, regional accreditors assumed the role of an external stimulus in higher education, and institutional accreditation has been referred to as “an aggressive federal quality assurance tool” (Ewell, 2009, p. 14). According to Peter Ewell, the accrediting bodies’ “requirements regarding assessment are, thus, largely about process: to ensure that the institution has valid and vital mechanisms for
establishing and determining student achievement of key learning outcomes and to determine the extent to which the institution is using the resulting information to improve curricula and pedagogy” (p. 14). As such, higher education institutions often engage in compliance mode when it comes to institutional accountability, publically reporting their assessment efforts and thus meeting accountability demands placed by states and accrediting bodies, but not always using the data for institutional improvement. Overall, the stakes associated with accountability tests in higher education are lower than those in K-12 mainly because no federally imposed sanctions for student performance are in place. Nonetheless, higher education institutions do have to meet accountability demands or they risk losing accreditation, and, in some cases, financial support, thus making accountability tests fairly high-stakes for institutions. Beyond that, the use of assessment data, including the extent of student-level consequences attached to the test scores, is the prerogative of individual institutions (Erwin, 2002). So, what tends to be the stakes of accountability tests in higher education for students?

The report published in 2009 by the National Institute of Learning Outcomes Assessment (NILOA) titled More Than You Think, Less Than We Need: Learning Outcomes Assessment in American Higher Education summarizes the results of surveying administrators from regionally accredited, undergraduate-degree-granting institutions (n = 2,809) about their assessment practices. The findings revealed that the most common use of accountability test data is for institutional or program accreditation. In that study, the use of accountability test data for student-level decisions, such as transfer policies, was found to be very limited. Although this study did not examine the usage rate of accountability test data for student promotion or graduation, other evidence
indicates that such usage is very rare. For example, nationally normed standardized measures of general knowledge and skill such as Collegiate Learning Assessment (CLA), College Basic Academic Subjects Examination (College Base), Collegiate Assessment of Academic Proficiency (CAAP), and the Measure of Academic Proficiency and Progress (MAPP) are administered under low-stakes conditions, with no student-level consequences attached to performance. About two-fifths (39%) of all colleges and universities surveyed by NILOA in 2009 reported using such measures for institutional accountability (Kuh & Ikenberry, 2009). Institutions that choose to develop their own measures also tend to administer them in low-stakes conditions. For example, in one public state institution, students are mandated to attend a university-wide assessment testing session, but their performance on these assessments does not have any consequences; the scores are not reported on the academic transcript, are not factored into any grades, and no passing score is required for progressing to the next academic level. Although students are debriefed on the purposes of assessment, their scores are often not available to them at all (e.g., as in James Madison University). Hence, it appears that the accountability tests in higher education tend to be low-stakes for students.

**Implications of low stakes on test performance.** In summary, accountability tests administered both in K-12 and higher education contexts can be considered low-stakes for the majority of students. Even though these tests are low-stakes for students, they are high-stakes for other stakeholders, such as institutions and teachers. It is essential to ensure that scores from such low-stakes accountability tests are accurate representations of students’ knowledge, so that high-stakes decisions based on such scores (e.g., accreditation and state funding) are also accurate. If the stakes of a test
undermine students’ performance, then the scores are no longer accurate representations of students’ knowledge, and the validity of such test scores suffers. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) hold measurement specialists and test users responsible for collecting ample validity evidence supporting the uses of test scores. Therefore, investigating the implications of low-stakes on test performance is of utmost importance.

Research shows that, on average, students earn higher scores on tests that have personal consequences (i.e., high-stakes) than on tests that do not have personal consequences (i.e., low-stakes) (Cole & Osterlind, 2008; Huffman et al., 2011; Napoli & Raymond, 2004). Therefore, in aggregate, students’ scores on low-stakes tests are not indicative of their true ability or knowledge. As such, the issue of under-performance in low-stakes testing conditions becomes an issue of the validity of test scores. Although scores, on average, tend to be lower in low-stakes versus high-stakes contexts, it should be noted that, similar to high-stakes setting, there is variability in test scores in low-stakes settings (i.e., students score differently on the test even though they all take it under low-stakes conditions). For this reason, investigating variables that explain variability in low-stakes test scores, as well as ways to ameliorate the problem of underperformance, becomes a priority. Besides ability, the most straight-forward explanation for students’ variability in low-stakes test scores is test-taking motivation.

**Test-taking motivation in low-stakes settings.** Students completing low-stakes accountability tests may not be properly motivated to demonstrate what they know and are able to do, compromising the validity of test scores (Haladyna & Downing, 2004;
Liu, Bridgeman, & Adler, 2012). Given the importance of test-taking motivation for the validity of test scores, it is essential to consider how test-taking motivation is conceptualized, both theoretically and empirically.

**Theoretical understanding of test-taking motivation.** Wise and DeMars (2005) define test-taking effort as the extent to which students give their “best effort to the test, with the goal being to accurately represent what one knows and can do in the content area covered by the test” (p. 2). If the tests are low-stakes for students and no incentives are in place to perform to one’s best potential, the effort given to the test may not be a student’s “best”. As such, low-stakes tests demonstrate not what students know and can do in the content area covered by the test, but rather “what students will demonstrate with minimal effort” (O’Neil, Sugrue, & Baker, 1995/1996, p. 135). As noted by Eklof (2007), “ignoring the test-taking motivation component in low-stakes achievement testing could lead to a confounding of knowledge and motivation and thereby be a threat to the validity of the results” (p. 644).

Nonetheless, there is considerable variability in students’ test-taking effort in low-stakes contexts (Thelk et al., 2009). From a theoretical standpoint, the expectancy-value model of achievement offers a convenient framework for explaining variability in students’ test-taking effort. According to this theory, an individual’s behavior is a function of (a) expectancy (a person’s belief in his or her own ability to complete the task) and (b) value (one’s appraisal of the importance of a given task) (Pintrich, 1999). In turn, value is composed of the perceived importance of the task to an individual and the task’s perceived “cost”. Cost is defined as a perceived amount of effort required for the completion of the task. In the context of low-stakes settings, the majority of students are
likely to perceive the costs of trying on the tests as fairly high – the tests are long and mentally taxing. Value, on the other hand, is likely to be low for most students because performance does not have any direct consequences in low-stakes conditions. Therefore, theoretically, test scores are, on average, lower in low-stake settings than in high-stakes settings because students’ test-taking effort is, on average, lower. Test-taking effort is, on average, lower because value is lower in these contexts (i.e., lower perceived importance of test and higher cost). Empirical evidence further supports this theoretical claim that test-taking effort partially explains variability in performance on low-stakes tests.

**Empirical understanding of test-taking motivation.** A substantial body of empirical research provides ample evidence that unmotivated students score lower on low-stakes tests than motivated students, after controlling for ability (Eklof, 2010; Steedle, 2010; Sundre & Kitsantas, 2004; Wise & DeMars, 2005). Wise and DeMars (2005) summarized 12 empirical studies that examined the effects of test-taking motivation on low-stakes test performance, and concluded that, on average, unmotivated students perform one half of a standard deviation lower than motivated students. Importantly, research also indicates that there is a high degree of variability in college students’ motivation to perform well on low-stakes accountability tests (Lau & Pastor, 2007; Thelk et al., 2009; Zerpa et al., 2011), with some students even reporting investing maximal effort (Sundre & Kitsantas, 2004) and other students giving about 60-70% of their total effort on low-stakes tests (Cole & Bergin, 2005). Thelk and her colleagues (2009) reported considerable variability in both perceived importance of the test and test-taking effort for college first-year and sophomore samples. Given this variability, it is essential to examine potential reasons why students differ in their test-taking motivation.
in a low-stakes context. Similar to how motivation theory offers an explanation for lower performance in low-stakes context, other theories may explain variation in test-taking motivation. Prior to considering these theories in detail, it is essential to first discuss the measurement of test-taking motivation.

**Measuring test-taking motivation.** With respect to self-report measures of test-taking motivation, one popular measure is the 10-item Student Opinion Scale (SOS) consisting of two subscales: Effort and Importance (Sundre & Moore, 2002). The effort subscale measures the extent to which students put forth their best effort on the test, whereas the importance subscale assesses the extent to which students perceive the test as important. This measure has been extensively used at one American institution to assess student test-taking motivation on low-stakes accountability tests and is empirically supported with adequate validity and reliability evidence (Sundre & Moore, 2002; Sundre & DeMars, 2005; Thelk, 2006; Thelk et al., 2009). In the current program of research, the SOS will be used as a measure of students’ test-taking motivation.

**Strategies to increase test-taking motivation.** After presenting evidence of the effects of subpar test-taking motivation on low-stakes tests, as well as measures used to assess it, the next logical step is to consider strategies to improve students’ motivation. Although not explicitly stated in all studies aimed at boosting motivation, the implied premise of these studies is that a motivation-improving strategy will improve students’ attitudes. As such, the very nature of such studies suggests that attitudes affect motivation. Although attitudes were not measured in the studies described below, attempts to change them via various strategies have been examined.
Wise and DeMars (2005) provide a parsimonious yet comprehensive overview of motivation-improving strategies, including their benefits and limitations. These strategies include, but are not limited to material incentives, feedback, and interventions. The effects of incentives have been inconsistent and their use in higher education academic contexts is both impractical (e.g., choosing the appropriate incentive to motivate hundreds of students is quite difficult) and unbefitting in the academic context because it may be perceived as inappropriate incitement for students’ intellectual efforts. When it comes to providing individual feedback to a large number of students, it may not be appropriate to do so using the measures created for program assessment purposes because such measures often don’t have the necessary psychometric properties needed for reporting and interpreting individual scores (e.g., internal consistency). Plus, the positive effects of informational feedback on performance are not well-evidenced (Baumert & Demmrich, 2001).

Other studies investigating strategies to improve test-taking motivation focused on students’ perceptions of assessment and incorporated motivational interventions created to improve students’ attitudes toward accountability testing and in turn boost test-taking motivation. A recent study conducted by Brown (2010) explored such motivational strategies as highlighting the value of the test and appealing to the students’ willingness to improve their institutions. In a study conducted by Zilberberg and her colleagues (2009), the effects of various ways of framing the purpose of accountability testing to college students were investigated using focus groups. This qualitative inquiry revealed that framing assessment as a way to provide service to the university is most likely to resonate with students and positively impact their attitudes. However, attitudes
were not measured in that study but only inferred. In yet another study, Huffman et al. (2011) found that college students who viewed a motivational PowerPoint Presentation geared toward increasing a positive attitude toward tests had significantly higher scores and stronger score-to-GPA correlations than students who received monetary incentives. In yet another study, Swing (2001) succeeded at bolstering students’ test-taking motivation on low-stakes tests by making a public announcement highlighting assessment activity as campus pride, civic responsibility, and a chance to shape institutional policies. This line of research suggests that one approach for combatting poor test-taking motivation in low-stakes contexts is through changing perceptions or affect towards tests that students are bringing into the testing session. Unfortunately, none of the studies reviewed measured students’ attitudes or change in them directly, further signaling the need for an empirical investigation of this construct.

**Students’ Attitudes towards Institutional Accountability Testing**

Prior to considering students’ attitudes toward accountability tests, it is important to frame the construct of attitudes in general. The social psychology definition of attitude states that it is a “favorable or unfavorable evaluative reaction toward something or someone, exhibited in one's beliefs, feelings, or intended behavior” (Myers, 2002). Unlike personality traits, attitudes are malleable and can be changed over time (Kobella, 1989). Social psychologists hold various views regarding the processes underlying the formation of attitudes. For example, persuasion theories suggest communication plays a key role in attitude formation (Petty and Cacioppo, 1986), whereas a frame factor theory purports a variety of shaping forces in one’s learning environment (e.g., prior experience, external influences) are responsible for attitude formation (Lundgren, 1972). Irrespective
of the theoretical reasons as to what is causing attitude formation, researchers tend to agree that investigating attitude change in students is vital for understanding students’ motivation for academic success and engagement in academic-related activities (e.g., Anderson, Teisl, Criner, Tisher, Smith, Hunter, Norton, Jellison, Alyokhin, Gallandt, Haggard, & Bicknell, 2007; Kobella, 1989; Sundre, Barry, Gynnild, & Ostgard, 2012). Moreover, evidence indicates that attitudes are predictive of students’ academic behavior. For example, a meta-analysis of the effects of study attitude (defined as a positive attitude toward education in general and studying in particular), study habits, and prior performance on college students’ academic performance revealed that study attitude is a strong predictor of performance, even after controlling for the effects of other variables (Crede & Kuncel, 2008).

With respect to the role of attitudes on the validity of low-stakes tests scores, several researchers called for such investigations. For example, Sundre and Kitsantas (2004) noted the need to understand the “psychology of the examinee” in the context of low-stakes accountability testing. From their perspective, three components influence test-taking motivation: expectancy for success, value of the test, and affect toward the test. The role of student affective feelings, or attitudes, on test-taking motivation is not well-understood, despite their potential importance. As emphasized by Weekers, Brown, and Veldkamp (2009), “The role of personal beliefs and attitudes in determining test scores is less well understood” than various other factors influencing test scores (p. 133). The research on the role of attitudes in low-stakes accountability testing contexts in higher education is nascent at best (Zilberberg et al., in press; Zilberberg et al., in progress).
Although the research related to students’ attitudes towards low-stakes accountability testing is scarce, evidence on students’ views on assessment in general is very rich. Students’ attitudes toward institutional accountability testing are best understood in light of the current understanding of students’ attitudes toward assessment in general. For this reason, this literature is considered next in more detail.

**College students’ attitudes toward assessment in general.** In their comprehensive review of students’ perceptions about evaluation and assessment in higher education, Struyven, Dochy and Janssens (2005) summarize research studies conducted in this domain over several decades. The literature search that serves as a foundation to the Struyven et al. (2005) review employed the following keywords: “student perception”, “assessment”, and “higher education”. Not surprisingly, the search resulted in a collection of articles covering a broad range of assessments used in college classrooms, including traditional multiple-choice, constructed-response, and portfolio assessments. Although none of the articles reviewed by Struyven et al. (2005) pertained to institutional accountability testing specifically, the findings on college students’ perceptions and attitudes toward assessment in general inform researchers’ hypotheses regarding students’ attitudes toward accountability testing. Overall, they found that students favor multiple-choice format exams, but question the fairness of such tests. Next, we take a closer look at some of the studies reviewed by Struyven et al. (2005), as well as other studies pertinent to students’ attitudes toward institutional accountability testing.

A study by Sambell, McDowell, and Brown (1997) titled “‘But is it fair?’: An exploratory study of student perceptions of the consequential validity of assessment”
explored undergraduate students’ preferences and attitudes pertaining to assessment in a variety of subject matters and of various response formats (e.g., multiple-choice, essays). This team of researchers found that students prefer assessments that they perceive to be *fair*, and by that they mean - (1) relates to authentic tasks; (2) represents reasonable demands; (3) encourages students to apply knowledge to realistic contexts; (4) emphasizes the need to develop a range of skills; and (5) is perceived to have long-term benefits. The same study found students have more negative perceptions of tests over which they have little control (i.e., the tests are “done to them”). Other evidence indicates students dislike tests that do not result in any feedback on performance, suggesting evaluation plays a major role. For example, Janssens, Boes, and Wante (2001) found when students were not graded on their portfolio assessments, their engagement levels went down and much less effort was invested in the task as compared to when they received a grade. This finding is congruent with previous research indicating that students under-perform in low-stakes testing conditions (e.g., Napoli & Raymond, 2004) and also implies that students are more skeptical of the value of such assessments, further necessitating the need to study students’ attitudes toward low-stakes testing. In addition, Janssens et al. (2001) also found that students dislike “single-shot” evaluations such as end-of-course exams. Pertaining to standardized achievement tests (without distinguishing the difference between the stakes), some evidence indicates that as students get older, they become more disillusioned about such tests, in particular as measures of their worth (Paris, Roth, & Turner, 2000). Specific to the current study, Huffman et al. (2011) reported that college students resent the fact that they are mandated
to spend time and effort on an institutional accountability test with no benefit to them personally.

What about response format; does it matter whether the assessment is multiple-choice or constructed-response? Interestingly, some evidence indicates that even though the essay format is viewed by students as a more fair assessment and is preferred by those with higher academic self-efficacy (Birenbaum & Feldman, 1998), the majority prefer multiple-choice response format because such tests are not as cognitively taxing as constructed response formats. More specifically, multiple-choice tests are preferred on the following dimensions: perceived difficulty, anxiety, complexity, success expectancy, and feeling at ease. Such tests are perceived by students as easier to prepare for and yield higher overall scores. However, students also feel that multiple-choice tests encourage surface learning only and overemphasize recollection of isolated details and facts (Struyven et al., 2005). DeMars (2010) investigated the interaction between response format (multiple-choice and constructed-response) and stakes of test (low-stakes and high-stakes) using a high school student sample. She found that increasing the stakes of the test had a stronger effect on the constructed-response test scores than multiple-choice test scores. DeMars concluded that studying perceptions and attitudes toward tests of different response formats is worthwhile, presumably because such attitudes may explain the interaction between response format and stakes.

What can we infer about students’ attitudes toward accountability testing? In light of the research on students’ attitudes towards assessment in general, what can we infer about students’ attitudes toward tests used for institutional accountability purposes in particular? Such mandatory accountability tests tend to be standardized, multiple-
choice, often resulting in no individual feedback, are “done to students”, and are often
given as single-shot evaluations. Recall that research on students’ perceptions of
assessment in general indicates that these are exactly the types of tests likely to elicit
negative affect from students.

Furthermore, the population of interest here is college students, meaning that they
are young adults who tend to have disillusioned views regarding standardized
achievement testing of all kinds (Paris et al., 2000). Having grown up under the test-
based accountability system of NCLB, these young adults are accustomed to taking low-
stakes accountability tests, but that does not mean that they appreciate them. On the
contrary, it is likely that by the time students reach college, they have developed strong
negative attitudes toward such tests in response to the skeptical attitudes towards K-12
accountability tests emanating from those in authority. Such authority figures (i.e.,
parents and teachers) tend to question the validity of K-12 accountability tests and
perceive them as detrimental to classroom climate (Abrams, Pedulla, & Madaus, 2003).

It follows that college students are likely to perceive accountability tests as not fair, view them as a waste of time, and thus have negative attitudes towards such tests.

Whether such negative student attitudes actually exist and how they affect test-taking
motivation and performance on accountability testing in higher education are empirical
questions that can only be examined if adequate measures of the construct exist. Next, the
current state of measurement of students’ attitudes toward testing is reviewed.

**Measurement of Students’ Attitudes toward Testing**

Two instruments exist that measure American college students’ attitudes toward
accountability testing – one for the K-12 context (SAIAT-K12) and one for the higher
education context (SAIAT-HE) (Zilberberg et al., in press). For the purposes of the current research, these measures and the substantive research pursuits enabled by them are of primary interest. However, other measures created to represent similar constructs warrant discussion. Thus, the subsequent sections describe these other measures, followed by the discussion of the psychometric properties of the SAIAT-K12 and SAIAT-HE and culminating with the discussion of why the SAIAT-K12 and the SAIAT-HE are deemed as the most appropriate for current purposes.

**Students’ perceptions of standardized tests.** Paris and his colleagues (1991; 2000) created an instrument, for use with the K-12 student population (grades 4, 7, 10), measuring students’ attitudes toward both classroom tests and standardized achievement tests. Intended to “measure developmental changes in students’ perception of standardized achievement tests” (Paris et al., 2000, p. 18), the instrument included a range of constructs, including awareness of test strategies, test anxiety, and perceptions of test validity and reliability. In regards to the standardized achievement tests, the authors referred primarily to tests used for institutional accountability purposes, such as the California Achievement Test (CAT), Iowa Test of Basic Skills (ITBS), and Michigan Educational Assessment Program (MEAP); however, these tests also had some direct consequences for students in Paris’s studies.

Investigation of the psychometric qualities of this instrument revealed less-than-desirable results. The subscales yielded low internal consistency (ranging from 0 to .51) and only two subscales were supported by the results of a principal component analysis: *Perceived Value of the Test* (e.g., “My test scores show that I am a good student”) and *Students’ Affect about the Test* (e.g., “I feel good about my test scores”). The measure
created by Paris’s research team cannot be used for investigating college students’
attitudes towards accountability testing because the construct defined by Paris et al. is
more broad (i.e., all standardized achievement testing), the population studied by Paris et
al. differs from the one of interest here, and the psychometric properties of the measure
are less than desirable. Nonetheless, the research conducted by Paris’s team greatly
informed the current line of research.

Students’ Conceptions of Assessment inventory (SCoA-VI). The Students’
Conceptions of Assessment inventory currently in its sixth version (SCoA-VI) was
developed by a team of researchers in New Zealand (Weekers, Brown, & Veldkamp,
2009) to assess conceptions of assessment among high school students in New Zealand.
Designed to measure four inter-correlated constructs (“Assessment Improves Learning
and Teaching [Improvement]”, “Assessment Relates to External Factors [External]”,
“Assessment has Affective Benefit [Affect]”, and “Assessment is Irrelevant”
[Irrelevance]) the SCoA-VI refers to assessment broadly, without specifying the type
(e.g., low-stakes or high-stakes).

The internal structure of the measure has been validated with high school students
in New Zealand, and both high school and college students in various other countries
(Brown, 2011). Moreover, it has been used at one American institution during a
university-wide low-stakes assessment session (Wise & Cotton, 2009). The latter study
revealed that less guessing behavior on tests (i.e., longer response time) is associated with
greater belief that assessment leads to improvement, whereas more guessing (i.e., shorter
response time) is associated with lower affective benefit and greater belief in irrelevance
of assessment. Further, those students who scored higher on “Improvement” and lower on
“Irrelevance” were more likely to attend the scheduled low-stakes mandatory assessment session, as opposed to a make-up testing session.

Even though the SCoA-VI had been successfully used to measure American college students’ *conceptions of assessment*, this construct differs somewhat from *students’ attitudes towards accountability testing in higher education* (the attitudes of interest in the current line of research). That is, the theoretical assumption is that the construct of interest herein - *students’ attitudes towards accountability testing in higher education* - is qualitatively distinct from its theoretical relatives such as conceptions of assessment and perceptions of standardized achievement testing, and thus warrants its own measure. Next, the psychometric properties of the measures created to assess this construct specifically are outlined.

**Psychometric properties of the SAIAT-K-12 and the SAIAT-HE**

As reviewed above, several instruments, although helpful for understanding students’ attitudes toward assessment in general, are not suitable for measuring the specific construct of *college students’ attitudes toward institutional accountability testing* because these instruments are either too broad in scope (e.g., SCoA-VI; Brown, 2006), are designed for a younger population, or lack adequate psychometric study (e.g., Paris et al., 2000). Furthermore, the construct of *students’ attitudes toward institutional accountability testing* might differ depending on the context –K-12 versus higher education – signaling a need to engage in the instrument development process separately for each context in order to assess their similarity or distinctiveness. To address this need, a team of higher education assessment professionals developed two new measures: one designed for measuring college students’ attitudes toward institutional accountability
tests administered in K-12 and another one designed for measuring college students’ attitudes toward institutional accountability tests administered in higher education (Zilberberg, Anderson, Finney, & Marsh, in press). The development and properties of these measures are outlined below.

**SAIAT-K-12.** A measure of students’ attitudes toward institutional accountability tests administered in K-12 (SAIAT-K12) was developed by a team of subject matter experts (i.e., higher education assessment specialists well-versed in the practice and theory of large-scale low-stakes testing, examinee test-taking motivation, and scale development). Coupling existing theories of students’ attitudes toward assessment with their professional experience in large-scale low-stakes testing, the test developers identified the following five dimensions of student attitudes toward institutional accountability tests: (1) Validity, (2) Use, (3) Purpose, (4) Disillusionment, and (5) Parents. Validity was defined as students’ perceptions of the accuracy of institutional accountability tests administered in a K-12 context. Use was defined as students’ perceptions of the extent to which accountability tests were used for identifying areas of improvement. Purpose was defined as students’ perceived understanding of the accountability tests’ purpose. Disillusionment was defined as students’ dissatisfaction with accountability tests. Parents, an “influence” dimension that reflected the possibility that external parties may influence students’ attitudes, was defined as the students’ perceptions of the extent to which their parents paid attention to the K-12 accountability test results.

Items representing these dimensions were created by adapting some items from other measures (e.g., Paris et al., 2000; Weekers, Brown, & Veldkamp, 2009; Wong &
Paris, 2000) and creating new items to cover the breadth of five proposed dimensions. The SAIAT-K12 was administered to two large college student samples: incoming first-year students and mid-career students (i.e., sophomores and juniors who earned between 45 and 70 academic credits). Both exploratory (EFA) and confirmatory factor analyses (CFA) were conducted to investigate the factor structure of the SAIAT-K12. Based on the results of these analyses, the Use factor was omitted altogether and several items from other factors were omitted due to low factor pattern coefficients or cross-loadings. With respect to the Use items, the majority of which focused on rating how others (e.g., teachers) used test results, students tended to select “neither agree nor disagree”, suggesting they did not know how test scores were used by others. Thus, most of the Use items had low pattern coefficients. With respect to the other items that were omitted, some were ambiguously worded and others were too extreme in wording, causing cross-loadings and low pattern coefficients.

After removing these poorly performing items, results from both samples confirmed a correlated four-factor model, with the majority of items having strong relationships with their respective factors. One item from the Purpose factor correlated weakly with the other items (“Someone (teacher, principal, guidance counselor) explained to me why I take standardized tests”) in both samples. Given the responses to this item were approximately normally distributed in both samples, the lack of correlation was not due to floor or ceiling effects. If this item continues to perform poorly in the subsequent studies, it will be removed from the scale. The current version of the SAIAT-K12 consists of sixteen items representing four factors: (1) Validity, (2) Purpose, (3) Disillusionment, and (4) Parents (see Appendix B).
SAIAT-HE. Building upon the SAIAT-K12, a similar instrument designed to measure students’ attitudes toward institutional accountability tests administered in higher education (Zilberberg et al., in press). The following six inter-related dimensions were specified: (1) Validity, (2) Purpose, (3) Disillusionment, (4) Parents, (5) Professor, and (6) Students. The first four dimensions are defined as they were for the SAIAT-K12 measure. In addition to the Parents dimension, two new “influence” dimensions (Professor and Students) were added to reflect the notion that college students may have other external influences affecting their attitudes, namely, other students and professors.

Besides the two additional “influence” dimensions, the SAIAT-HE differed from the SAIAT-K12 in the following ways. First, the instructions preceding the measure prompted students to think specifically of accountability assessments completed throughout the course of college, with examples of these types of tests (in contrast to the SAIAT-K12 instructions which gave examples of accountability tests in K-12). Second, the SAIAT-HE items referred to tests as “assessment tests”, whereas the SAIAT-K12 items referred to tests as “standardized tests”. This change was intended to prompt students to think of higher education accountability testing because such tests are often referred to as “assessments” by both students and faculty.

The SAIAT-HE was administered to a sample of mid-career students, with the sample being independent from that used to study the SAIAT-K12. CFA was used to test the correlated six-factor structure. Overall, the model fit the data moderately well. However, all the external influence factors had low reliability indices, calling for a more careful examination of these factors. The reliability of the Parents factor was low (ω = .56), due to moderate to low item-level $R^2$ values. For example, the item “My parents are
unaware of the assessment tests I complete at the University” was unrelated ($R^2 = .06$) to
the Parents factor, with 57.4% of students strongly or moderately agreeing with this item.
The item “My parents would be proud of me if I performed well on the assessment tests”
was not a strong indicator of the Parents factor ($R^2 = .32$). In contrast, this item had a
stronger saliency to the corresponding Parents factor on the SAIAT-K12 ($R^2 = .50$). This
finding is not surprising given parents may not be aware of accountability tests
administered in college.

The reliability of factors Professors ($\omega = .44$) and Students ($\omega = .50$) was also
very low, thus raising doubts about their utility. Two of the three Professor items had
extremely low $R^2$ values, which was most likely due to lack of variability (floor effects).
Specifically, the majority of students (81%) strongly or moderately disagreed with the
item “My professor(s), an academic advisor, or resident advisor encouraged me to
prepare for the assessment tests”. Notably, the item “My professors don’t value the
assessment tests I complete at the University” was associated with 45% of the students
selecting “Neither agree nor disagree”, implying students are not certain of their
professors’ views of assessment. Two of the items on the Students subscale elicited
extreme responses. The majority of students (76%) moderately or strongly disagreed with
the item “Fellow students urged me to try my best on the assessment tests” and the
majority (78%) slightly, moderately, or strongly agreed with the item “Fellow students
speak negatively about the assessments tests at the University”. Low factor pattern
coefficients and reliability associated with these three influence factors are a cause for
concern. If the three factors continue to perform poorly, they will be omitted from the
scale.
In addition, a few items from other factors were flagged as poorly functioning. The item “Assessment tests are unfair to some students” had $R^2$ value of .15, thus contributing little to the Validity factor. Similar to the SAIA-T-K12, the item “Someone (professor, academic advisor, Resident Advisor) explained to me why I take assessment tests” also had a low $R^2$ value of .10, thus contributing little to the Purpose factor.

In summary, the initial study of the psychometric properties of the SAIAAT-HE revealed further investigation of the internal structure is necessary before the measure can be used in research. It follows that garnering more validity evidence for the SAIAAT-HE is necessary. To this end, Stage I of the current program of research focuses on the evaluation of psychometric properties and revisions of the SAIAAT-HE; this was described in detail in Chapter 1.

To date, no other measure exists that breaks down the construct of students’ attitudes toward accountability testing in higher education in the same way as the SAIAAT-HE. Given adequate psychometric properties, the SAIAAT-HE will be used to answer substantively important questions regarding students attitudes, as detailed in Chapter 1 (Stage II of the current program of research).
Methods

Data for the current study were collected at a mid-sized southeastern university that uses large-scale low-stakes accountability assessments for programmatic improvement and meeting external mandates. For clarity, descriptions of the measures are presented first, followed by the description of the samples that completed them and data collection procedures. A description of the data analyses concludes this section.

Measures

Students’ Attitudes toward Institutional Accountability Testing –Higher Education (SAIAT-HE: Zilberberg et al., in press). The SAIAT-HE is a self-report measure developed based on the SAIAT-K-12. In its current form, the SAIAT-HE consists of 22 items and measures the following six dimensions: Validity, Purpose, Disillusionment, Parents, Students, and Professors. The first psychometric evaluation of the SAIAT-HE provided preliminary evidence for the factorial structure of the scale and identified areas for improvement (Zilberberg et al., in press). Students responded to a series of statements using a Likert scale ranging from 1 (“Strongly Disagree”) to 7 (“Strongly Agree”).

Stage I (RQs 1, 2, and 3) of the current program of research focused on gathering validity and reliability evidence for the SAIAT-HE. Upon completion of Stage I, the structure of the scale was modified as a result of these analyses. The revised version of the SAIAT-HE was used to address several substantive research questions pertaining to students’ attitudes toward higher education accountability testing (RQs 4, 5 (Models 1, 2, 3), RQ6). The SAIAT-HE can be found in Appendix A.

Students’ Attitudes toward Institutional Accountability Testing-K-12 (SAIAT-K-12: Zilberberg et al., in press). The SAIAT-K-12 is a newly developed self-
report measure of students’ attitudes toward institutional accountability testing administered in K-12. In its current form, the measure consists of 16 items and measures the following four dimensions: Validity, Purpose, Disillusionment, and Parents. Students responded to a series of statements using a Likert scale ranging from 1 (“Strongly Disagree”) to 7 (“Strongly Agree”). The first study reporting the psychometric properties and the use of the SAIAT-K-12 offers promising, yet preliminary, evidence in support of validity and reliability of the SAIAT-K-12 (Zilberberg et al., in press). Recall that data from this measure was collected to investigate the strength of the relationship between first-year students’ attitudes toward K-12 accountability tests and their attitudes toward higher education accountability tests (RQ4) as well as the causal chain between first-year students’ attitudes toward K-12 accountability tests and their performance on college accountability tests (RQ5 (Model 3)). The SAIAT-K-12 can be found in Appendix B.

Global Experience (GLEX: James Madison University). The GLEX is a multiple-choice measure created by James Madison University faculty for assessment of general education learning outcomes focused on critical thinking regarding students’ own society and its relationship to the larger global community. More specifically, GLEX measures students’ understanding of basic global problems, global political, social, cultural and economic systems, issues involved in analyzing societies different from their own, the global forces that shape societies, theoretical models used in studying global problems, and the strengths and limitations of alternative solutions to global problems across and within cultures (DeMars, 2012). The GLEX was chosen because it is a measure used to assess general education outcomes, it is psychometrically sound, and a substantial number of students who took the SAIAT-HE also took the GLEX.
The GLEX test consists of 32 items; total score is interpreted as an overall measure of global experience. For first-year students completing the GLEX between Fall 2010 and Fall 2012, the mean total score ranged between 20.34 with the standard deviation of 4.88 and 21.06 with the standard deviation of 4.76. These descriptive statistics indicate that GLEX tends to be fairly difficult for first-year students, but also indicates the presence of moderate variability in the GLEX scores. The most recent administration of GLEX to first-year students revealed internal consistency of the scale to be .74, which is acceptable for assessment purposes (DeMars, 2012). Item-total correlation exceeded .20 for 24 out of 32 GLEX items in the first-year student sample.

In the most current mid-career student sample (i.e., students who completed the relevant general education requirement), GLEX had the following properties. The mean score was 23.01 with the standard deviation of 4.90. Again, these descriptive statistics indicate that the GLEX tends to be fairly difficult for mid-career students (but less so than it is for first-year students) and that there is moderate variability in the GLEX scores. The internal consistency of the scale was .80. Item-total correlation exceeded .20 for 28 out of 32 GLEX items in the mid-career student sample. GLEX was moderately correlated with grades from relevant courses, which adds validity evidence to the inferences drawn from its scores (DeMars, 2012).

In the current program of research, the GLEX score was used as the performance measure for all three path models subsumed under RQ5 (i.e., What is the nature of the relationship between students’ attitudes, test-taking motivation, and performance on higher education accountability test?). The goal of this research question is to explain variability in the GLEX scores for both first-year and mid-career students, thus the
presence of variability in GLEX scores (i.e., lack of floor/ceiling effects) in both samples is desirable.

**The Student Opinion Scale (SOS: Thelk, Sundre, Horst, & Finney, 2009; Sundre & Moore, 2002).** The SOS is a self-report measure of test-taking motivation consisting of 10 items and measuring two dimensions of test-taking motivation: perceived importance of the test (e.g., “Doing well on these tests was important to me”) and test-taking effort (e.g., “I engaged in good effort throughout these tests”). The SOS instructions directed students to think about all the tests they completed during the assessment session (see Appendix C). Students responded to a series of statements using a Likert scale ranging from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”). The scale has been extensively used for measuring college students’ test-taking motivation in low-stakes settings, providing ample support for the validity and reliability of the scale (e.g., Thelk et al., 2009). The SOS was used to investigate the nature of the relationship between students’ attitudes, test-taking motivation, and performance on a higher education accountability test (RQ5 (Models 1, 2, 3)). The SOS can be found in Appendix C.

**SAT-math and SAT-verbal.** SAT-math was included as a control for quantitative ability, with possible scores ranging from 200 to 800. SAT-verbal was included as a control for verbal ability, with possible scores ranging from 200 to 800. Recall that a research question addressing the nature of the relationship between students’ attitudes, test-taking motivation, and performance on higher education accountability test (RQ5 (Models 1, 2, 3)) examines the direct and indirect effects of attitudes on GLEX test scores after controlling for prior ability. Both measures were included given that the
GLEX requires interpretation of graphs and numerical information as well as advanced reading and text comprehension.

**Big Five Inventory – Conscientiousness (John & Srivastava, 1999).** The Big Five Inventory is a widely used measure of personality traits consisting of 44 items and measuring the following five personality aspects: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Students responded to a series of statements using a Likert scale ranging from 1 (“Disagree Strongly”) to 5 (“Agree Strongly”). Evidence exists in support of the factorial structure and reliability of the measure (John & Srivastava, 1999).

For the current study, only Conscientiousness (characterized by traits such as orderly, responsible, and dependable) was utilized in order to assess the differences in attitudes among students who attend a scheduled assessment session (i.e., compliant students) and those who attend a make-up session (i.e., non-compliant students) (RQ6). The Big Five Conscientiousness subscale can be found in Appendix D.

**Participants and Procedures**

This section describes different samples and corresponding data collection methods. Please see Table 2 and Table 3 for an overview of the alignment between samples, measures, and research questions. Demographic information for all student samples are presented in Table 4 and summarized below.

**First-year students.** Upon arriving on campus, all first-year students at this university have to participate in their first higher education accountability assessment session - a university-wide mandatory “Assessment Day”, during which students complete a mixture of cognitive and attitudinal accountability tests. Prior to starting the
tests, students are presented with an instructional video which clearly explains the purpose and importance of accountability testing in higher education and provides detailed instructions on how to fill out the testing scantrons. In addition, trained proctors reiterate the same message in person and are available throughout the testing session to answer students’ questions. Although the tests administered on Assessment Day are high-stakes for the university because the data are used to inform programmatic changes and meet accountability demands, they are low-stakes for students because no important individual decisions hinge on these test scores. Furthermore, students do not typically have access to their test results, with some exceptions. As such, these accountability tests are non-consequential for students. Although proctors emphasize the importance of such tests and participation is mandatory, in the end students choose how much effort to invest in each test. The same test administration procedure is used for first-year and mid-career students.

Students were randomly assigned to different test configurations, so not all students take the same tests. The Students’ Attitudes towards Institutional Accountability Tests in Higher Education (SAIAT-HE) measure was included on one of the test configurations on Assessment Day. First-year students who took the SAIAT-HE also took several other measures employed in the current program of research: SOS, SAT, and GLEX.

Initially, 894 responses were obtained on the measures listed above from first-year students. After screening the data (i.e., omitting outliers and nonsensical responses) and selecting students who had complete data on all the measures listed above, the total sample size was reduced to 857. Of these students, 62.66% were female, 85.76% were
Caucasian. This sample was used to answer RQ1, 2, 3 (i.e., psychometric properties of the SAIAT-HE) and RQ5 (Models 1, 2) (i.e., hypothesized relationships between these variables).

**Matched incoming and first-year student sample.** Prior to starting their first-year orientation at the university, all newly accepted incoming students have to prepare for their first college semester by completing a set of tasks pertaining to residence life, academics, and computing – all outlined in the manual called One Book (James Madison University, 2012-2013). Students are highly encouraged to complete all the steps outlined in One Book - they are given deadlines for each task and receive several reminders as the deadlines approach, which results in a high completion rate. One of the surveys that students have to complete over the summer as part of One Book is the “First Year Student Survey” which includes demographic questions and a set of attitudinal measures that inquire about students’ “high-school experiences, personal values and beliefs and decision to attend college at JMU” (JMU, 2012-2013). The Students’ Attitudes towards Institutional Accountability Tests in K-12 (SAIAT-K-12) measure was included on the First-Year Survey. Initially, 4338 responses were obtained on the SAIAT-K-12 from incoming students.

Out of the sample of the incoming students described above, a subset also completed the SAIAT-HE measure on Assessment Day, prior to starting their college career. Having the SAIAT-K-12 and the SAIAT-HE scores from *the same individuals* allows us to investigate the relationships among students’ attitudes toward K-12 accountability tests and their attitudes toward such tests encountered in college (RQ4). In a similar vein, having complete data on SAIAT-K-12, SAIAT-HE, SOS, SAT, and
GLEX from the same individuals allows us to investigate the influence of attitudes towards K-12 accountability tests on the performance on college accountability tests (i.e., GLEX) (RQ5 (Model 3)). To this end, first-year students’ records from Assessment Day were matched with their records from the First-Year Survey, resulting in the matched incoming and first-year student sample.

Initially, this matching resulted in 746 responses; after screening the data (i.e., omitting outliers and nonsensical responses) and selecting students who had complete data on all the measures listed above, the total sample size was 678. Of these students, 64.60% were female, 78.76% were Caucasian.

Mid-career Assessment Day students. Students at this university are mandated to participate in two “Assessment Day” sessions: once as first-year students (as described above) and the second time after they have accumulated between 45 and 70 academic credits, and thus can be considered mid-career students. Notably, mid-career students completed this round of Assessment Day tests under the same low-stakes conditions as they did the first time they took these tests as first-year students. An independent sample of mid-career Assessment Day students, unrelated to the matched incoming and first-year student sample described above, was used to answer several research questions. A total of 2765 mid-career Assessment Day students had complete data on the SAIAT-HE, and some of the students had complete data on other measures. This sample was broken down strategically in order to address different research questions. This break-down is explained next.

Mid-career Assessment Day sample 1. A total subset of 758 mid-career Assessment Day students had complete data on SAIAT-HE, SOS, SAT, and GLEX and
all 758 were retained after screening the data (i.e., omitting outliers and nonsensical responses). Of these students, 66.09% were female, 72.82% were Caucasian. This sample was used to answer a research question pertaining to the relationship between mid-career students’ attitudes toward accountability tests in higher education, test-taking motivation, and performance (RQ5 (Models 1, 2)).

**Mid-career Assessment Day sample 2.** A total subset of 300 mid-career Assessment Day students had complete data on SAIAT-HE and Big Five Conscientiousness subscale. Of these 300 students, all 300 were retained after screening the data (i.e., omitting outliers and nonsensical responses). Of these students, 60.67% were female, 74% were Caucasian. This sample was used to answer a research question about the differences in attitudinal profiles between Assessment Day attendees and make-up session attendees, after controlling for conscientiousness (RQ6).

**Mid-career calibration and validation samples.** A total subset of 1,707 mid-career Assessment Day students had complete data on SAIAT-HE and all 1,707 were retained after screening the data (i.e., omitting outliers and nonsensical responses). This sample was then divided into two: Mid-career Calibration sample ($N = 857$) and Mid-career Validation sample ($N = 850$). Of 857 students in the Mid-career Calibration sample, 62.54% were female and 74.33% were Caucasian. This sample was used for initial psychometric analyses of the SAIAT-HE (RQs 1, 2, 3). Of 850 students in the Mid-career Validation sample, 58.82% were female and 73.29% were Caucasian. This sample was used for replicating initial psychometric analyses.

**Mid-career make-up sample.** Some mid-career students chose not to attend their scheduled “Assessment Day” session and instead attended a make-up testing session. A
total of 169 mid-career make-up attendees had complete data on the SAIAT-HE and Big Five Conscientiousness, and all 169 were retained after screening the data (i.e., omitting outliers and nonsensical responses). Of these 169, 37.87% were female, 67.46% were Caucasian. This sample was used to answer a research question about the differences in attitudinal profiles between Assessment Day attendees and make-up session attendees, after controlling for Conscientiousness (RQ6).

Data Analyses

The current program of research consists of two stages: Stage I involves collecting validity evidence for the SAIAT-HE and Stage II involves addressing substantive research questions about students’ attitudes toward higher education accountability testing using the SAIAT-HE-revised. Please see Table 2 and Table 3 for an overview of the alignment between samples, measures, and research questions. Below, the data screening procedures are outlined followed by the proposed data analysis for each research question.

Data screening and assumptions. For all samples, raw data were available. Thus, SPSS 18.0 and PRELIS were used to screen the data. Covariance matrices were produced and submitted to LISREL 8.80 (Jöreskog & Sörbom, 2006) for any analyses involving structural equation modeling (RQs 1, 2, 3, 5). Covariance matrices, as opposed to correlation matrices, were used as input data for structural equation modeling analyses because estimation methods assume the use of unstandardized variables (Tanaka, Panter, Winborne, & Huba, 1990). Prior to conducting any analyses (structural equation modeling or otherwise), the following data screening procedures were conducted.
First, data were screened for univariate and multivariate outliers. Data on each individual variable were examined for extreme scores in order to identify univariate outliers. In order to identify multivariate outliers, a regression procedure was used to obtain the Mahalanobis distance (the distance of a case from the centroid in a multivariate space). A break in the list of the top ten Mahalanobis distance values was used to detect multivariate outliers. Values identified as both univariate and multivariate outliers were examined closely to assess if their responses were anomalies. “Outliers” were excluded from the analyses if the responses followed a nonsensical pattern (e.g., 1, 2, 3, 4, 5, 6, 7) or a response set (e.g., 2, 2, 6, 2, 2, 6, 2).

Second, data were screened for both univariate and multivariate normality because severe non-normality can affect the accuracy of Maximum Likelihood (ML) estimation (Finney & DiStefano, 2006). The following absolute values of univariate skewness and kurtosis were used to detect severe non-normality: above or below |3| for skew and above or below |10| for kurtosis (Kline, 1998). Histograms, bivariate scatterplots, and residual plots were also examined for potential outliers, shape of the distributions, linear relationship among observed variables, and homoscedasticity. In order to screen for multivariate non-normality, Mardia’s normalized multivariate kurtosis values larger than 10 was used as a criteria. In situations when data were found to be severely non-normal, the Satorra-Bentler adjustment was applied to fit indices and standard errors (Finney & DiStefano, 2006).

Extreme multicollinearity may be an issue potentially leading to a non-positive definite matrix in situations when variables are highly correlated (above .85) (Tabachnick & Fidell, 2001). In order to screen for multicollinearity, bivariate correlations among
observed variables were examined. Specifically, SAIAT-HE items were considered during the psychometric analyses subsumed under Stage I (RQs 1, 2, 3). The SAIAT-HE-revised subscales were considered during addressing substantive research questions subsumed under Stage II (RQs 3, 5, 6). Subscales of other measures (SOS, SAT, GLEX, Conscientiousness) were screened for multicollinearity in the analyses they were used in (see Table 2). None of the bivariate correlations were expected to be stronger than $r = .85$. In order to check for multivariate multicollinearity, a regression procedure in SPSS was used. Tolerance levels (percent of variance not explained by other variables) for each measured variable were reported, with the desired values larger than .10.

For analyses that involved assessing item functioning (RQs 1, 2, 3), descriptive statistics (means, standard deviations, kurtosis, and skewness) were reported for every SAIAT-HE item in each sample. For analyses that involved the use of SAIAT-HE-revised subscales (RQs 4, 5, 6) descriptive statistics were reported for each subscale. After data screening and assumption checking, the following analyses were conducted to answer each one of the research questions.

**Evaluation of model-data fit.** Several research questions (RQs 1, 2, 3, 5) involved evaluating model-data fit in the structural equation modeling framework. In order to evaluate model-data fit, the following fit indices were used. The $\chi^2$ goodness-of-fit test, an absolute fit index, which assesses the overall discrepancy between observed and implied covariance matrices, was reported. Smaller $\chi^2$ values indicate better model-data fit. This test is overly sensitive to sample size, so it was supplemented by other fit indices (Kline, 1998). Namely, another absolute fit index, the Standardized Root Mean Square Residual (SRMR) was used, with the desired values falling below approximately
In addition, the Root Mean Square Error of Approximation (RMSEA), an absolute fit index that takes into account model complexity was also used, with the desired values falling below approximately .05 (Mueller & Hancock, 2010). *Relative or incremental* fit indices assess model fit relative to a baseline model that specifies no relations among the observed variables. The Comparative Fit Index (CFI) was used, with the desired values falling at or above approximately .95. In addition, correlation residuals were used to detect localized misfit, with values greater than |.15| considered large.

**Evaluation of model parameters.** Estimated parameters for adequately-fitting models were interpreted. Specifically, both standardized and unstandardized pattern coefficients, standard errors, z-tests and p-values for every estimated parameter were reported and interpreted. In addition, the amount of variance explained in the outcome was reported and interpreted. Results were reported both in tables and figures to facilitate readability.

**Nested models comparison.** Several research questions involved testing increasingly complex nested models against simpler models in which they are nested (RQs 1, 2, 3, 5 (Models 1 and 2)). Although a more complex model always fits the data better than the less complex model, it is important to evaluate whether the difference in fit is significant enough to justify championing a more complex (but less parsimonious) model (Steenkamp & Baumgartner, 1998). The following criteria were used when comparing relative fit of nested models in the structural equation modeling framework: non-significant chi-square difference (Δχ²), ΔCFI. Based on Chen’s (2007) recommendations for invariance testing with continuous data and ML estimation, ΔCFI <
.01 criterion was used for tests of factor pattern coefficient invariance and for tests of intercept or scalar invariance. In addition, the decrease in correlation residuals as well as substantive changes in parameter estimates when moving from a fuller to reduced models, were considered. The test of measurement invariance of the SAIAT-HE-revised across first-year and mid-career student groups was conducted within a single Structural Means Modeling (SSM) analysis used to answer research questions 1, 2, and 3, which are described next.

**Research questions 1, 2, 3.** Recall that the three research questions subsumed under Stage I focus on the psychometric properties and group differences of the SAIAT-HE-revised among first-year and mid-career students. More specifically, RQ1 focuses on the factorial structure of the SAIAT-HE in two student groups and the necessary scale revisions; RQ2 focuses on the measurement invariance of the SAIAT-HE-revised across these two groups; RQ3 focuses on the latent mean differences of these two student groups on the SAIAT-HE-revised factors. These three questions were addressed within a single structural equation modeling framework: Structural Means Modeling (SSM) (Steenkamp, & Baumgartner, 1998; Sass, 2011).

**Research question 1.** First, dimensionality of the SAIAT-HE was assessed via confirmatory factor analysis (CFA) and scale revisions were performed independently for first-year and mid-career student samples. In order to set the metrics of latent variables, all factor variances were set to unity. For the factor-item alignment of the two models, see Table 1. For theoretical justifications of the proposed factor models and expected scale revisions, see Chapter 1. Item functioning in both models was assessed by examining descriptive statistics (mean and standard deviation), standardized factor
pattern coefficients, and amount of variance explained in each item by the corresponding factor.

For each latent factor, McDonald’s (1999) ω was calculated using unstandardized parameter estimates and error variances. More accurate than Cronbach’s alpha when used with congeneric items, McDonald’s ω is considered acceptable at or above .70 (McDonald, 1999). Variance extracted in each factor (the average squared standardized loading for a factor’s items) was also reported. Also, means, standard deviations, and inter-factor correlations for each one of the SAIAT-HE subscales for every sample were reported.

In the case when different factor structures are found in these two samples and configural invariance does not hold, the implications become quite interesting. Such a finding may indicate that the scale functions differently for these two groups or that certain dimensions of the construct are distinct for one group but not for the other. Assuming that configural invariance does hold across first-year and mid-career students, the following questions were addressed next.

**Research question 2.** In order for researchers to interpret the SAIAT-HE scores the same way for both first-year and mid-career students, it is necessary to establish measurement invariance of the SAIAT-HE-revised across two student groups (Brown, 2006). More specifically, if one’s goal is to evaluate mean differences across the groups on the latent level (RQ 3), it is necessary to establish configural, and at least partial (but preferably full) metric and scalar invariance for at least two items per factor (Steenkamp & Baumgartner, 1998).
The test of metric invariance answers the question of whether the SAIAT-HE items have equal saliency to their corresponding factors across two student samples. It involves constraining unstandardized pattern coefficients to be equal and evaluating the difference in fit between this constrained model and the factor model championed in the previous step. If established, metric invariance indicates that the SAIAT-HE items and factors have equivalent relationships across first-year and mid-career students. If metric invariance does not hold, the next step is to examine whether partial metric invariance holds (see Table 4) by individually constraining each item’s pattern coefficient to be equivalent across groups. At least two items per factor must be invariant to establish partial metric invariance. Nonetheless, violating metric invariance would be interesting, because it implies some items are more applicable for one group than another (Sass, 2011).

Assuming at least partial metric invariance holds the next step is to examine scalar invariance. The test of scalar invariance answers the question of whether observed mean differences on the SAIAT-HE subscales across first-year and mid-career student samples are accounted for by the differences in factor means across groups. In other words, it involves constraining item intercepts to be equal across groups and evaluating the difference in fit between this model and the metric invariance model. If established, scalar invariance indicates that both pattern coefficients and intercepts are equivalent across the two groups, and a test of latent mean differences can be meaningfully carried out.

**Research question 3.** Recall that theoretically expected group differences on the SAIAT-HE-revised are the following. Mid-career students are expected to have lower
scores on Validity and Disillusionment and higher scores on Purpose than first-year students. Assuming that all invariance assumptions necessary for testing latent mean differences (i.e., configural, and at least partial metric and scalar) hold, the next step is to test latent mean differences on the SAIAT-HE factors across first-year and mid-career student samples. These tests were actually conducted at the same time as the scalar invariance test.

Direction and significance of the mean difference on each SAIAT-HE-revised factor was interpreted and supplemented with the standardized effect size (average of the factor variances weighted by sample sizes). Importantly, the metric of the latent factors was set using an invariant indicator (Cordon & Finney, 2008; Rensvold & Cheung, 2001). Invariant indicators were chosen by testing every item for metric and scalar invariance (Rensvold & Cheung, 2001). In the SSM framework, latent means themselves are not estimated, only latent mean difference. On each one of the SAIAT-HE subscales, the latent mean for one group (first-year students) was fixed to zero, so that the latent mean for the other group represents the difference between the groups.

**Research question 4.** How strongly are first-year students’ attitudes toward K-12 accountability tests related to their attitudes toward higher education accountability tests? In order to ascertain the strength of the association between incoming students’ attitudes toward K-12 accountability tests and attitudes of the same students toward accountability testing in higher education, *bivariate correlations* among the SAIAT-K-12 subscales and the SAIAT-HE subscales both on the observed and latent levels were examined. Pearson product moment correlations were used to rank-order consistency of scores (i.e., are those
who are relatively high on Validity for K-12 accountability tests also relatively high on Validity for higher education accountability tests?).

**Research question 5.** What is the nature of the relationship between students’ attitudes towards accountability testing, test-taking motivation, and performance on higher education accountability test? As detailed in Chapter 1, several competing theoretically-justified models explaining the relationships between students’ attitudes towards accountability assessment, test-taking motivation, and performance on the accountability test have been proposed (*Figures 2, 3, 4*). In addition, alternative models were specified post hoc. In order to formally test the utility of each one of the a priori specified models, as well as post hoc models, a single-indicator latent variable approach was used because it allows for partialing out measurement error (Brown, 2006). The following process was used in estimating relationships between single-indicator latent variables. First, the subscale scores of each measure used (i.e., SAIAT-K-12 subscales, SAIAT-HE-revised subscales, SAT math, SAT verbal, GLEX total score, Importance, and Effort) were calculated. Second, the percentage of the subscale score variance attributable to measurement error was calculated as \((1 - r_{xx}) \times (\text{Var}_x)\), where \(r_{xx}\) equals reliability of the total subscale score (i.e., Cronbach’s coefficient alpha) and \(\text{Var}_x\) is the variance of the total composite score. Third, the error variance of each subscale score was fixed to its corresponding error score variance calculated in the previous step. Fourth, the observed subscale scores were used as single indicators of the latent variables, with the paths from observed to latent variables fixed to one. As such, “error-free” relationships among single-indicator latent variables were modeled.
After the best-fitting model was identified using the criteria specified above and championed as one viable representation of the phenomenon, all direct and indirect effects were reported and interpreted. Moreover, the amount of unexplained variance in Importance, Effort and Performance was examined to shed light on whether variables other than students’ attitudes need to be considered in the low-stakes contexts.

**Research question 6.** In order to evaluate whether students who attended their scheduled assessment session (i.e., compliant students) differed in their attitudinal profile from students who skipped the assessment session (i.e., non-compliant students), *descriptive analyses* of SAIT-HE-revised subscales were conducted first. More specifically, means (with Confidence Intervals), standard deviations, skew and kurtosis values for each of the SAIT-HE-revised subscales were compared and contrasted across these two samples.

In addition, analysis of covariance (ANCOVA) was conducted to test whether mean differences on each one of the SAIT-HE-revised subscales between compliant and non-compliant mid-career students were larger than expected by chance, after removing the effect of a covariate, Conscientiousness. Given the quasi-experimental design of the current study (i.e., students could not be randomly assigned to compliant and non-compliant groups), ANCOVA was used as a statistical adjustment for group differences on conscientiousness; adjusted means on the SAIT-HE-revised subscales were compared as if groups had the same level of Conscientiousness (Tabachnik & Fidell, 2001).

ANCOVA was conducted as a hierarchical multiple regression. Group membership (compliant and non-compliant) was dummy coded (1 assigned to compliant
and 0 assigned to a non-compliant group). Given that there are only two levels of group membership, a single interaction term (between group membership and a covariate) was computed. To avoid potential multicollinearity issues between the covariate and the interaction term, conscientiousness was first centered so that the mean equals zero (Aiken & West, 1991).

One of the ANCOVA assumptions is homogeneity of regression slopes, which states that the relationship between a covariate (i.e., conscientiousness) and dependent variables (i.e., SAIAT-HE-revised subscale) is similar across compliant and non-compliant groups, indicating that the covariate functions similarly across groups. This assumption of null interaction was tested first. After supporting homogeneity of the regression slopes assumption, the differences in adjusted means were then interpreted. That is, the answer to RQ6 offered insight into whether compliant and non-compliant students differed on the SAIAT-HE-revised subscales after controlling for group differences on conscientiousness.
Results

Recall that the current program of research consists of two stages with three research questions subsumed under each stage. Stage I pertains to building a validity argument for the SAIAT-HE. More specifically, research questions under Stage I focus on the psychometric properties of the SAIAT-HE for first-year and mid-career students and resulting scale revisions; invariance of the SAIAT-HE across these two student populations; and known-groups validity evidence. Stage II pertains to substantive research questions related to students’ attitudes toward institutional accountability testing, including the strength of the relationship between attitudes towards testing in K-12 and attitudes towards testing in HE; the nature of the relationship between students’ attitudes, test-taking motivation, and performance; and attitudinal profiles of compliant and non-compliant students. Below, the results for each research question are presented, followed by a general discussion of all findings.

Research Question #1

What revisions to the SAIAT-HE are necessary based on the results from first-year and mid-career student samples? As detailed in Chapter 2, the initial investigation of the 22-item SAIAT-HE using a mid-career student sample (Zilberberg et al., in press) supported a correlated six-factor model (see Table 1 for item-factor mapping). However, several poorly functioning factors and items were also identified, necessitating the need to investigate whether these items function consistently poorly in first-year and mid-career student samples. As such, the results of the current analyses, in conjunction with the previous analysis of the 22-item SAIAT-HE, are used to make necessary scale revisions and propose a revised scale with the revised factor structure. Furthermore, the
current study is the first evaluation of the original and modified SAIAT-HE using first-year students.

The results are presented in the following order. First, descriptive statistics for the first-year and mid-career student samples are discussed, foreshadowing the results of factor and item analysis. Second, the fit of the six-factor model in both samples is discussed. Third, functioning of the six factors and 22 items in both samples is presented. Fourth, the fit of the three-factor model in both samples is discussed. Fifth, functioning of the three factors and ten items in both samples is presented.

**Descriptive statistics.** Descriptive statistics for the first-year, mid-career calibration, and mid-career validation student samples are summarized in Tables 5, 6, and 7, respectively. None of the standard deviations were below 1 in any of the samples, indicating sufficient variability in students’ responses to all items and lack of extreme floor/ceiling effects. Examination of the three correlation matrices revealed that data were not multicollinear in the bivariate sense ($r < .90$ for all pairs of variables in both samples). Moreover, tolerance values for all items in both samples were below .10, indicating that data were also not multivariately collinear. Skewness and kurtosis values did not exceed $|2|$ and $|7|$ in any of the samples, respectively, indicating that all 22 items followed univariate normal distributions. However, data were multivariately non-normal with Mardia’s standardized coefficient equal to 31.33 ($>3$) in the first-year student sample, 45.43 ($>3$) in the mid-career validation student sample, and 51.84 ($>3$) in the mid-career validation, thus necessitating the use of Satorra-Bentler adjustment to the chi-square statistic, approximate fit indices, and standard errors when employing confirmatory factor analysis.
Prior to testing the a priori specified factor structure, a closer examination of the pattern of correlations among the items in all samples suggested the existence of six distinct factors underlying the data. That is, correlations among the items written to represent each of the six factors tended to be higher than correlations among other items, thus creating six “clusters” of correlations that foreshadowed the proposed six-factor dimensionality. Specifically, correlations among most of the items assigned to Validity (1, 2, 3, 4, 5), Purpose (6, 7, 8, 9) and Disillusionment (10, 11, 12) were higher in magnitude with items written to represent the same factor than with other items, and tended to “cluster” together, in accordance with these three proposed factors. With that said, item 1 (Validity) correlated weakly with other Validity items in the first-year student sample (r ranging from .17 to .39) and in both mid-career student samples (r ranging from .11 to .32), suggesting that this item will likely emerge as a non-salient indicator in the factor analysis. Similarly, item 8 (Purpose) correlated weakly with other Purpose items in the first-year student sample (r ranging from .33 to .37) and in both mid-career student samples (r ranging from .18 to .29), foreshadowing its low utility as an indicator of Purpose. Moreover, the magnitude of the correlations within the three remaining “clusters” were quite low, indicating that the corresponding factors will not be well represented by this set of items and the common variance defining these factors will be small. That is, correlations among items 13 – 22 designed to represent the “influence” factors (Parents, Professor, Students) were weak in all samples, suggesting that these items do not share a lot of common variability and thus may not be salient to their respective factors. For example, item 20 “Fellow students urged me to try my best on the assessment tests” correlated with the other two “Student” items at $r = .10$ in the first-year
student sample, at \( r = .12 - .28 \) in the mid-career validation sample, and at \( r = .06 - .25 \) in the mid-career calibration sample. In sum, this pattern of correlations supports the predicted factor structure of the scale (i.e., six interrelated factors) and highlights items that will likely have low factor pattern coefficients.

In addition, several items had problematic response patterns. For example, a large proportion of students selected “neither agree nor disagree” option in response to item 14 “My parents don’t value the assessment tests I complete at the University” (41.7% of first-year students; 49.4% and 46.8% of mid-career students in the calibration and validation samples, respectively) and in response to item 18 “My professors don’t value the assessment tests I complete at the University” (34.3% of first-year students; 43.5% and 46.2% of mid-career students, in calibration and validation samples, respectively). This pattern of responses suggests that students may not know whether their parents and professors value assessments, leading to the ill-defined “influence” factors.

**Fit of the six-factor model.** A six-factor correlated model was fit to two samples: first-year student sample and mid-career calibration student sample. Factor variances were fixed to a value of 1 to set the metric of the latent variables. The top section of Table 8 presents the global fit indices with Satorra-Bentler adjustments for multivariate non-normality for the six-factor model. Overall, the six-factor model fit the data in both samples adequately, as predicted. Specifically, absolute fit indices (Robust RMSEA, SRMR) indicated adequate model-data fit in both samples. However, the values of a CFI did not reach the cut-off value of 0.95. An examination of the correlation residuals presented in Table 9 provides rich information regarding local model-data fit. In the first-year student sample, about 14% (33 out of 231 correlation residuals exceeded the
absolute value of .10) of all inter-tem relationships were poorly reproduced by the six-factor model. Specifically, several of the relationships between items on the same factor were underestimated by the model, indicating that these items correlated stronger than predicted by the model (Validity items 3 & 5; Purpose items 8 & 9; Disillusionment items 10 & 11). The majority of the residuals associated with the items corresponding to Parents, Professors, and Students factors (13, 14, 18, 19, 20, 21, and 22) were above the absolute value of .10, with some of the relationships being underestimated and others being overestimated.

A somewhat similar pattern of results emerged in the mid-career calibration student sample, with the local fit being slightly better than in the first-year student sample. Specifically, about 11% (26 out of 231 correlation residuals exceeded the absolute value of .10) of all inter-tem relationships were poorly reproduced by the six-factor model. Importantly, none of the relationships between the items corresponding to the factors of Validity, Purpose, or Disillusionment were under or overestimated. However, two relationships between the items across these three factors were underestimated: the relationship between item 3 (Validity) and item 8 (Purpose); and between items 1 (Validity) and 6 (Purpose). The majority of the residuals were associated with the items corresponding to Parents, Professors, and Students factors (13, 14, 15, 17, 18, 19, 20, 21), with some of the relationships being underestimated and others being overestimated.

Taken together, global and local fit indices indicate that the six-factor model fit the data adequately globally but also contains areas of misfit, especially with respect to
the “influence” factors. Next, the utility of the factors and the functioning of individual items are considered.

**Functioning of six factors and 22 items.** Table 10 presents unstandardized and standardized pattern coefficients, $R^2$ values for each item, reliability indices, and amount of variance accounted for by each factor in the six-factor model for the two student samples. First, the factors whose utility has been questioned in past study (Zilberberg et al., in press) due to low reliability and $R^2$ values of the items are examined: Parents, Professors, and Students. Given the results from the previous study in addition to the low correlations described above, it was not unexpected to find that the Student factor accounted for only 36.68% and 39.94% of the variance in the corresponding items in the first-year and the mid-career calibration student samples, respectively. As foreshadowed by low bivariate correlations among the Student items, all items except item 22 had low factor pattern coefficients and low $R^2$ values in both samples, indicating poor saliency of these items to the Students factor. The reliability (McDonald’s Omega) associated with these Students items was only .51 and .56 for first-year and mid-career students, respectively. For these reasons, the entire Students factor is omitted from the SAIAT-HE-revised.

With respect to the Professor factor, it accounted for only 36.56% and 32.61% of the variance in the corresponding items in the first-year and the mid-career calibration student samples, respectively. Most of the Professor items had low standardized pattern coefficients and low $R^2$ values, as foreshadowed by low bivariate correlations among these items. Moreover, a high rate of endorsement for “neither agree nor disagree” option for item 18 (“My professors don’t value the assessment tests I complete at the
University”) across both samples suggests that students are not aware of their professors’ views on assessment. The reliability (McDonald’s Omega) associated with these Professor items was only .55 and .52 for first-year and mid-career students, respectively. Based on these findings and the previous study of these items, the Professors factor was omitted from the SAIAT-HE-revised.

With respect to the Parents factor, it accounted for 42.53% and 36.55% of the variance in the corresponding items in the first-year and mid-career calibration student samples, respectively. Aligning with the original hypothesis, the Parents factor functioned slightly better in the first-year rather than in the mid-career student sample, most likely because parents have a more influential role on younger students’ attitudes toward all academic activities, including their performance on assessment tests. However, the Parents items functioned poorly in both samples, as indicated by low standardized pattern coefficients and low $R^2$ values, as foreshadowed by low bivariate correlations. Similar to the finding for the Professor factor, there was a high rate of endorsement for “neither agree nor disagree” option for item 14 (“My parents don’t value the assessment tests I complete at the University”) across both samples, suggesting that students are not aware of their parents’ views on assessment. The reliability (McDonald’s Omega) associated with these Parents items was only .62 in both first-year and mid-career student samples. Based on these findings and the previous study of these items, the Parents factor was omitted from the SAIAT-HE-revised.

With respect to Validity, Purpose, and Disillusionment factors, they performed rather well, all accounting for over 45% of variance in the corresponding items in the first-year and mid-career calibration student samples. Moreover, the reliability indices
(McDonald’s Omega) associated with these factors were above .71 in both samples (see Table 10). Minor revisions were necessary to Validity and Purpose factors. Items previously identified as poorly functioning continued to exhibit poor psychometric properties in these two samples, leading to their omission from the scale. Specifically, Validity item 1 “Assessment tests are unfair to some students” had a low standardized pattern coefficient and a low $R^2$ value in both samples. Another Validity item, item 3 “Assessment test results accurately reflect basic skills and knowledge of the subject” had a fairly low factor pattern coefficient for first-year but not mid-career students. Item 3 was not identified as poorly functioning in the initial study of the SAIAT-HE and thus was kept on the scale. Purpose item 8 “Someone (professor, academic advisor, Resident Advisor) explained to me why I take assessment tests” functioned very poorly in the mid-career sample, with only 5.9% of its variance explained by the Purpose factor. This low common variance is most likely due to low bivariate correlations between item 8 and other Purpose items. Although item 8 functioned slightly better in the first-year student sample, only 19% of its variance was explained by the Purpose factor. Based on these findings and the previous study of these items, items 1 and 8 were omitted from the SAIAT-HE-revised. In summary, the following revisions resulted from the analyses described above and align with recommendations from previous study of the measure: Students, Parents, and Professor factors were omitted; Validity item 1 and Purpose item 8 were also omitted.

**Fit of the three-factor model.** Given the results from the current and previous study of the 22-item, six-factor SAIAT-HE, a 10-item, three-factor version (SAIAT-HE-revised; see Table 1) was created and consisted of Validity (four items), Purpose (three
items), and Disillusionment (three items). The three-factor model was fit to the 10-item SAIAT-HE-revised using data from three samples: first-year, mid-career calibration, and mid-career validation (which is an independent sample not used in the prior CFA analyses). The bottom section of Table 8 presents the global fit indices with Satorra-Bentler adjustments for multivariate non-normality for the three-factor model. The three-factor model yielded good model-data fit in the first-year and mid-career calibration student samples, which was expected given that this three-factor model was derived using these two samples. Importantly, the three-factor model also fit very well in the independent mid-career validation student sample. The correlation residuals presented in Table 11 also supported the fit of the three-factor model. In the first-year student sample, the relationship between Validity items 3 and 5 was slightly underestimated, as indicated by a positive correlation residual of .13, and the relationship between Validity item 5 and Disillusionment item 10 was slightly overestimated, as indicated by the negative correlation residual of -.12. In the mid-career calibration sample, the relationship between Validity items 3 and 5 was also slightly underestimated, as indicated by a positive correlation residual of .17; and the relationship between Disillusionment items 10 and 11 was slightly underestimated, as indicated by the positive correlation residual of .15. To test whether the shared variance between these items is ignorable (i.e., if the residuals are practically small), a three-factor model with correlated errors added between items 3 and 5; items 10 and 5; and items 10 and 11 was tested. The overall model fit and the parameter estimates changed only minimally with the addition of these correlated error terms, thus this slight local misfit was deemed negligible. Notably, none of the correlation residuals from the independent mid-career validation sample exceeded |.05|. 
offering strong validity evidence for the three-factor model of the SAIAT-HE-revised.

Given good global and local model-data fit, the next step is to consider utility of the three factors and functioning of each of the ten items.

**Functioning of three factors and 10 items.** Table 12 presents unstandardized and standardized factor pattern coefficients, $R^2$ values for each item, reliability indices, and amount of variance accounted for by each factor in the three-factor model for the following three samples: first-year, mid-career calibration, and mid-career validation samples. The utility of the Validity factor was fairly high; it accounted for 51.52%, 58.60%, and 57.10% of the variance in the corresponding items for first-year, mid-career calibration, and mid-career validation samples, respectively. Internal consistency of the Validity factor was acceptable, with McDonald’s Omegas reaching .72, .76, and .76 in the first-year, mid-career calibration, and mid-career validation samples, respectively. The standardized pattern coefficients for all four Validity items exceeded .4, and $R^2$ values exceeded .25 for both mid-career samples. Items 3 and 5 functioned less well for the first-year student sample, with $R^2$ values of .15 and .23, respectively.

The utility of the Purpose factor was also high; it accounted for 72.93%, 75.00%, and 74.20% of the variance in the corresponding items for first-year, mid-career calibration, and mid-career validation samples, respectively. Internal consistency of the Purpose factor was acceptable, with McDonald’s Omegas reaching .80, .81, and .81 in the first-year, mid-career calibration, and mid-career validation samples, respectively. All three Purpose items functioned well across the samples, as indicated by standardized factor pattern coefficients at or above .58 with $R^2$ values at or above .32.
The utility of the Disillusionment factor was comparable to the other two factors; it accounted for 67.53%, 66.40%, and 71.80% of the variance in the corresponding items for first-year, mid-career calibration, and mid-career validation samples, respectively. Internal consistency of the Disillusionment factor was acceptable, with McDonald’s Omegas reaching .72, .76, and .79 in the first-year, mid-career calibration, and mid-career validation samples, respectively. Disillusionment items 11 and 12 functioned well in all three samples, as indicated by standardized factor pattern coefficients at or above .65 and $R^2$ values at or above .40. Disillusionment item 10 functioned slightly better in the mid-career student samples than it did in the first-year student sample, as indicated by larger standardized residuals and $R^2$ values.

Inter-factor correlations indicate that students’ attitudes are inter-related. As expected, Validity and Purpose were moderately and positively related, with $r$ being .62, .60, and .68 in the first-year, mid-career calibration, and mid-career validation student samples, respectively. Validity and Disillusionment were rather strongly and negatively related, with $r$ being -.74, -.77, and -.77 in the first-year, mid-career calibration, and mid-career validation student samples, respectively. Purpose and Disillusionment were also rather strongly and negatively related to each other in all three samples, with $r$ being -.73, -.65, and -.67 in the first-year, mid-career calibration, and mid-career validation student samples, respectively.

**Research Question #2**

Does the SAIAT-HE function invariantly for first-year students and mid-career students? In order for researchers to interpret the SAIAT-HE scores the same way for both first-year students and mid-career students, it is necessary to establish measurement
invariance of the SAIAT-HE across these two student groups (Brown, 2006). As detailed in Chapter 3, the following levels of invariance were evaluated: configural (i.e., equivalence of the factor structure across the two samples), metric (i.e., equivalence of item saliency to their corresponding factors across the two samples), and scalar (i.e., equivalence of item intercepts across the two samples). The following two samples are used to test measurement invariance of the SAIAT-HE: first-year and mid-career validation student samples (Tables 2 and 3).

**Configural invariance.** Global and local fit of the three-factor model to the first-year and two mid-career student samples was established and was discussed in detail in the previous section; fit indices are presented in Table 8 and item functioning is presented in Table 12. Based on these results, configural invariance for the SAIAT-HE-revised across first-year and mid-career students was confirmed. Given that the configural invariance levels were upheld, the next step is to examine metric invariance.

**Metric invariance.** In the metric invariant model, the unstandardized factor pattern coefficients (Table 12) were constrained to be equal across the two groups, with all other parameters freely estimated. The metric invariant model fit data well overall and did not fit significantly worse than a configural model in which it was nested (Table 13). Moreover, correlation residuals did not increase in number or in size when the factor pattern coefficients were constrained to be equal. For these reasons, metric invariance for the SAIAT-HE-revised across first-year and mid-career students was established. Given that the metric invariance was upheld, the next step is to examine scalar invariance, or the equivalence of item intercepts.
Scalar invariance. In the scalar invariant model, the unstandardized factor pattern coefficients and the item intercepts were constrained to be equal across the two groups, with all other parameters freely estimated. The scalar invariant model fit the data well overall (Table 13). Although the significant $\Delta \chi^2$ value indicated a potential violation of scalar invariance, the $\Delta$CFI was negligible ($<$ .01). Two of the largest unstandardized mean residuals representing the difference between observed and model-implied means on the metric of 1-7 Likert scale were associated with Disillusionment items 10 (.12 point difference) and 11 (.15 point difference). These residuals did not reach a quarter of a point (.25) on the Likert scale, signifying that their size is negligible and thus this amount of uniform DIF is negligible. Nonetheless, partial scalar models with freed intercepts for items 10 and 11 were estimated in order to investigate whether the minimal uniform DIF present in these items had an impact on substantive conclusions. Although these partial scalar invariance models fit the data statistically better than the full scalar model, the difference in model fit was not practically better (e.g., CFI changed by less than .01 units; RMSEA changed by .007 units). Importantly, the latent mean difference in Disillusionment across first-year and mid-career students did not change substantively when including or excluding items 10 and 11 in the computation, suggesting any uniform DIF was ignorable. For these reasons, the misfit associated with these items was considered negligible and full scalar invariance for the SAIAT-HE-revised across first-year and mid-career students was supported. Given that configural, metric, and scalar invariance was upheld, the research question pertaining to latent mean differences in attitudes across student populations is addressed next.
**Research Question #3**

Recall the theoretically expected group differences with respect to attitudes toward higher education accountability tests: mid-career students are expected to have lower scores on Validity and higher scores on Purpose and Disillusionment than first-year students. Table 14 presents the latent and observed mean differences across first-year and mid-career students. All latent and observed mean differences were statistically significant, which is not surprising given the sample size. As predicted, mid-career students scored statistically significantly lower than first-year students on Validity, with the latent effect size indicating a practically significant difference of half a standard deviation. Also, as predicted, mid-career students scored statistically and practically higher on Disillusionment, with the latent effect size indicating a difference of over half a standard deviation. Contrary to the hypothesis, mid-career students were statistically and practically lower than first-year students on Purpose, although the latent effect size was smaller than for the other two factors, indicating a difference of about a quarter of a standard deviation. Comparison of observed and latent mean differences reveals that, although observed mean differences are contaminated with measurement error as expected, they are not that grossly attenuated. This observation is due to the high internal consistency of the subscales. Therefore, if sample size restrictions preclude researchers from using latent variable analyses, the current analysis indicates that observed variables methods using SAIAT-HE-revised subscales are likely to provide accurate results.

**Research Question #4**

How strongly are first-year students’ attitudes toward K-12 accountability tests related to their attitudes toward higher education accountability tests? As detailed in
Chapter 1, the theoretical expectation was that newly minted first-year students who encounter college accountability testing for the first time will view it through the lens of prior experience with accountability testing in K-12. Given the similarities between the testing contexts and proximity in time from high school graduation to the beginning of college, the relationships between the two sets of students’ attitudes were predicted to be positive and range in magnitude from moderate to strong.

To empirically investigate this hypothesis, the strength and direction of bivariate correlations between incoming students’ attitudes toward K-12 accountability testing (SAIAT-K-12) and attitudes of the same students toward accountability testing in higher education (SAIAT-HE-revised) were evaluated. Given that a previous study of the SAIAT-K-12 indicated a stable factor structure and adequate subscale reliability (Zilberberg et al., in press) and the current study of the SAIAT-HE-revised supported the expected factor structure and uncovered high internal consistency, subsequent analyses employed these two measures. Prior to discussing the relationships between attitudes toward higher education accountability testing and attitudes toward K-12 accountability testing, descriptive statistics for the SAIAT-K-12 and the SAIAT-HE-revised dimensions are presented in order to gauge the average levels of students’ attitudes and foreshadow the results of subsequent analyses.

**Descriptive statistics.** Table 15 contains descriptive statistics, internal consistency indices, and correlations among each of the SAIAT-K-12 and the SAIAT-HE-revised dimensions. Internal consistencies for all subscales were adequate (Cronbach's $\alpha > .70$) except for the SAIAT-K-12 Parents subscale (Cronbach's $\alpha = .68$). Therefore, the score for the Parents subscale should be interpreted with caution.
Moreover, in order to address less than perfect reliability, latent correlations were computed, in addition to the observed correlations. None of the standard deviations for any of the subscales were below 1, indicating sufficient variability. On average, incoming college students were neutral in their perception of validity of K-12 accountability tests, neutral to positive in their perceived understanding of the tests’ purpose, and slightly disillusioned about accountability testing. In addition, they indicated that their parents exhibit slight concern with regards to performance on such tests. The profile of attitudes toward accountability tests in higher education was similar. On average, these first-year college students were neutral in their attitudes toward the validity of college accountability tests, slightly positive in their perceived understanding of the tests’ purpose, and had neutral affect towards such tests.

**Correlations between the SAIAT-K-12 and SAIAT-HE-revised subscales.**

Table 15 displays bivariate correlations between all SAIAT-K-12 and SAIAT-HE-revised subscales, both at the observed and latent levels. Expectedly so, correlations computed among latent factors were higher in magnitude than correlations computed among the observed subscales because of the correction due to the measurement error in the observed scores. Nonetheless, as discussed above, most of the SAIAT-K-12 and the SAIAT-HE-revised subscales were adequate in terms of internal consistency, indicating data from both measures can be analyzed using observed variable statistical techniques (e.g., ANOVA, multiple regression) if low sample size prohibits the use of latent variable techniques (e.g., latent means modeling, latent structural equation modeling, single-indicator path analysis).
The correlation values between corresponding SAIAT-K-12 and SAIAT-HE-revised subscales (Validity, Purpose, and Disillusionment) are of primary research interest and are thus bolded in Table 15. As predicted, students’ attitudes toward K-12 accountability testing were positively and moderately correlated with the corresponding attitudes toward accountability testing in college. Specifically, the observed correlation between the Validity subscales was $r = .33$, 95% CI [.26, .39], indicating that two subscales share about 11% of their observed variance (latent correlation = .36). The observed correlation between Purpose subscales was $r = .31$, 95% CI [.25, .36], indicating that two subscales share about 10% of their observed variance (latent correlation = .41). The observed correlation between Disillusionment subscales was $r = .33$, 95% CI [.27, .39], indicating that two subscales share about 11% of their variance (latent correlation = .34). The SAIAT-K-12 Parents subscale was weakly and positively related to the SAIAT-HE-revised Validity and Purpose subscales at $r = .15$, 95% CI [.10, .20] and $r = .16$, 95% CI [.10, .20] (only about 2% of the variance shared across constructs) The SAIAT-K-12 Parents subscale was also weakly and negatively related to the SAIAT-HE-revised Disillusionment subscale at $r = -.18$, 95% CI [-.25, -.20] (about 3% of the variance shared). Relationships among the remaining SAIAT-K-12 and SAIAT-HE-revised subscales ranged from |.15| to |.27|. In summary, the correlational analyses revealed that two sets of attitudes measured by the SAIAT-K-12 and the SAIAT-HE-revised are clearly distinct, though related, constructs.

**Research Question #5**

What is the nature of the relationship between students’ attitudes toward accountability testing in higher education, test-taking motivation, and performance on a higher education accountability test? As detailed in Chapters 1 and 2, detangling the
effects of students’ attitudes on test performance is of utmost importance. As detailed in Chapter 3, a single-indicator latent variable approach, which partials out measurement error, is best suited for formally testing this relationship. Two competing models were hypothesized a priori to underlie such relationships (see Figures 2 and 3) and were fit to two independent student samples. In addition, an expanded model that included students’ attitudes toward accountability testing in K-12 was specified (see Figure 4) and fit to a first-year student sample. These models were slightly modified, yielding two new modified models (Figures 5 and 6). All of these models and their fit to the data are discussed below in the following order. First, fully-mediated Model 1 (Figure 2) and its fit in both first-year and mid-career student samples are detailed. Second, partially mediated Model 2 (Figure 3) and its fit in both first-year and mid-career student samples are outlined. Third, modified fully-mediated Model 4 (Figure 5) and its fit in both first-year and mid-career student samples are detailed. Fourth, partially-mediated Model 3 with SAIAT-K-12 and its fit in the first-year student sample are described. Fifth, modified fully-mediated Model 5 with SAIAT-K-12 and its fit in the first-year student sample is discussed.

**Model 1: Fully-mediated model.** This model, depicted in Figure 2, posits that the relationship between attitudes towards higher education accountability tests and test performance is fully mediated via test-taking Effort and perceived test Importance (two components of test-taking motivation), after controlling for the effects of verbal (SAT verbal) and math (SAT math) ability on Importance and Performance. Furthermore, this model posits that students’ attitudes affect test-taking Effort only *indirectly*, via Importance. In turn, Importance influences Performance *indirectly*, via Effort, not
directly. According to this model, Effort is the only variable (besides the control variables of SAT math and SAT verbal) affecting Performance directly.

If this model accurately represents reality, the observed correlations among the variables hypothesized to be directly related (e.g., Importance and Effort) will be larger in magnitude compared to the correlations among the variables hypothesized to be related indirectly, or mediated via other variables (e.g., indirect relationship between Importance and Performance mediated via Effort). Therefore, examining correlational patterns among the variables is informative for foreshadowing the results of subsequent single-indicator latent variable analysis and for identifying potential areas of model misfit (Kelloway, 1995). Thus, an overview of the descriptive statistics as well as an examination of the correlational patterns both in the first-year and mid-career student samples follows, preceding the results of the single-indicator latent variable analysis.

**Matched incoming and first-year student sample.** Descriptive statistics and correlations among all the variables are reported in Table 16. Estimates of skewness and kurtosis did not exceed |3| and |8|, respectively, for all measures, indicating that non-normality will have negligible effects on fit indices and standard errors (Finney & DiStefano, 2006). Average subscale scores on Importance and Effort indicated that students were neutral with respect to how important they perceived college accountability tests to be ($M = 3.18$ on a 5-point Likert scale) but invested a moderate amount of effort into completing these tests ($M = 3.80$ on a 5-point Likert scale). There was sufficient variability in Importance and Effort subscale scores, with standard deviation estimates of 0.75 and 0.71, respectively. Students’ SAT math ($M = 579.56$, $SD = 66.70$) and SAT verbal ($M = 573.95$, $SD = 68.83$) scores were typical of the university’s student body.
The average score on the college accountability test designed to measure knowledge of global experience (GLEX) was 21.00, indicating that first-year college students answered about 65.6% of the GLEX items correctly. Although this test is fairly difficult for first-year students, there was moderate amount of variability in GLEX scores ($SD = 4.82$), indicating that students varied substantially in their performance on this test. With respect to attitudes toward accountability tests in higher education, this group of students held neutral to positive perception of the validity of the tests ($M = 4.08$ on a 7-point Likert scale), reported their perceived understanding of the tests’ purpose to be moderately high ($M = 5.07$ on a 7-point Likert scale), and held neutral to slightly negative affect towards both types of tests ($M = 4.39$ on a 7-point Likert scale). Internal consistency estimates (Cronbach’s alphas) were adequate (ranging from .72 to .83) for all measures.

Correlations among the variables (see Table 16) shed light on the accuracy of hypothesized relationships specified in Model 1 (Figure 2). Recall that variables hypothesized to be related directly should correlate stronger than the variables expected to be related only indirectly. As expected, SAT verbal had the strongest relationship with Performance ($r = .64$) followed by SAT math ($r = .39$), thus aligning with the model-implied direct effects of ability (SAT scores) on test Performance. Notably, correlations between SAT math and Importance ($r = -.04$) as well as between SAT verbal and Importance ($r = .03$) were essentially zero, indicating that direct paths between SAT scores and Importance are not necessary and thus would likely be non-significant when estimating the model. Also, as previously found, SAT math was unrelated with Effort ($r = .02$) (Wise & Kong, 2005) and SAT verbal was only weakly related to Effort ($r = .14$), further evidencing that direct effects are not needed between these variables. Jibing with
previous findings, Effort and Importance were moderately inter-related \( r = .39 \); Importance was weakly related to Performance \( r = .14 \) and Effort was moderately related to Performance \( r = .29 \) (Sundre & Kitsantas, 2004; Wise & DeMars, 2005; Thelk et al., 2009). Again, this pattern supports the hypothesized direct effects between Effort and Performance, Importance and Effort, as well as an indirect effect of Importance on Performance via Effort. With respect to students’ attitudes, Validity, Purpose, and Disillusionment were related more strongly to Importance (all \( r s \) above |.32|) than to Effort (all \( r s \) above |.25|), suggesting that direct effects from attitudes to Importance are needed, and that the weaker relationships between the attitudes and Effort are likely to be well-represented by indirect effects mediated via Importance. However, the correlation between Purpose and Effort was larger than expected \( r = .28 \), indicating that modeling a direct (in addition to indirect) relationship between these two variables may be necessary. Moreover, two SAIAT-HE subscales (Validity and Purpose) were weakly related to Performance (\( r s \) slightly above .15), whereas Disillusionment was not related to Performance at all \( r = -.06 \). Therefore, direct effects between attitudes and Performance are not likely to be needed in order to accurately represent the observed relationships. In other words, this correlational pattern suggests that student attitudes toward college accountability tests affect students’ performance on such tests \textit{indirectly}, via Importance and Effort, not directly, aligning with Model 1.

\textit{Mid-career student sample.} Descriptive statistics and correlations among all the variables obtained from the mid-career student sample are presented in Table 17. Estimates of skewness and kurtosis did not exceed |3| and |8|, respectively, for any of the variables, indicating that non-normality is not likely to cause estimation issues (Finney &
DiStefano, 2006). Aligning with the previous finding that mid-career students have more skeptical attitudes toward testing than their first-year peers (see Research Question #3 which uses the same first-year sample, but a different independent mid-career student sample), mid-career students’ average scores on the Validity ($M_1 = 3.42$ on a 7-point Likert scale) and Purpose ($M_1 = 4.87$ on a 7-point Likert scale) subscales were lower than first-year students’, and their average score on Disillusionment ($M_1 = 4.88$ on a 7-point Likert scale) was higher than that of the first-year students’. There was a sufficient amount of variability in all SAIT-HE-revised subscales, with all standard deviations exceeding 1. Notably, mid-career students perceived college accountability tests to be less important than their first-year peers, as indicated by the low average on Importance ($M_1 = 2.66$ on a 5-point Likert scale), but reported investing similarly moderate amount of test-taking effort, as indicated by their average on Effort ($M_1 = 3.72$ on a 5-point Likert scale). There was sufficient amount of variability in Importance and Effort subscales, with standard deviation estimates of 0.89 and 0.81, respectively. Students’ SAT math ($M = 577.65$, $SD = 68.84$) and SAT verbal ($M = 565.86$, $SD = 70.05$) scores were typical of this university student body. The average score on the college accountability test designed to measure knowledge of global experience (GLEX) was 22.81, indicating that mid-career students answered about 71.3% of the GLEX items correctly, which is slightly higher than 65.6% of items answered correctly by the first-year students. Evidently, this test was fairly difficult for mid-career students, but there was still moderate amount of variability in GLEX scores ($SD = 4.81$), indicating that students vary substantially in their Performance. Internal consistency estimates (Cronbach’s alpha) ranged from adequate to good for all measures used (.73 to .86).
Correlations among the variables (see Table 17) shed light on the accuracy of the relationships specified in Model 1. Recall that variables hypothesized to be related directly should correlate stronger than the variables expected to be related only indirectly. Similar to the findings observed in the first-year student sample, SAT verbal had the strongest relationship with Performance \( (r = .56) \) followed by SAT math \( (r = .35) \), thus aligning with the model-implied direct effects from prior knowledge (SAT scores) to test Performance. Also similar to the first-year sample, correlations between SAT math and Importance \( (r = -.07) \) as well as between SAT verbal and Importance \( (r = -.07) \) were essentially zero, indicating that the direct paths between SAT scores and Importance will likely not be significant. Similar to first-year students, both SAT verbal and SAT math were only weakly related to Effort \( (r = .12 \) for both), further corroborating lack of a direct relationship between ability and effort (Wise & Kong, 2005). Aligning with first-year student results, Effort and Importance were moderately inter-related \( (r = .43) \); Importance was weakly related to Performance \( (r = .16) \) and Effort was moderately related to Performance \( (r = .34) \). Again, this pattern supports the hypothesized (Model 1, Figure 2) direct effects between Effort and Performance, Importance and Effort, as well as an indirect effect of Importance on Performance via Effort (Sundre & Kitsantas, 2004; Wise & DeMars, 2005; Thelk et al., 2009). Similar to the first-year students, Validity, Purpose, and Disillusionment were related more strongly to Importance (all \( rs \) above \(|.37|\)) than to Effort (all \( rs \) above \(|.33|\)), suggesting that direct effects from attitudes to Importance are needed, and that the weaker relationships between the attitudes and Effort are likely to be indirect, mediated via Effort. However, correlations between attitudes and Effort were still sizeable, indicating that modeling direct (in addition to indirect) relationships
between these variables may be necessary. Similar to findings in the first-year student sample, two SAIAT-HE subscales (Validity and Purpose) were weakly related to Performance ($r_s$ slightly above .12), whereas Disillusionment was not related to Performance at all ($r = - .02$), indicating that, as predicted, direct effects between attitudes and Performance are also not needed in the mid-career sample.

**Model 1: Fit and parameters.** After examining descriptive statistics, fully-mediated Model 1 (Figure 2) was fit to both first-year and mid-career student samples. According to the global fit indices presented in Table 18, Model 1 yielded excellent model-data fit in the first-year sample and acceptable model-data fit in the mid-career sample. Most importantly, localized model-data misfit was minimal (89% and 82% of correlation residuals did not exceed $| .10 |$ in the first-year and mid-career student samples, respectively), as foreshadowed by the inspection of the correlational patterns. That is, most observed correlations were well-reproduced by the fully-mediated model in both samples, suggesting that most of the relationships between the variables were correctly specified in Model 1. Standardized parameter estimates obtained from both samples are presented in Figure 2. Specifically, direct effects from SAT math and SAT verbal to Performance were statistically and practically significant, with SAT verbal having a strong effect on Performance. Furthermore, there was a practically and statistically significant direct effect from Importance to Effort (over half of a standard deviation increase in Effort for every standard deviation increase in Importance) and from Effort to Performance (approximately a third of a standard deviation increase in Performance for every standard deviation increase in Effort) in both samples. Importantly, the indirect effect of Importance on Performance via Effort was statistically and practically
significant (standardized indirect effect = .14 and .19 in the first-year and mid-career samples, respectively) and the lack of model misfit associated with the Importance-Performance relationship suggests this relationship is fully mediated. With respect to attitudes, a strong direct negative effect of Disillusionment on Importance was supported in both samples. Interestingly, Purpose had a non-significant and practically small direct effect on Importance in the first-year student sample, but this effect reached significance in the mid-career student sample. Also, contrary to the prediction, Validity had a negligible direct effect on Importance in the first-year student sample, and slightly larger (but non-significant) direct effect on Importance in the mid-career student sample. Indirect standardized effects of all three attitudes on Performance were small and statistically non-significant in both samples.

Despite the decent global and local fit of Model 1, there were some areas of concern, as illustrated by the non-significant parameter estimates (presented in Figure 2) and a few correlation residuals. Specifically, the direct paths from SAT math and SAT verbal to Importance were statistically and practically non-significant, as foreshadowed by lack of the correlation between these variables. Recall that there was no prior research or theory regarding this direct effect of prior ability (as measured by the SAT scores) on Importance, thus the direct effects were included to control for any potential effect of ability on perceived importance of the test. The results indicate that ability does not directly affect perceived test Importance, making these direct paths unnecessary. Moreover, some of the variable relationships were underestimated by the fully-mediated model, as indicated by positive correlation residuals. Specifically, the relationship between Purpose and Effort was underestimated in both samples (correlation residual =
.15 and .19 in the first-year and mid-career student samples, respectively), indicating that a direct path, in addition to an indirect path, is needed in order to accurately model the relationship between these two variables. Although this was the largest residual in both samples, there were other sizeable correlation residuals in the mid-career student sample. Specifically, the relationships between Effort and the following variables were underestimated in the mid-career student sample: SAT math (.15), SAT verbal (.15), and Validity (.17), whereas the relationship between Effort and Disillusionment was slightly overestimated by Model 1 (correlation residual = -.16). Taken together, these results suggest that a more plausible model should have direct effects between the variables whose relationships were underestimated in Model 1, and paths found to be non-significant in Model 1 should be eliminated. Prior to using this information to modify Model 1 and propose a modified model, it is necessary to test the a priori specified partially mediated model.

**Model 2: Partially-mediated model.** This model, depicted in Figure 3, posits that the effects of attitudes towards higher education accountability tests on Performance are only partially mediated by test-taking Effort and perceived Importance, after controlling for SAT math and SAT verbal. Model 2 differs from Model 1 by simply adding a direct path from each of the three attitudes directly to Performance. As foreshadowed by the low correlations between attitudes and Performance, as well as the absence of large positive correlation residuals between these variables in Model 1 (which would indicate underestimated relationships), Model 2 did not adequately address misfit and thus did not represent the observed data better than the simpler Model 1. Expectedly so, Model 2 did not yield significantly or practically better model-data fit than Model 1 in
the first-year student sample ($\Delta \chi^2_{ML}(3) = .39, p = .942; \Delta CFI < .01$) or in the mid-career student sample ($\Delta \chi^2_{ML}(3) = 1.95, p = .583; \Delta CFI < .01$). Given the lack of utility of Model 2, parameter estimates are not presented and this model will no longer be discussed.

**Model 4: Modified fully-mediated model.** Given the adequate global fit of the fully-mediated Model 1 (Figure 2) in both samples, Model 1 was slightly modified to produce Model 4 (see Figure 5). Specifically, the following modifications were made based on the observed correlational pattern among the variables as well as correlation residuals observed in Model 1. First, direct paths from SAT math and SAT verbal to Importance were eliminated due to lack of the strong relationship between these variables. Second, a direct path was added from Purpose to Effort, in order to better estimate this relationship that was underestimated in Model 1 in both samples, as illustrated by high positive correlation residuals. The modified fully-mediated Model 4 was then fit to both first-year and mid-career student samples, and yielded good model-data fit (Table 18). Recall that in the mid-career student sample, relationships between Effort and the following variables were also underestimated when fitting Model 1: SAT math (.15), SAT verbal (.15), Validity (.17), whereas the relationship between Effort and Disillusionment was slightly overestimated (correlation residual = -.16). Models with the added direct paths between these variables were tested, but they did not fit the data statistically or practically significantly better than Model 4, and parameter estimates did not change dramatically when these direct effects were added. Conceptually, the main difference between Model 4 and Model 1 is that Model 4 posits that the relationship between Purpose and Effort is not only *indirect* (mediated via Importance) but also
Model 4 yielded good model-data fit in both samples (see Table 18) and none of the correlation residuals exceeded |.15|. Model 4 was championed in both samples.

Standardized parameter estimates of direct effects obtained in both samples are presented in Figure 5. Tables 19 and 20 also contain standardized and unstandardized estimates of direct, indirect, and total effects obtained from Model 4 in the first-year and mid-career student samples, respectively. Whereas unstandardized parameters are accompanied by significance tests, standardized parameters can be directly compared to one another, thus allowing for the assessment of the relative importance of each effect in the model. As foreshadowed by the correlational pattern and similar to Model 1, SAT verbal was the strongest predictor of Performance in both samples. Effort was the second strongest direct predictor of Performance. Similar to Model 1, Importance had a significant direct effect on Effort. Notably, the direct effect of Importance on Effort decreased slightly from Model 1 (.50/.54) to Model 4 (.41/.39), due to the added path from Purpose to Effort. However, the direct effect of Importance on Effort remained statistically and practically significant after controlling for Purpose in Model 4. Additionally, and similar to Model 1, an indirect effect of Importance on Performance (via Effort) was also large and statistically significant in Model 4 in both samples (.11/.14).

In the case of Importance-Performance relationship, an indirect effect equals the total effect because there is only a single path from Importance to Performance (through Effort). However, in other cases, an indirect effect will not equal a total effect. For example, unlike in Model 1, the total effect (.22/.32) of Purpose on Effort in Model 4 is a combination of an indirect effect (via Importance) (.04/.05) and the added direct effect
(18/27). Notably, both direct and total effects of Purpose on Effort were statistically significant and practically sizeable in both samples. That is, similar to Model 1, indirect effects of Purpose on Effort via Importance in Model 4 were not significant in both samples, whereas a direct effect of Purpose on Importance reached a significance level in the mid-career sample (.15) but did not in the first-year student sample (.10). These findings indicate that a direct path from Purpose to Importance may be unnecessary. In other words, it appears that students’ perceived understanding of the tests’ purpose directly affects their test-taking effort (which then directly affects performance), but does not directly affect how important students perceive the tests to be.

In contrast to Purpose, Disillusionment had a significant negative direct effect on Importance (-.54/.40), a significant total effect on Effort (via Importance; -.22/-16), and a significant total effect on Performance in both samples (via Importance and Effort; -.06/- .05). This pattern of results was similar in Models 1 and 4. It appears that the mediated effect of attitudes on performance via motivation was observed most strongly for Disillusionment. That is, disillusionment toward the tests directly affected how important students perceived the tests to be, which in turn affected their test-taking effort, which in turn affected performance on the test. Even though the effects were statistically significant, they were still fairly small in size. For example, the strongest unstandardized direct effect of Disillusionment on Importance in the mid-career sample signifies that for every 1 point (on a 7-point Likert scale) increase on the SAIAT-HE-revised Disillusionment scale, there is .35 point (on a 5-point Likert scale) decrease on the SOS Importance subscale.
Interestingly, Validity had a statistically non-significant and practically small direct effect on Importance in both samples (-.03/.10), and it follows that indirect effects from Validity on Effort and Performance were also trivial in size and non-significant. This pattern of results was similar in Models 1 and 4. If replicated, these results would indicate that the extent to which students view college accountability tests as fair and valid does not affect how important they perceive these tests to be, how much effort they invest in them, and, ultimately, how well they perform.

Comparison of standardized parameter estimates allows for rank-ordering relative importance of each one of the attitudes, controlling for other variables in the model. As mentioned above, Validity had the lowest effect on all variables. With respect to Performance, the total effect of Purpose on Performance was the strongest (.06/.11), followed by Disillusionment (-.06/-05), and followed by Validity (.00/.01) in both samples (the last one being practically nonexistent). It wasn’t surprising that the total effects of Purpose on Performance were the strongest in Model 4 given that Purpose has both a direct and an indirect effect on the variable that affects Performance directly – Effort, whereas the effects of the other attitudes are mediated via two variables: Effort and Importance. With respect to Effort, Purpose had a stronger total effect in the mid-career sample (.32) compared to the first-year sample (.22), and Disillusionment had a stronger total effect in the first-year student sample (-.22) compared to the mid-career sample (-.16). With respect to Importance, Disillusionment had the strongest effect in both samples (-.54/-04). The direct effect of Purpose on Importance reached statistical significance in the mid-career sample (.15), but did not in the first-year student sample.
Again, the direct effects of Validity on Importance were non-significant and trivial in both samples (-.03/.10).

Examining the percent of variance explained in the variables is another method of evaluating the utility of the model. That is, a model could fit the data well, but not explain much variance in the outcomes (e.g., Performance, Effort). Notably, a relatively high percentage of the variance in Performance (GLEX) was explained by Model 4 in both first-year and mid-career student samples: 69% and 46%, respectively, with SAT verbal being the strongest predictor in both samples (standardized estimates of .69 and .61). A smaller amount of variance, yet still large, was explained in Effort: 26% and 32% in the first-year and mid-career student samples, respectively. Moreover, 34% of variance in Importance was explained by Model 4 in both samples. Although it is likely that necessary variables that would have explained more variance in the Performance, Effort, and Importance were omitted from the current model, the explanatory utility of the variables included in Model 4 is quite high.

**Model 3: Partially-mediated model with SAIAT-K-12.** As detailed in Chapter 1, students’ attitudes toward accountability testing in K-12 are theoretically expected to influence their attitudes toward similar tests administered in college. An expanded model, including three SAIAT-K-12 attitudes (Validity, Purpose, Disillusionment), was specified a priori (see Figure 4). This partially mediated model is identical to Model 2, but also includes direct effects from the SAIAT-K-12 subscales to the corresponding SAIAT-HE subscales. Although this model produced adequate global model-data fit (see Table 18), it is evident from previous analysis that the modified fully-mediated model (Model 4 in Figure 5) with the SAIAT-K-12 variables added was more plausible. Indeed, the
partially-mediated model with SAIAT-K-12 included had similar areas of concern as Model 4: non-significant paths from SAT scores to importance and local misfit associated with Purpose and Effort. For this reason, modified Model 5 was specified and is described next.

**Model 5: Modified fully-mediated model with SAIAT-K-12.** This model, depicted in Figure 6, differs from Model 4 in that three SAIAT-K-12 attitudes (Validity K-12, Purpose K-12, and Disillusionment K-12) were added in the model and predicted to *directly* affect corresponding SAIAT-HE attitudes. Given that descriptive statistics and correlations among all the variables except for SAIAT-K-12 subscales were described for this sample earlier and presented in Table 16, only descriptives pertinent to SAIAT-K-12 subscales are outlined next.

As detailed in Research Question 4, the relationships between students’ attitudes toward accountability tests in K-12 (Validity K-12, Purpose K-12, and Disillusionment K-12) and students’ attitudes toward accountability tests in higher education (Validity HE, Purpose HE, and Disillusionment HE) were positive in direction and moderate in magnitude. Notably, the SAIAT-K-12 subscales were weakly related to Importance (ranging from .17 to .19) which is consistent with the previous finding (Zilberberg et al., in progress). The SAIAT-K-12 subscales were virtually unrelated to Effort. In the current study, two of the SAIAT-K-12 subscales were weakly and positively related to Performance: Validity (r = .10) and Purpose (r = .17), which is partially consistent with the previous study that found weak positive relationships (above .10) between all SAIAT-K-12 subscales except for the Parents subscale.
Model 5 was fit to the same sample of first-year students and yielded good model-data fit (see Table 18) and no correlation residuals exceeding |.15|. Standardized and unstandardized model parameters are presented in Table 21. Supporting the hypothesis that attitudes toward K-12 accountability tests directly predict attitudes toward higher education accountability tests, all three direct paths from K-12 to HE attitudes were positive, significant, and moderate in magnitude (standardized estimates ranging from .36 to .42). Importantly, total and indirect effects of Validity K-12 on Importance, Effort, and Performance were small and non-significant, evidencing that students’ perception of K-12 tests’ validity does not directly or indirectly affect their test-taking motivation or performance on a college accountability test. Further, a total effect of Disillusionment K-12 on Importance was statistically and practically significant (-.21), so was the total effect of Disillusionment K-12 on Effort (-.09), and the total effect of Disillusionment K-12 on Performance (-.02). These results suggest that the disillusionment that students feel towards K-12 tests indirectly (via attitudes toward college accountability tests) affects how important they perceive college accountability tests to be, which in turn affects how much effort they invest in them, which in turn affects their actual performance on such tests. With respect to Purpose K-12, its total effect on Importance was non-significant (.04), suggesting that students’ perceived understanding of the K-12 accountability tests’ purpose does not affect how important they view college tests to be. However, Purpose K-12 did have significant (though small) total effects on Effort (.08) and Performance (.02), suggesting that students’ perceived understanding of the K-12 accountability tests’ purpose indirectly (via attitudes toward college accountability tests) affect how much effort students invest in a college accountability test and, ultimately, how well they
perform on it. This significant effect of Purpose K-12 on Performance is not surprising given that this variable directly influences Purpose HE, which affects Performance through two pathways: via Importance and then Effort, and via Effort directly. Notably, adding K-12 attitudes did not substantially change the effects of attitudes toward accountability testing in higher education on Importance, Effort, and Performance specified in Model 5.

**Research Question #6**

Do students who attended their scheduled assessment session (i.e., compliant students) differ in their attitudinal profile from students who skipped the assessment session (i.e., non-compliant students)? As detailed in Chapter 1, non-compliant students tend to differ from compliant students in academic entitlement, perceived test importance, test-taking effort, and GPA (Brown & Finney, 2011; Swerdzewski et al., 2009). Furthermore, it is of interest to examine whether differences in attitudes toward higher education accountability tests exist after controlling for conscientiousness, a variable that has been repeatedly shown to be a strong positive predictor of academic success and behavior (Poropat, 2009).

Tables 22 and 23 contain descriptive statistics, inter-subscale correlations, and internal consistency indices for the SAIAT-HE-revised subscales and Conscientiousness for compliant and non-compliant mid-career students. The internal consistency of all subscales was adequate (Cronbach’s alphas > .74). The correlational pattern among the SAIAT-HE-revised subscales was similar to that previously observed and similar across compliant and non-compliant students. Specifically, Validity and Purpose were moderately and positively related in both samples ($r = .48/r = .58$); Validity and
Disillusionment were moderately and negatively related \((r = -.66/r = -.51)\); and Purpose and Disillusionment were also moderately and negatively related \((r = -.51/r = -.50)\).

Notably, Conscientiousness was unrelated to Validity \((r = .07)\) and Purpose \((r = .06)\) and weakly negatively related to Disillusionment \((r = -.14)\) in the compliant sample. In the non-compliant student sample, Conscientiousness was weakly related to Validity \((r = .12)\) and Purpose \((r = .15)\) and also weakly negatively related to Disillusionment \((r = -.12)\).

Table 24 contains mean differences on the SAIAT-HE-revised subscales across compliant and non-compliant students before and after adjusting for Conscientiousness. On average, and prior to adjusting for Conscientiousness, both compliant and non-compliant students held neutral views regarding the tests’ validity \((M = 3.51/M = 3.28)\), had some perceived understanding of the tests’ purpose \((M = 4.61/M = 4.15)\), and were somewhat to strongly disillusioned by testing \((M = 4.98/M = 5.26)\). As expected, compliant students were more conscientious \((M = 3.75)\) than their non-compliant peers \((M = 3.47)\). Prior to adjusting for Conscientiousness, non-compliant students scored somewhat lower than compliant students on Validity; this difference was not statistically or practically significant \((d = 0.19)\). Non-compliant students scored significantly and practically lower on Purpose \((d = 0.30)\) and higher on Disillusionment \((d = -0.23)\). Moreover, non-compliant students scored practically and significantly lower on Conscientiousness than compliant students \((d = 0.28)\).

The interactions between student type (compliant vs. noncompliant) Conscientiousness were non-significant and practically trivial, indicating that the relationships between Conscientiousness and attitudes are similar across compliant and
non-compliant students. The null interactions supported homogeneity of the regression slopes assumption, leading to the conclusion that Conscientiousness functions well as a covariate. Therefore, means on the SAIA-T-HE-revised subscales for two groups were *adjusted*, controlling for group differences on Conscientiousness (Table 24). As expected, the pattern of mean differences changed somewhat after taking Conscientiousness into account. Specifically, the magnitude of mean differences among all three attitudes decreased in size. The minor non-significant difference between compliant and non-compliant students on Validity became even smaller after controlling for Conscientiousness ($d = 0.15$) than prior to adjustment ($d = 0.19$). The score of non-compliant students on Purpose was still significantly lower than that of compliant students after controlling for Conscientiousness ($d = 0.25$), but lower compared to the effect observed prior to adjustment ($d = 0.25$). With respect to Disillusionment, the difference between non-compliant and compliant students ceased to be statistically significant after controlling for Conscientiousness and diminished in practical significance ($d = -0.15$ compared to $d = -0.23$ prior to adjustment).
Discussion

The current program of research focused on the measurement of college students’ attitudes toward institutional accountability testing in higher education and implications of such attitudes on test score validity. Consisting of two stages, this research program focused on the psychometric properties and validity of the self-report measure designed to assess students’ attitudes toward accountability testing and a substantive investigation of such attitudes in relation to other variables. The discussion of the results is organized in the following way. First, the results of the research questions subsumed under Stage I are discussed and conclusions regarding the use of the scale are made. Second, the results of the research questions subsumed under Stage II are discussed and supported with connections to the theoretical predictions and existing literature. In addition, the results from both stages are used to make recommendations to higher education assessment professionals interested in improving validity of low-stakes test scores by boosting test-taking motivation. Finally, limitations of the current studies and suggestions for future research studies are outlined.

Stage I

Stage I focused on garnering validity evidence for the 22-item self-report measure designed to assess college students’ attitudes towards institutional accountability testing administered in higher education (SAIAT-HE). Three research questions subsumed under Stage I pertained to 1) the psychometric properties of the SAIAT-HE for first-year and mid-career students and resulting scale revisions; 2) invariance of the measure across two student groups; and 3) known-groups validity evidence. The implications of the findings from each research question are discussed next.
The psychometric properties and dimensionality of the 22-item SAIAT-HE, in conjunction with the previous analysis of the SAIAT-HE using mid-career students, guided scale revisions and modifications of the factor structure. Based on these results, three factors were omitted from the scale due to low proportion of variance accounted for in the corresponding items: Parents, Professors, and Students. In addition, two items (1 and 8) exhibited poor psychometric properties in both samples and were thus omitted from the scale. The resulting SAIAT-HE-revised consists of three inter-related factors (Validity, Purpose, and Disillusionment) and 10 items. Each of the three subscale scores had adequate internal consistency in both samples, and all items functioned well. These findings have the following implications for research and practice. The SAIAT-HE-revised is a psychometrically sound self-report measure of students’ attitudes toward accountability testing in higher education that yields three reliable subscale scores. When scoring the measure, three subscale scores should be computed: one for each attitude. Computing the total score is not appropriate for this measure. Even though the subscales are inter-related, they are conceptually distinct, thus a total score would mask differential relationships with external criteria.

The extent to which the SAIAT-HE-revised is invariant across first-year and mid-career students was evaluated. Specifically, configural, metric, and scalar invariance was upheld indicating that attitudes towards higher education accountability testing are conceptualized equivalently and the SAIAT-HE-revised functions invariantly across first-year and mid-career students. The implications of full measurement invariance of the SAIAT-HE-revised are two-fold. First, the measure can be used with both first-year and mid-career students to examine research questions pertaining to the relationship between
such attitudes and other variables of interest (e.g., performance on an accountability test). Second, differences in the level of these attitudes across first-year and mid-career students can be meaningfully evaluated using the SAIAT-HE-revised, thereby allowing for the collection of known-groups validity evidence for the scale, in addition to answering substantive questions regarding differences in attitudes across the two student populations.

Differences between first-year and mid-career students were investigated, thus providing known-groups validity evidence. As predicted, mid-career students perceived accountability testing in higher education to be less valid than first-year students. Also, as predicted, mid-career students were more disillusioned by accountability testing than their first-year peers. Contrary to the hypothesis, mid-career students’ perceived understanding of the tests’ purpose was lower than that of the first-year students, although this difference was quite small. Despite the fact that the direction of this mean difference was not predicted a priori, this finding is not necessarily indicative of lack of external validity. Rather, this finding may suggest that this group of first-year students was slightly more informed about the purposes of the tests than the group of mid-career students. If that is the case, and the same group of first-year students is reassessed at the mid-point of their college career, their perceived understanding of the tests’ purpose would be predicted to either remain the same or increase. True longitudinal designs are needed to fully interpret the meaning of the group difference uncovered for the Purpose subscale.

Nonetheless, current findings provide important known-groups validity evidence for the SAIAT-HE-revised scores. Moreover, the results of the latent means analysis
demonstrated that students’ attitudes toward low-stakes accountability assessments follow a downward spiral, aligning with theoretically supported predictions (Olsen & Wilson, 1991; Paris et al., 2000; Zilberberg et al., in press). Assuming that mid-career and first-year students differ in their attitudes due to differences in age and academic levels, a conclusion can be made that something occurs over the course of the first two years spent in college that worsens students’ attitudes toward accountability testing. The implication of this finding is that an intervention aimed at improving students’ attitudes is needed and should occur sometime between students’ first year in college and the midpoint of their academic career. Similar interventions, aimed at framing good test-taking behavior as a service to university, have been proposed in the past (e.g., Zilberberg et al., 2009; Wise & Cotten, 2009).

Taken together, results from Stage I indicate that the SAIAT-HE-revised reliably assesses three dimensions of students’ attitudes toward accountability testing in higher education: Validity (i.e., perceived accuracy and fairness of accountability tests), Purpose (i.e., perceived understanding of the tests’ purpose) and Disillusionment (i.e., dissatisfaction with the tests). Therefore, the SAIAT-HE-revised can be used to address substantive research questions pertaining to the attitudes of American college students toward institutional accountability testing in higher education.

Stage II

Stage II focused on addressing substantive research questions related to students’ attitudes toward accountability testing in higher education, as measured by the SAIAT-HE-revised. Three research questions subsumed under Stage II pertained to 1) the relationship between students’ attitudes toward K-12 accountability testing and their attitudes toward testing with the similar purpose in college; 2) the nature of the
relationship between students’ attitudes, test-taking motivation, and performance on the test; and 3) differences in attitudinal profiles among compliant and non-compliant students. The implications of the findings corresponding to each one of the aforementioned research questions are discussed next.

It was hypothesized that incoming college students would have had developed a cognitive schema of “accountability testing” over the course of K-12 and view college accountability testing through the lens of their prior experience. This hypothesis was somewhat supported by moderate positive relationships between corresponding SAIAT-K-12 and SAIAT-HE subscales. That is, the extent to which incoming college students perceived K-12 accountability tests to be valid, understood their purpose, and were disillusioned was positively related to their views on accountability testing in higher education. However, these relationships were only moderate in strength, indicating that the two sets of attitudes are conceptually distinct. Despite the similarities between accountability testing in K-12 and college (e.g., large-scale, low-stakes, use of test scores for measuring institutional effectiveness), it appears that the differences between the academic context (college vs. high-school) meaningfully differentiated the two sets of students’ attitudes. Thus, when designing interventions to boost test-taking motivation on low-stakes college accountability tests, it is useful to know students’ attitudes toward K-12 testing, but attitudes toward K-12 testing cannot be substituted for attitudes toward higher education accountability testing (i.e., they are not interchangeable).

The examination of the nature of the relationships between students’ attitudes towards accountability testing (both in K-12 and higher education), test-taking motivation, and test performance was very informative. As foreshadowed by the answer
to the previous research question, students’ attitudes toward K-12 tests moderately affected their corresponding attitudes toward higher education tests (Figure 6). The indirect effects of attitudes toward K-12 testing on performance on a college accountability test were small and practically non-significant. Notably, students’ perception of the validity of K-12 did not affect students’ perceived importance of college accountability tests, test-taking effort, or performance. The extent to which students were disillusioned by K-12 testing did have significant, though small, indirect effects on how important they perceived college accountability tests to be, how much effort they invested, and how well they performed. The extent of students’ perceived understanding of K-12 tests’ purpose did not have a significant indirect effect on how important they perceived college accountability tests to be, but it did have a significant, though small, indirect effect on test-taking effort and performance. Most importantly, the relationships between attitudes toward K-12 accountability testing and performance on college accountability testing were fully mediated by student attitudes toward college accountability testing. The main implication of this finding is that the focus of any intervention at the higher education level should be on improving students’ attitudes toward college accountability tests, not K-12 accountability tests.

Pertaining to the relationships among attitudes toward college accountability tests, perceived importance of the test, test-taking effort, and actual performance, a modified mediated model had the most utility, both in the first-year and mid-career student samples (Figure 5). Several components of this model warrant further discussion. First, test-taking effort had a consistently significant and sizeable direct effect on performance which is congruent with the expectancy-value theory (Wigfield & Eccles, 2000) and prior
empirical findings (Thelk, 2006; Wise & Kong, 2005). Second, the relationship between perceived importance of the test and test performance was fully mediated by test-taking effort. That is, the extent to which students view accountability tests as important directly affects how much effort they invest in such tests, which, in turn, directly affects how well they perform. This finding aligns with the theoretical prediction of how test-taking motivation affects performance in low-stakes testing conditions (Sundre & Kitsantas, 2004; Wise & DeMars, 2005). Therefore, assessing both test-taking effort and perceived importance is crucial when examining factors influencing performance on college accountability tests. In addition, and as expected, verbal and mathematical ability were directly related to performance but unrelated to perceived importance of the test or test-taking effort, aligning with prior literature (Robbins et al., 2004; Steedle, 2010; Wise, 2009; Wise & DeMars, 2005).

With respect to attitudes, some of the results were intriguing. Contrary to the hypothesis, the extent to which students perceived college accountability tests to be valid and fair did not affect how important students perceived the test to be, how much effort they expended on it, or their performance. This finding contradicts that of Sambell and his colleagues who found that “test fairness” was a determinant of perceived importance of the test (1997), but aligns with the more recent finding that college students’ perception of K-12 tests’ validity does not affect their performance on college accountability tests (Zilberberg et al., in progress). Notably, this lack of relationship between perceived validity of the test and perceived importance was stable across first-year and mid-career student populations. The main implication of this finding is if one’s
goal is to boost students’ test-taking motivation, arguing for psychometric soundness of the test is futile.

Perhaps the most interesting finding pertained to students’ perceived understanding of the tests’ purpose. Contrary to the hypothesis, the direct effect of perceived understanding of purpose of the test on the perceived importance of the test was rather small in both student samples. That is, students who reported a greater perceived understanding of the tests’ purpose were only slightly more likely to perceive these tests as more important. However, students who reported a greater perceived understanding of the tests’ purpose were more likely to report investing more test-taking effort in such tests, and thus scored higher on the test. That is, there was an indirect effect of perceived understanding of purpose on performance, but this relationship was mediated via test-taking effort, not via both importance and effort, as hypothesized. As such, perceived understanding of the tests’ purpose and test performance were more strongly related than originally hypothesized. Thus, addressing students’ perceived understanding of the tests’ purpose is likely to boost test-taking effort, and, in turn, increase performance. The most straightforward way of achieving this goal is via clearly communicating to students the purpose of accountability testing, which has been repeatedly suggested by many researchers (e.g., Brown & Gaxiola, 2010; Olsen & Wilson, 1991; Swing, 2001; Zilberberg et al., 2009).

Only one of the attitudes – disillusionment – had a mediated relationship with performance through perceived importance and effort, aligning with the theoretical hypothesis (Paris et al., 1990; 1991; 2000). That is, higher disillusionment with respect to accountability testing led to lower perceived importance of the test, which, in turn,
negatively affected test-taking effort, which ultimately biased performance on the test downward. Notably, the level of disillusionment did not affect test-taking effort or performance directly, suggesting that the mediated relationship was accurately specified. Thus, decreasing students’ levels of dissatisfaction with college accountability tests is likely to improve test performance, not directly, but through increasing perceived importance of the tests and test-taking effort. Together, these results emphasize the necessity to measure test-taking motivation when studying the impact of attitudes in low-stakes accountability testing. The current studies modeled this mediating effect of test-taking motivation, thereby advancing prior research of students’ attitudes toward testing that did not include test-taking motivation (e.g., Brown & Hirschfeld, 2008).

Contrasting the attitudinal profile of compliant students (i.e., those who attend their scheduled assessment session) to that of non-compliant students (i.e., those who attend a make-up session instead) offers more insight into how attitudes toward accountability testing in higher education differ among various groups of students. Non-compliant students tend to be male, slightly older, less conscientious, have a lower GPA, perceive accountability tests to be less important, and invest less test-taking effort (Swerdzewski et al., 2009). Moreover, non-compliant students tend to be more academically entitled and more psychologically reactant than compliant students (Brown & Finney, 2011; Kopp & Finney, in press). Controlling for conscientiousness, non-compliant students reported significantly lower perceived understanding of the tests’ purpose than compliant students. Aligning with the previous finding in which perceived understanding of purpose was directly related to actual test-taking behavior (test-taking effort), it appears that perceived understanding of purpose is also predictive of another
test-related behavior – attending the scheduled assessment session. Differences among compliant and non-compliant students on disillusionment, after controlling for conscientiousness, were minimal. Moreover, compliant and non-compliant students did not differ in their perception of tests’ validity. The implications of this finding reinforce a previously discussed suggestion to focus on communicating the purpose of college accountability testing to students, as it is most likely to promote good test-taking behavior such as attendance and maximal effort (e.g., Brown & Finney, 2011).

Limitations of the Current Program of Research and Directions for Future Studies

The current program of research has several limitations that could not be surmounted due to sample availability and other factors. First, direct inferences regarding the change in attitudes over the course of a college career could not be made due to the cross-sectional nature of the design. In a true longitudinal design, it would be possible to assess attitudinal change among the same students across time. Having a third measurement point (during their senior year in college) would allow for an even more robust evaluation of change over time. Such a longitudinal design would enable an empirical assessment of whether the hypothesized downward spiral of attitudes really exists (Paris et al., 2000; Wong & Paris, 2000). Having data on multiple occasions from the same sample would also allow for testing longitudinal invariance of the SAIAT-HE-revised, which is a necessary assumption for testing change over time. Unfortunately, only independent samples of first-year and mid-career students were available for the current studies. Therefore, all observed differences can be attributed to any differences across samples, not only to the variation in students’ academic level.
Second, and related to the limitations associated with the available samples, the time of testing and students’ academic level were confounded, complicating the task of attributing the observed effects to students’ academic level. That is, mid-career students were assessed in the Spring whereas first-year students were assessed in the Fall. Time of year could affect attitudes and test-taking motivation, but isolating this effect was not possible in the current studies. Similarly, it was not feasible to ascertain whether the deterioration of attitudes was due to repeated testing or maturation effects. Designing a study with random assignment and a control group would address these issues.

Third, the explanatory utility of the mediated models (Figures 5 and 6) needs to be addressed. That is, a substantial amount of variability in test-taking effort (over 65%) remained unexplained, indicating there are variables omitted from the model that could explain why students vary in the amount of effort they invest in college accountability tests. For example, academic citizenship, or students’ willingness to help improve the performance of their institution, could be another factor affecting test-taking effort (Wise, 2009). Conscientiousness, a variable shown to be predictive of academic success (O’Connor & Paunonen, 2007; Poropat, 2009) and the quality of expended effort on various academic and personal activities (Bauer & Liang, 2003) is another factor that may influence test-taking effort. It has been suggested that personal pride or a tendency to try one’s best on a task regardless of consequences is characteristic of high-achieving students and explains why some students invest more effort than others (McGuinness, 2010). An alternative model, specified based on the current results and including omitted variables such as these, may have more explanatory utility for predicting test-taking effort, and, ultimately, performance.
Fourth, it may be of interest to examine how attitudes differ for male and female students. We know that, in general, males are socialized to engage in more independent behaviors and thoughts, whereas females are socialized to express socially approved behaviors and attitudes that are expected of them (Guttentag & Bray, 1976). Karrmos and Karmos studied gender differences in attitudes toward standardized school accountability testing among 6-9th graders and found that females had more positive attitudes than boys (e.g., scored higher on “It’s important for me to do well” and lower on items such as “Tests are a waste of time”) (1984). Given the current finding that non-compliant students, who tend to be male, have worse attitudes than compliant students, it would be informative to examine whether attitudes differ depending on a student’s gender. This information could potentially aid in developing more tailored interventions.

Fifth, the extent to which current results can generalize to other higher education institutions needs to be carefully considered. The “Assessment Day” model, in which a full day is designated for accountability assessments, is fairly uncommon. It is a unique model that reflects a deep institutional commitment to assessment, which involves trained proctors, several e-mail reminders sent out to students, and an informational video shown on the day of testing. At the institution where the current studies were conducted, a culture of assessment is strong and long-standing, spanning more than two decades. At universities with a different set of values and assessment history, students’ attitudes toward testing may differ drastically. Moreover, results may vary if a different accountability test (e.g., a music or math test) is used as a measure of performance.

In summary, future studies are likely to advance the current work and allow for a more rigorous and informed investigation of the role of students’ attitudes on the validity
of accountability test scores. Although the studies in the current program of research have their limitations, they make an important contribution to the growing body of research related to non-cognitive factors influencing test performance.

**General Conclusion**

Thirty years after *The Nation at Risk* touched off a wave of educational reforms in America, the progress made in the area of direct assessment of students’ learning outcomes is evident. At the higher education level, the voluntary efforts to boost the validity of accountability assessment by employing psychometrically sound measures are becoming widespread. In addition to considering the psychometric integrity of accountability tests, it is becoming imperative to also consider the psychology of the examinees and various non-cognitive factors affecting students’ test-taking motivation and performance on accountability tests (e.g., Sundre & Kitsantas, 2004). The current program of research laid the foundation for the exploration of college students’ attitudes toward accountability testing in higher education and the role of such attitudes in building a validity argument for accountability testing.
Appendix A

Students’ Attitudes toward Institutional Accountability Testing in Higher Education

(SAIAT-HE)

Directions: Below are a series of questions related to assessments completed throughout your college or university experience. Assessments are defined as the tests you completed as Freshman on Assessment Day, the Information Seeking Skills Tests (ISST) and Tech1 tests, and the tests you are completing here today.

When completing the items below, think only of these types of tests. Do not think of college entrance tests (e.g., SAT) or tests for your courses.

1 = strongly disagree
2 = moderately disagree
3 = slightly disagree
4 = neither agree nor disagree
5 = slightly agree
6 = moderately agree
7 = strongly agree

VALIDITY

1. Assessment tests are unfair to some students (R).

2. Assessment test results are not accurate (R).

3. Assessment test results accurately reflect basic skills and knowledge of a subject.

4. Assessment tests are not valid (R).

5. Assessment test scores don't reflect my true ability (R).

PURPOSE

6. I don’t understand how assessment tests are related to my education (R).

7. I don’t understand the need for assessment tests (R).

8. Someone (professor, academic advisor, Resident Advisor) explained to me why I take assessment tests.
9. I understand the purpose of assessment tests

DISillusionMENT

10. The more assessment tests I complete, the more I dislike assessment tests.

11. There is too much assessment testing.

12. Assessment tests are a waste of my time.

PARENTS

13. My parents would be disappointed if I performed poorly on the assessment tests.

14. My parents don't value the assessment tests I complete at the University (R).

15. My parents would be proud of me if I performed well on the assessment tests.

16. My parents are unaware of the assessment tests I complete at the University (R).

PROFESSOR

17. My professor(s), an academic advisor, or resident advisor encouraged me to prepare for the assessment tests.

18. My professors don't value the assessment tests I complete at the University (R).

19. If I performed poorly on the assessment tests, my professors would be disappointed.

STUDENTS

20. Fellow students urged me to try my best on the assessment tests.

21. Fellow students discouraged me from taking the assessment tests seriously (R).

22. Fellow students speak negatively about the assessment tests at the University (R).

Note: “R” indicates reverse coded items. Items were randomized when administered to respondents.
Appendix B

Students’ Attitudes toward Institutional Accountability Testing in K-12

(SAIAT-K-12)

Directions: Below are a series of questions related to standardized testing. Here standardized testing is defined as tests you completed as measures of your progress through school (kindergarten – 12th grade). These tests were created to meet the mandates of No Child Left Behind. For example, in the Commonwealth of Virginia, these are called the SOL (Standard of Learning) tests. These tests were often not used to calculate GPA.

When completing the items below, think only of these types of tests (K – 12 standardized tests). Do not think of tests like the PSAT, SAT, or ACT used for the college admission process or AP tests.

1 = strongly disagree
2 = moderately disagree
3 = slightly disagree
4 = neither agree nor disagree
5 = slightly agree
6 = moderately agree
7 = strongly agree

VALIDITY

1. Standardized tests are unfair to some students (R).
2. Standardized tests results are not accurate (R).
3. Standardized test results accurately reflect basic skills and knowledge of a subject.
4. Standardized tests are not valid (R).
5. Standardized test scores don’t reflect my true ability (R).

PURPOSE

6. I don’t understand how standardized tests are related to my education (R).
7. I don’t understand the need for standardized tests (R).

8. Someone (teacher, principal, guidance counselor) explained to me why I take standardized tests.

9. I understand the purpose of standardized tests.

DISILLUSIONMENT

10. The more standardized tests I complete, the more I dislike standardized tests.

11. There is too much standardized testing.

PARENTS

12. My parents would be disappointed if I performed poorly on a standardized test.

13. My parents encouraged me to prepare for standardized tests.

14. My parents didn’t value the standardized tests I completed in school (R).

15. My parents were proud of me when I performed well on standardized tests.

16. My parents were unaware of the standardized tests I completed in school (R).

7 “R” indicates reverse coded items. Items were randomized when administered to respondents.

2 Items are presented by subscale; however, items were randomized prior to administering the test to participants.
Appendix C

Student Opinion Survey (SOS)

Please think about the tests that you just completed. Mark the answer using the 1-5 point scale that best represents how you feel about statements 1 through 10 below.

1 = strongly disagree
2 = disagree
3 = neutral
4 = agree
5 = strongly agree

IMPORTANCE

1. Doing well on these tests was important to me.
2. I am not curious about how I did on these tests relative to others. R
3. I am not concerned about the scores I receive on these tests. R
4. These were important tests to me.
5. I would like to know how well I did on these tests.

EFFORT

6. I engaged in good effort throughout these tests.
7. I gave my best effort on these tests.
8. While taking these examinations, I could have worked harder on them. R
9. I did not give these tests my full attention while completing them. R
10. While taking these tests, I was able to persist to completion of the tasks.

Note: “R” indicates reverse coded items. Items were randomized when administered to respondents.
Appendix D

Big Five – Conscientiousness subscale

The following items contain a number of characteristics that may or may not apply to you. Please select the number from 1 to 5 that indicates the extent to which you disagree or agree with each statement. Please take your time and answer thoughtfully.

1 = strongly disagree
2 = disagree
3 = neutral
4 = agree
5 = strongly agree

I See Myself as Someone Who...

1. Does a thorough job.
2. Can be somewhat careless. R
3. Is a reliable worker.
4. Tends to be disorganized. R
5. Tends to be lazy. R
6. Perseveres until the task is finished.
7. Does things efficiently.
8. Makes plans and follows through with them.
9. Is easily distracted. R

*Note:* “R” indicates reverse coded items. Items were randomized when administered to respondents.
References


Table 1

Factors-Item Map for Two Models and Two Versions of the SAIAT-HE

<table>
<thead>
<tr>
<th>Six-factor model and SAIAT-HE</th>
<th>Three-factor model and SAIAT-HE-revised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>Items</td>
</tr>
<tr>
<td>Validity</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>Purpose</td>
<td>6, 7, 8, 9</td>
</tr>
<tr>
<td>Disillusionment</td>
<td>10, 11, 12</td>
</tr>
<tr>
<td>Parents</td>
<td>13, 14, 15, 16</td>
</tr>
<tr>
<td>Professors</td>
<td>17, 18, 19</td>
</tr>
<tr>
<td>Students</td>
<td>20, 21, 22</td>
</tr>
<tr>
<td>Total number of items</td>
<td>22</td>
</tr>
</tbody>
</table>

Note. All factors are inter-related. For a complete list of the SAIAT-HE items, see Appendix A.
Table 2

Sample Sizes

<table>
<thead>
<tr>
<th>Measures</th>
<th>Incoming</th>
<th>First-year</th>
<th>Matched incoming and first-year</th>
<th>Mid-career A-Day</th>
<th>Mid-career A-Day 1</th>
<th>Mid-career A-Day 2</th>
<th>Mid-career calibration</th>
<th>Mid-career validation</th>
<th>Mid-career make-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAIAT-K12</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SAIAT-HE</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SOS</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SAT</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GLEX</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Big Five</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Collection Date(s) and Modes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Cleaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Data (N)</td>
<td>4286</td>
<td>894</td>
<td>746</td>
<td>3165</td>
<td>3165</td>
<td>184</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete data on SAIAT-K12 (N)</td>
<td>4194</td>
<td>N/A</td>
<td>678</td>
<td>N/A</td>
<td>N/A</td>
<td>169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete data on SAIAT-HE (N)</td>
<td>N/A</td>
<td>857</td>
<td>678</td>
<td>2765</td>
<td>2765</td>
<td>169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete data on all measures (N)</td>
<td>4194</td>
<td>857</td>
<td>678</td>
<td>2765</td>
<td>758</td>
<td>169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final sample size for data screening (N)</td>
<td>4194</td>
<td>857</td>
<td>678</td>
<td>2765</td>
<td>758</td>
<td>300</td>
<td>857</td>
<td>850</td>
</tr>
<tr>
<td>Data Screening</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outliers removed (N)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Final Sample Size (N)</td>
<td>856</td>
<td>678</td>
<td>758</td>
<td>300</td>
<td>857</td>
<td>850</td>
<td>169</td>
<td></td>
</tr>
</tbody>
</table>

Note. 1 - Computer-Based First-Year Survey; 2 - Assessment Day Paper and Pencil
### Table 3

Alignment Between Samples, Measures, and Research Questions

<table>
<thead>
<tr>
<th>Measures</th>
<th>Incoming</th>
<th>First-year</th>
<th>Matched incoming and first-year</th>
<th>Mid-career A-Day</th>
<th>Mid-career A-Day 1</th>
<th>Mid-career A-Day 2</th>
<th>Mid-career calibration</th>
<th>Mid-career validation</th>
<th>Mid-career make-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAIAT-K12</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SAIAT-HE</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOS</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>G&amp;EX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Big Five</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Final N</td>
<td>4194</td>
<td>856</td>
<td>678</td>
<td>2765</td>
<td>758</td>
<td>380</td>
<td>857</td>
</tr>
</tbody>
</table>

**Research Questions Analysis**

1. What revisions to the SAIAT-HE are necessary based on results from first-year and mid-career student samples?
   - Confirmatory Factor Analysis (CFA)  
   - X

2. Does the SAIAT-HE function invariantly for first-year students and mid-career students?
   - Structural Means Modeling (SSM)  
   - X

3. Do mid-career students have worse attitudes about accountability tests than first-year students?
   - Bivariate correlations  
   - X

4. How strongly are first-year students' attitudes toward K-12 accountability tests related to their attitudes toward higher education accountability tests?
   - Path Analysis  
   - Model 1: Fully mediated model  
   - Model 2: Partially-mediated model  
   - Model 3: Partially-mediated model with the SAIAT-K12  
   - X

5. What is the nature of the relationship between students’ attitudes towards accountability testing, test-taking motivation, and performance on higher education accountability test?
   - Descriptives and ANCOVA  
   - X

6. Do students who skip versus attend A-Day have different SAIAT-HE score profiles?
   - X

*Note:* 1 - Computer-Based First-year survey; 2 - Assessment Day Paper and Pencil. Final sample sizes to be used in the analysis are bolded.
<table>
<thead>
<tr>
<th>Sample</th>
<th>First-year</th>
<th>Matched incoming and first-year</th>
<th>Mid-career</th>
<th>Mid-career</th>
<th>Mid-career</th>
<th>Mid-career</th>
<th>Mid-career</th>
<th>Mid-career</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td></td>
<td></td>
<td>A-Day 1</td>
<td>A-Day 2</td>
<td>Calibration</td>
<td>Validation</td>
<td>Make-up</td>
</tr>
<tr>
<td>White</td>
<td>735</td>
<td>85.76%</td>
<td>534</td>
<td>78.76%</td>
<td>552</td>
<td>72.82%</td>
<td>222</td>
<td>74.00%</td>
</tr>
<tr>
<td>Black</td>
<td>44</td>
<td>5.13%</td>
<td>26</td>
<td>3.83%</td>
<td>32</td>
<td>4.22%</td>
<td>13</td>
<td>4.33%</td>
</tr>
<tr>
<td>Latino/Hispanic</td>
<td>52</td>
<td>6.07%</td>
<td>34</td>
<td>5.01%</td>
<td>35</td>
<td>4.62%</td>
<td>10</td>
<td>3.33%</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>76</td>
<td>8.87%</td>
<td>46</td>
<td>6.78%</td>
<td>55</td>
<td>7.26%</td>
<td>10</td>
<td>3.33%</td>
</tr>
<tr>
<td>American Indian/Alaskan</td>
<td>12</td>
<td>1.40%</td>
<td>3</td>
<td>0.44%</td>
<td>5</td>
<td>0.66%</td>
<td>4</td>
<td>1.33%</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0.00%</td>
<td>33</td>
<td>4.87%</td>
<td>43</td>
<td>5.67%</td>
<td>40</td>
<td>13.33%</td>
</tr>
<tr>
<td>Not Specified</td>
<td>19</td>
<td>2.22%</td>
<td>2</td>
<td>0.29%</td>
<td>36</td>
<td>4.75%</td>
<td>1</td>
<td>0.33%</td>
</tr>
</tbody>
</table>

| Male                       | 320        | 37.34% | 240    | 35.40%     | 257        | 33.91%     | 117        | 39.00%     | 313        | 36.52%     | 346        | 40.71%     | 105        | 62.13%     |
| Female                     | 537        | 62.66% | 438    | 64.60%     | 501        | 66.09%     | 182        | 60.67%     | 536        | 62.54%     | 500        | 58.82%     | 64         | 37.87%     |
| Not Specified              | 0          | 0.00%  | 0      | 0.00%      | 0          | 0.00%      | 1          | 0.33%      | 8          | 0.93%      | 4          | 0.47%      | 0          | 0.00%      |

*Note.* Demographic statistics were calculated for samples after data cleaning and screening, but prior to outlier removal. A-Day - Assessment Day.
Table 5

Correlations and Descriptive Statistics for the SAIAT-HE Items for First-year Student Sample

| Item | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  | 21  | 22  |
|------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1    | 1  |    |    |    |    |    |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 2    | .357 | 1  |    |    |    |    |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 3    | .171 | .253 | 1  |    |    |    |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 4    | .388 | .661 | .293 | 1  |    |    |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 5    | .268 | .387 | .351 | .367 | 1  |    |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 6    | .283 | .300 | .247 | .486 | .238 | 1  |    |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 7    | .278 | .347 | .251 | .492 | .265 | .690 | 1  |    |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 8    | .154 | .136 | .125 | .216 | .065 | .330 | .367 | 1  |    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 9    | .152 | .176 | .179 | .282 | .105 | .467 | .501 | .370 | 1  |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 10   | -0.152 | -0.232 | -0.190 | -0.271 | -0.266 | -0.217 | -0.244 | -0.095 | -0.153 | 1  |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 11   | -0.275 | -0.307 | -0.233 | -0.407 | -0.266 | -0.356 | -0.412 | -0.210 | -0.269 | -0.474 | 1  |     |     |     |     |     |     |     |     |     |     |     |     |
| 12   | -0.254 | -0.463 | -0.284 | -0.568 | -0.315 | -0.501 | -0.563 | -0.224 | -0.397 | -0.419 | -0.536 | 1  |     |     |     |     |     |     |     |     |     |     |     |
| 13   | -0.065 | 0.038 | 0.109 | 0.088 | 0.028 | 0.109 | 0.127 | 0.021 | 0.171 | -0.113 | -0.135 | -0.189 | 1  |     |     |     |     |     |     |     |     |     |     |
| 14   | 0.072 | 0.211 | 0.136 | 0.281 | 0.082 | 0.258 | 0.248 | 0.106 | 0.140 | -0.211 | -0.302 | -0.417 | -0.377 | 1  |     |     |     |     |     |     |     |     |     |
| 15   | 0.007 | 0.083 | 0.132 | 0.127 | 0.010 | 0.168 | 0.182 | 0.142 | 0.217 | -0.072 | -0.125 | -0.235 | -0.475 | -0.439 | 1  |     |     |     |     |     |     |     |     |
| 16   | 0.043 | 0.053 | 0.031 | 0.049 | 0.024 | 0.059 | 0.038 | 0.055 | 0.071 | -0.054 | -0.093 | -0.116 | 0.099 | 0.258 | -0.153 | 1  |     |     |     |     |     |     |     |
| 17   | -0.041 | 0.044 | 0.073 | 0.051 | 0.080 | 0.046 | 0.058 | 0.065 | 0.152 | -0.163 | -0.081 | -0.193 | 0.296 | 0.235 | 0.164 | -0.089 | 1  |     |     |     |     |     |
| 18   | 0.153 | 0.278 | 0.106 | 0.306 | 0.143 | 0.269 | 0.283 | 0.184 | 0.168 | -0.132 | -0.197 | -0.343 | 0.082 | 0.298 | 0.166 | 0.113 | 0.107 | 1  |     |     |     |
| 19   | -0.116 | -0.022 | 0.038 | 0.045 | 0.043 | 0.042 | 0.049 | 0.022 | 0.062 | -0.082 | -0.034 | 0.126 | 0.477 | 0.249 | 0.278 | 0.108 | 0.417 | 0.169 | 1  |     |     |
| 20   | 0.063 | 0.052 | 0.134 | 0.095 | 0.022 | 0.114 | 0.100 | 0.102 | 0.117 | -0.122 | -0.088 | 0.155 | 0.252 | 0.211 | 0.165 | 0.185 | 0.307 | 0.119 | 0.271 | 1  |     |
| 21   | 0.126 | 0.104 | 0.096 | 0.215 | 0.055 | 0.173 | 0.229 | 0.067 | 0.146 | -0.112 | -0.181 | -0.244 | 0.008 | 0.192 | 0.089 | 0.054 | -0.001 | 0.159 | 0.015 | 0.100 | 1  |     |
| 22   | 0.128 | 0.153 | 0.119 | 0.216 | 0.130 | 0.146 | 0.189 | 0.071 | 0.155 | -0.245 | -0.327 | -0.299 | 0.040 | 0.237 | 0.095 | 0.086 | 0.038 | 0.179 | 0.052 | 0.101 | 0.439 | 1  |

Note. Responses range from 1 (strong disagreement) to 7 (strong agreement). Items 1, 2, 4, 5, 6, 7, 14, 16, 18, 21, 22 had been reverse-scored. The sample does not include outliers. N = 856.
Table 6
Correlations and Descriptive Statistics for the SAIAT-HE Items for Mid-career Calibration Student Sample

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.279</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.161</td>
<td>.366</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.323</td>
<td>.665</td>
<td>.378</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.315</td>
<td>.421</td>
<td>.411</td>
<td>.413</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.337</td>
<td>.340</td>
<td>.275</td>
<td>.447</td>
<td>.273</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.261</td>
<td>.351</td>
<td>.317</td>
<td>.442</td>
<td>.281</td>
<td>.699</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.126</td>
<td>.153</td>
<td>.191</td>
<td>.171</td>
<td>.107</td>
<td>.188</td>
<td>.180</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.173</td>
<td>.257</td>
<td>.233</td>
<td>.291</td>
<td>.139</td>
<td>.505</td>
<td>.508</td>
<td>.215</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-.148</td>
<td>-.349</td>
<td>-.270</td>
<td>-.378</td>
<td>-.402</td>
<td>-.306</td>
<td>-.347</td>
<td>-.088</td>
<td>-.235</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-.165</td>
<td>-.328</td>
<td>-.256</td>
<td>-.351</td>
<td>-.308</td>
<td>-.349</td>
<td>-.376</td>
<td>-.106</td>
<td>-.268</td>
<td>.511</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-.172</td>
<td>-.479</td>
<td>-.314</td>
<td>-.568</td>
<td>-.359</td>
<td>-.437</td>
<td>-.479</td>
<td>-.359</td>
<td>-.302</td>
<td>-.535</td>
<td>.492</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>.006</td>
<td>.106</td>
<td>.212</td>
<td>.158</td>
<td>.178</td>
<td>.059</td>
<td>.087</td>
<td>.062</td>
<td>.119</td>
<td>-.247</td>
<td>-.171</td>
<td>-.251</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-.057</td>
<td>.165</td>
<td>.199</td>
<td>.172</td>
<td>.096</td>
<td>.138</td>
<td>.183</td>
<td>.119</td>
<td>.185</td>
<td>-.162</td>
<td>-.190</td>
<td>-.270</td>
<td>.417</td>
<td>.364</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>.030</td>
<td>.124</td>
<td>.072</td>
<td>.097</td>
<td>.153</td>
<td>.096</td>
<td>.081</td>
<td>.021</td>
<td>-.031</td>
<td>-.130</td>
<td>-.140</td>
<td>-.123</td>
<td>.104</td>
<td>.237</td>
<td>.106</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>.005</td>
<td>.107</td>
<td>.226</td>
<td>.124</td>
<td>.212</td>
<td>.076</td>
<td>.066</td>
<td>.192</td>
<td>.049</td>
<td>-.202</td>
<td>-.084</td>
<td>-.159</td>
<td>.224</td>
<td>.145</td>
<td>.116</td>
<td>.083</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>.016</td>
<td>.126</td>
<td>.229</td>
<td>.165</td>
<td>.205</td>
<td>.103</td>
<td>.149</td>
<td>.127</td>
<td>.114</td>
<td>-.248</td>
<td>-.154</td>
<td>-.226</td>
<td>.514</td>
<td>.232</td>
<td>.281</td>
<td>.061</td>
<td>.347</td>
<td>.229</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-.031</td>
<td>.154</td>
<td>.273</td>
<td>.149</td>
<td>.165</td>
<td>.078</td>
<td>.073</td>
<td>.141</td>
<td>.094</td>
<td>-.237</td>
<td>-.190</td>
<td>-.234</td>
<td>.272</td>
<td>.157</td>
<td>.163</td>
<td>.106</td>
<td>.391</td>
<td>.138</td>
<td>.287</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>.114</td>
<td>.196</td>
<td>.138</td>
<td>.242</td>
<td>.170</td>
<td>.222</td>
<td>.266</td>
<td>.056</td>
<td>.164</td>
<td>-.196</td>
<td>-.198</td>
<td>-.287</td>
<td>.067</td>
<td>.213</td>
<td>.106</td>
<td>.051</td>
<td>.042</td>
<td>.179</td>
<td>.112</td>
<td>.120</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>.107</td>
<td>.271</td>
<td>.186</td>
<td>.281</td>
<td>.306</td>
<td>.193</td>
<td>.205</td>
<td>.024</td>
<td>.136</td>
<td>-.408</td>
<td>-.449</td>
<td>-.398</td>
<td>.164</td>
<td>.304</td>
<td>.119</td>
<td>.122</td>
<td>.172</td>
<td>.151</td>
<td>.165</td>
<td>.280</td>
<td>.405</td>
<td>1</td>
</tr>
</tbody>
</table>

M  4.25  3.86  3.46  4.11  2.80  4.63  4.50  3.92  5.30  5.53  4.61  4.48  2.96  3.85  4.37  3.20  2.05  4.27  2.69  2.02  4.16  2.71
SD  1.70  1.60  1.54  1.49  1.54  1.69  1.71  2.17  1.70  1.48  1.53  1.67  1.82  1.56  1.76  2.23  1.47  1.53  1.61  1.42  1.85  1.52
Skew -.06  .01  .01  -.19  .65  -.49  -.51  -.08  -1.14  -.85  -.06  -.20  .51  -.11  -.40  .60  1.37  -.11  .53  1.41  .00  .82  1.42
Kurtosis -.77  -.54  -.72  -.16  -.08  -.54  -.48  -1.44  .53  .25  -.41  -.64  -.83  -1.11  -.54  -1.18  1.06  -.21  -.81  1.42  -.93  .35

Note. Responses range from 1 (strong disagreement) to 7 (strong agreement). Items 1, 2, 4, 5, 6, 7, 14, 16, 18, 21, 22 had been reverse-scored. The sample does not include outliers. N = 857.
Table 7
Correlations and Descriptive Statistics for the SAIAT-HE Items for Mid-career Validation Student Sample

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.310</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.112</td>
<td>.551</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.325</td>
<td>.653</td>
<td>.372</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.302</td>
<td>.414</td>
<td>.368</td>
<td>.411</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.287</td>
<td>.404</td>
<td>.268</td>
<td>.518</td>
<td>.317</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.233</td>
<td>.420</td>
<td>.302</td>
<td>.489</td>
<td>.341</td>
<td>.716</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.051</td>
<td>.062</td>
<td>.126</td>
<td>.082</td>
<td>.196</td>
<td>.217</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.124</td>
<td>.268</td>
<td>.249</td>
<td>.325</td>
<td>.180</td>
<td>.462</td>
<td>.510</td>
<td>.287</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-.170</td>
<td>-.368</td>
<td>-.272</td>
<td>-.400</td>
<td>-.356</td>
<td>-.321</td>
<td>-.341</td>
<td>-.053</td>
<td>-.219</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-.173</td>
<td>-.374</td>
<td>-.301</td>
<td>-.446</td>
<td>-.311</td>
<td>-.376</td>
<td>-.442</td>
<td>-.075</td>
<td>-.302</td>
<td>.527</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-.174</td>
<td>-.462</td>
<td>-.340</td>
<td>-.548</td>
<td>-.357</td>
<td>-.471</td>
<td>-.506</td>
<td>-.120</td>
<td>-.365</td>
<td>.570</td>
<td>.549</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-.024</td>
<td>.145</td>
<td>.224</td>
<td>.189</td>
<td>.227</td>
<td>.139</td>
<td>.190</td>
<td>.133</td>
<td>.175</td>
<td>-.208</td>
<td>-.198</td>
<td>-.295</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>.102</td>
<td>.324</td>
<td>.275</td>
<td>.377</td>
<td>.256</td>
<td>.329</td>
<td>.326</td>
<td>.063</td>
<td>.238</td>
<td>-.335</td>
<td>-.426</td>
<td>-.413</td>
<td>.431</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.019</td>
<td>.177</td>
<td>.225</td>
<td>.235</td>
<td>.149</td>
<td>.199</td>
<td>.187</td>
<td>.115</td>
<td>.213</td>
<td>-.162</td>
<td>-.220</td>
<td>-.230</td>
<td>.451</td>
<td>.481</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>.039</td>
<td>.122</td>
<td>.029</td>
<td>.116</td>
<td>.124</td>
<td>.049</td>
<td>.070</td>
<td>.025</td>
<td>.015</td>
<td>-.142</td>
<td>-.112</td>
<td>-.171</td>
<td>.117</td>
<td>.254</td>
<td>.106</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-.058</td>
<td>.099</td>
<td>.113</td>
<td>.101</td>
<td>.121</td>
<td>.030</td>
<td>.034</td>
<td>.221</td>
<td>.088</td>
<td>-.142</td>
<td>-.005</td>
<td>-.174</td>
<td>.265</td>
<td>.183</td>
<td>.123</td>
<td>.086</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-.070</td>
<td>.265</td>
<td>.141</td>
<td>.238</td>
<td>.163</td>
<td>.248</td>
<td>.206</td>
<td>.206</td>
<td>.198</td>
<td>-.205</td>
<td>-.190</td>
<td>-.272</td>
<td>.143</td>
<td>.290</td>
<td>.130</td>
<td>.151</td>
<td>.095</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-.012</td>
<td>.125</td>
<td>.172</td>
<td>.164</td>
<td>.134</td>
<td>.137</td>
<td>.169</td>
<td>.194</td>
<td>.169</td>
<td>-.190</td>
<td>-.168</td>
<td>-.273</td>
<td>.518</td>
<td>.309</td>
<td>.325</td>
<td>.119</td>
<td>.378</td>
<td>.153</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-.060</td>
<td>.122</td>
<td>.127</td>
<td>.156</td>
<td>.149</td>
<td>.063</td>
<td>.127</td>
<td>.161</td>
<td>.099</td>
<td>-.301</td>
<td>-.110</td>
<td>-.291</td>
<td>.296</td>
<td>.226</td>
<td>.186</td>
<td>.096</td>
<td>.299</td>
<td>.161</td>
<td>.308</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>.088</td>
<td>.206</td>
<td>.127</td>
<td>.210</td>
<td>.159</td>
<td>.200</td>
<td>.209</td>
<td>.053</td>
<td>.189</td>
<td>-.299</td>
<td>-.323</td>
<td>-.297</td>
<td>.118</td>
<td>.239</td>
<td>.087</td>
<td>.130</td>
<td>.041</td>
<td>.222</td>
<td>.069</td>
<td>.058</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>.126</td>
<td>.220</td>
<td>.236</td>
<td>.294</td>
<td>.182</td>
<td>.192</td>
<td>.227</td>
<td>.047</td>
<td>.163</td>
<td>-.461</td>
<td>-.441</td>
<td>-.395</td>
<td>.181</td>
<td>.327</td>
<td>.114</td>
<td>.136</td>
<td>.144</td>
<td>.267</td>
<td>.167</td>
<td>.250</td>
<td>.411</td>
<td>1</td>
</tr>
</tbody>
</table>

Note. Responses range from 1 (strong disagreement) to 7 (strong agreement). Items 1, 2, 4, 5, 6, 7, 14, 16, 18, 21, 22 had been reverse-scored. The sample does not include outliers. N = 850.
Table 8

Fit Indices from CFA for SAIAT-HE (22 items) and the SAIAT-HE-revised (10 items) for the First-year, Mid-career Calibration, and Mid-career Validation Student Samples

<table>
<thead>
<tr>
<th>Model</th>
<th>SAIAT-HE (22 items)</th>
<th>SAIAT-HE-revised (10 items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \chi^2_{ML} )</td>
<td>( \chi^2_{SB} )</td>
</tr>
<tr>
<td>a) Six-factor first-year ((N = 856))</td>
<td>960.34*</td>
<td>845.907*</td>
</tr>
<tr>
<td>b) Six-factor mid-career calibration ((N = 857))</td>
<td>1006.01*</td>
<td>837.807*</td>
</tr>
<tr>
<td>a) Three-factor first-year ((N = 856))</td>
<td>195.95*</td>
<td>136.27*</td>
</tr>
<tr>
<td>b) Three-factor mid-career calibration ((N = 857))</td>
<td>161.85*</td>
<td>131.25*</td>
</tr>
<tr>
<td>c) Three-factor mid-career validation ((N = 850))</td>
<td>113.63*</td>
<td>85.47*</td>
</tr>
</tbody>
</table>

Note. ML = Maximum Likelihood; SB = Satorra-Bentler; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error Approximation; SRMR = Standardized Root Mean Square Error Residual.

\(*p < .01\)
Table 9

Correlation Residuals for the Six-factor Model for the SAJT-HE (22 items)

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.007</td>
<td>-.010</td>
<td>-.014</td>
<td>.041</td>
<td>.050</td>
<td>.023</td>
<td>.029</td>
<td>-.021</td>
<td>.022</td>
<td>-.049</td>
<td>.043</td>
<td>1.145</td>
<td>-.012</td>
<td>-.072</td>
<td>.012</td>
<td>-.070</td>
<td>.139</td>
<td>-.152</td>
<td>-.092</td>
<td>.033</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-.018</td>
<td>-.034</td>
<td>.019</td>
<td>.026</td>
<td>-.073</td>
<td>-.055</td>
<td>-.065</td>
<td>-.099</td>
<td>.046</td>
<td>.054</td>
<td>.009</td>
<td>-.091</td>
<td>.079</td>
<td>-.041</td>
<td>.004</td>
<td>-.061</td>
<td>.258</td>
<td>-.080</td>
<td>.003</td>
<td>-.047</td>
<td>-.036</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-.037</td>
<td>-.022</td>
<td>-.036</td>
<td>.166</td>
<td>.056</td>
<td>.043</td>
<td>.022</td>
<td>-.038</td>
<td>-.046</td>
<td>-.045</td>
<td>-.040</td>
<td>.043</td>
<td>.071</td>
<td>.068</td>
<td>.006</td>
<td>.048</td>
<td>.096</td>
<td>.007</td>
<td>.110</td>
<td>.019</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-.06</td>
<td>.041</td>
<td>-.042</td>
<td>-.049</td>
<td>.058</td>
<td>.025</td>
<td>-.014</td>
<td>-.036</td>
<td>.048</td>
<td>.010</td>
<td>-.022</td>
<td>-.061</td>
<td>.130</td>
<td>-.017</td>
<td>-.008</td>
<td>.003</td>
<td>.281</td>
<td>-.022</td>
<td>.042</td>
<td>-.040</td>
<td>-.001</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.090</td>
<td>-.017</td>
<td>.117</td>
<td>-.061</td>
<td>-.003</td>
<td>-.004</td>
<td>-.064</td>
<td>-.070</td>
<td>-.084</td>
<td>-.033</td>
<td>-.011</td>
<td>-.054</td>
<td>-.003</td>
<td>-.003</td>
<td>-.070</td>
<td>-.010</td>
<td>.053</td>
<td>.128</td>
<td>.005</td>
<td>-.009</td>
<td>-.041</td>
<td>-.009</td>
</tr>
<tr>
<td>6</td>
<td>.135</td>
<td>-.048</td>
<td>.014</td>
<td>-.027</td>
<td>-.022</td>
<td>-.001</td>
<td>-.015</td>
<td>-.002</td>
<td>.072</td>
<td>-.022</td>
<td>-.009</td>
<td>.070</td>
<td>.078</td>
<td>-.002</td>
<td>-.011</td>
<td>-.020</td>
<td>.236</td>
<td>-.047</td>
<td>.064</td>
<td>.019</td>
<td>-.052</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.074</td>
<td>-.048</td>
<td>.048</td>
<td>.010</td>
<td>-.019</td>
<td>.002</td>
<td>-.008</td>
<td>-.012</td>
<td>.075</td>
<td>-.002</td>
<td>-.028</td>
<td>-.068</td>
<td>.053</td>
<td>-.005</td>
<td>-.038</td>
<td>-.015</td>
<td>.249</td>
<td>-.048</td>
<td>.047</td>
<td>.061</td>
<td>-.022</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.066</td>
<td>.038</td>
<td>.114</td>
<td>.048</td>
<td>.020</td>
<td>-.014</td>
<td>-.028</td>
<td>.117</td>
<td>.062</td>
<td>-.004</td>
<td>.043</td>
<td>.076</td>
<td>.008</td>
<td>.050</td>
<td>.017</td>
<td>.030</td>
<td>.164</td>
<td>-.071</td>
<td>.073</td>
<td>-.018</td>
<td>-.034</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.024</td>
<td>-.029</td>
<td>.042</td>
<td>-.016</td>
<td>-.079</td>
<td>.008</td>
<td>-.006</td>
<td>.069</td>
<td>.063</td>
<td>.012</td>
<td>-.031</td>
<td>.040</td>
<td>.006</td>
<td>.090</td>
<td>-.121</td>
<td>-.102</td>
<td>.142</td>
<td>-.002</td>
<td>.080</td>
<td>.029</td>
<td>.012</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.054</td>
<td>.043</td>
<td>-.004</td>
<td>.046</td>
<td>-.104</td>
<td>-.061</td>
<td>-.028</td>
<td>-.021</td>
<td>-.035</td>
<td>.143</td>
<td>-.010</td>
<td>.050</td>
<td>-.046</td>
<td>.082</td>
<td>.011</td>
<td>-.084</td>
<td>-.093</td>
<td>.027</td>
<td>-.075</td>
<td>.044</td>
<td>-.049</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.031</td>
<td>-.052</td>
<td>-.002</td>
<td>-.060</td>
<td>-.019</td>
<td>-.007</td>
<td>-.011</td>
<td>.0</td>
<td>-.007</td>
<td>.062</td>
<td>-.027</td>
<td>-.076</td>
<td>-.089</td>
<td>.081</td>
<td>-.011</td>
<td>-.026</td>
<td>-.148</td>
<td>.105</td>
<td>-.025</td>
<td>.022</td>
<td>-.073</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.063</td>
<td>-.023</td>
<td>-.009</td>
<td>-.076</td>
<td>-.014</td>
<td>-.013</td>
<td>-.041</td>
<td>-.033</td>
<td>-.011</td>
<td>-.005</td>
<td>-.031</td>
<td>.085</td>
<td>.139</td>
<td>.030</td>
<td>-.008</td>
<td>-.055</td>
<td>-.276</td>
<td>.054</td>
<td>-.071</td>
<td>.018</td>
<td>-.029</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>-.114</td>
<td>-.124</td>
<td>.056</td>
<td>-.093</td>
<td>-.003</td>
<td>-.139</td>
<td>-.120</td>
<td>.003</td>
<td>-.027</td>
<td>-.039</td>
<td>.106</td>
<td>.081</td>
<td>-.062</td>
<td>-.057</td>
<td>-.070</td>
<td>-.045</td>
<td>-.043</td>
<td>.141</td>
<td>.215</td>
<td>-.106</td>
<td>-.102</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>.018</td>
<td>.063</td>
<td>.076</td>
<td>.088</td>
<td>.140</td>
<td>.083</td>
<td>.114</td>
<td>.073</td>
<td>.040</td>
<td>-.080</td>
<td>-.092</td>
<td>-.089</td>
<td>-.057</td>
<td>.011</td>
<td>-.085</td>
<td>-.026</td>
<td>.175</td>
<td>-.092</td>
<td>.174</td>
<td>.075</td>
<td>.089</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-.161</td>
<td>-.038</td>
<td>-.063</td>
<td>-.045</td>
<td>-.056</td>
<td>-.037</td>
<td>-.003</td>
<td>-.067</td>
<td>-.058</td>
<td>-.087</td>
<td>-.052</td>
<td>-.019</td>
<td>-.088</td>
<td>.0</td>
<td>-.012</td>
<td>-.083</td>
<td>-.045</td>
<td>-.047</td>
<td>-.130</td>
<td>-.021</td>
<td>-.044</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-.017</td>
<td>-.031</td>
<td>-.008</td>
<td>-.005</td>
<td>-.082</td>
<td>.016</td>
<td>-.004</td>
<td>-.004</td>
<td>-.089</td>
<td>-.015</td>
<td>-.028</td>
<td>-.011</td>
<td>-.049</td>
<td>.070</td>
<td>-.027</td>
<td>-.011</td>
<td>-.065</td>
<td>-.023</td>
<td>-.172</td>
<td>.011</td>
<td>-.029</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-.067</td>
<td>-.029</td>
<td>.133</td>
<td>-.026</td>
<td>.106</td>
<td>-.032</td>
<td>-.045</td>
<td>.161</td>
<td>-.031</td>
<td>-.051</td>
<td>.058</td>
<td>.015</td>
<td>-.019</td>
<td>-.081</td>
<td>-.062</td>
<td>-.001</td>
<td>-.041</td>
<td>-.017</td>
<td>.201</td>
<td>-.052</td>
<td>-.028</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>.027</td>
<td>.139</td>
<td>.088</td>
<td>.121</td>
<td>.067</td>
<td>.159</td>
<td>.114</td>
<td>.102</td>
<td>-.066</td>
<td>-.062</td>
<td>-.046</td>
<td>-.147</td>
<td>-.042</td>
<td>-.096</td>
<td>-.027</td>
<td>-.055</td>
<td>-.072</td>
<td>-.111</td>
<td>.132</td>
<td>.149</td>
<td>.117</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-.093</td>
<td>-.086</td>
<td>.085</td>
<td>-.067</td>
<td>.064</td>
<td>-.065</td>
<td>-.026</td>
<td>.078</td>
<td>-.010</td>
<td>-.019</td>
<td>-.069</td>
<td>-.040</td>
<td>.197</td>
<td>-.117</td>
<td>.002</td>
<td>-.067</td>
<td>-.053</td>
<td>-.031</td>
<td>-.250</td>
<td>-.052</td>
<td>-.032</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-.110</td>
<td>.005</td>
<td>.173</td>
<td>-.011</td>
<td>.051</td>
<td>-.035</td>
<td>-.042</td>
<td>.109</td>
<td>-.011</td>
<td>-.038</td>
<td>-.003</td>
<td>-.003</td>
<td>.156</td>
<td>.027</td>
<td>.059</td>
<td>.058</td>
<td>.310</td>
<td>.069</td>
<td>.161</td>
<td>-.008</td>
<td>-.036</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>.019</td>
<td>.014</td>
<td>.016</td>
<td>.047</td>
<td>.033</td>
<td>.084</td>
<td>.124</td>
<td>.016</td>
<td>.063</td>
<td>.045</td>
<td>.035</td>
<td>-.009</td>
<td>-.076</td>
<td>-.054</td>
<td>-.020</td>
<td>-.007</td>
<td>-.057</td>
<td>.090</td>
<td>-.041</td>
<td>-.070</td>
<td>-.009</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-.044</td>
<td>-.022</td>
<td>-.010</td>
<td>-.037</td>
<td>.083</td>
<td>-.029</td>
<td>-.023</td>
<td>-.043</td>
<td>-.027</td>
<td>-.021</td>
<td>-.074</td>
<td>.051</td>
<td>-.064</td>
<td>.051</td>
<td>-.079</td>
<td>.021</td>
<td>.015</td>
<td>.013</td>
<td>-.080</td>
<td>-.026</td>
<td>-.035</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values from the mid-career calibration student sample (N = 857) are below the diagonal; values from the first-year student sample (N = 856) are above the diagonal. Positive values indicate that the relationship between two items was underestimated by the model. Values greater than |.10| are bolded.
Table 10

Item-level Model Parameters from the Six-factor CFA model on the SAIAT-HE from the First-year and the Mid-career Student Samples

<table>
<thead>
<tr>
<th>Items</th>
<th>First-year N = 856</th>
<th></th>
<th>Mid-career calibration N = 857</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized</td>
<td>Standardized</td>
<td>McDonald's Omega</td>
<td>R²</td>
</tr>
<tr>
<td></td>
<td>Pattern Coefficients</td>
<td>Coefficients</td>
<td></td>
<td>Omega</td>
</tr>
<tr>
<td>VALIDITY</td>
<td></td>
<td></td>
<td></td>
<td>.728</td>
</tr>
<tr>
<td>1. Assessment tests are unfair to some students (R).</td>
<td>.73</td>
<td>.47</td>
<td>.22</td>
<td>.67</td>
</tr>
<tr>
<td>2. Assessment test results are not accurate (R).</td>
<td>1.03</td>
<td>.75</td>
<td>.56</td>
<td>1.24</td>
</tr>
<tr>
<td>3. Assessment test results accurately reflect basic skills and knowledge of a subject.</td>
<td>.54</td>
<td>.38</td>
<td>.15</td>
<td>.78</td>
</tr>
<tr>
<td>4. Assessment tests are not valid (R).</td>
<td>1.19</td>
<td>.86</td>
<td>.74</td>
<td>1.31</td>
</tr>
<tr>
<td>5. Assessment test scores don’t reflect my true ability (R).</td>
<td>.71</td>
<td>.48</td>
<td>.23</td>
<td>.86</td>
</tr>
<tr>
<td>PURPOSE</td>
<td></td>
<td></td>
<td></td>
<td>.757</td>
</tr>
<tr>
<td>6. I don’t understand how assessment tests are related to my education (R).</td>
<td>1.16</td>
<td>.80</td>
<td>.63</td>
<td>1.39</td>
</tr>
<tr>
<td>7. I don’t understand the need for assessment tests (R).</td>
<td>1.28</td>
<td>.87</td>
<td>.75</td>
<td>1.50</td>
</tr>
<tr>
<td>8. Someone (professor, academic advisor, Resident Advisor) explained to me why I take assessment tests.</td>
<td>.77</td>
<td>.43</td>
<td>.19</td>
<td>.55</td>
</tr>
<tr>
<td>9. I understand the purpose of assessment tests.</td>
<td>.80</td>
<td>.59</td>
<td>.35</td>
<td>1.00</td>
</tr>
<tr>
<td>DISILLUSIONMENT</td>
<td></td>
<td></td>
<td></td>
<td>.722</td>
</tr>
<tr>
<td>10. The more assessment tests I complete, the more I dislike assessment tests.</td>
<td>.72</td>
<td>.50</td>
<td>.25</td>
<td>1.03</td>
</tr>
<tr>
<td>11. There is too much assessment testing.</td>
<td>.94</td>
<td>.66</td>
<td>.43</td>
<td>1.10</td>
</tr>
<tr>
<td>12. Assessment tests are a waste of my time.</td>
<td>1.27</td>
<td>.85</td>
<td>.73</td>
<td>1.37</td>
</tr>
<tr>
<td>PARENTS</td>
<td></td>
<td></td>
<td></td>
<td>.619</td>
</tr>
<tr>
<td>13. My parents would be disappointed if I performed poorly on the assessment tests.</td>
<td>1.21</td>
<td>.66</td>
<td>.43</td>
<td>1.24</td>
</tr>
<tr>
<td>14. My parents don’t value the assessment tests I complete at the University (R).</td>
<td>1.02</td>
<td>.67</td>
<td>.45</td>
<td>1.17</td>
</tr>
<tr>
<td>15. My parents would be proud of me if I performed well on the assessment tests.</td>
<td>0.93</td>
<td>.64</td>
<td>.41</td>
<td>1.09</td>
</tr>
<tr>
<td>16. My parents are unaware of the assessment tests I complete at the University (R).</td>
<td>0.53</td>
<td>.26</td>
<td>.07</td>
<td>0.56</td>
</tr>
<tr>
<td>PROFESSOR</td>
<td></td>
<td></td>
<td></td>
<td>.548</td>
</tr>
<tr>
<td>17. My professor(s), an academic advisor, or resident advisor encouraged me to prepare for the assessment tests.</td>
<td>0.97</td>
<td>.55</td>
<td>.30</td>
<td>0.69</td>
</tr>
<tr>
<td>18. My professors don’t value the assessment tests I complete at the University (R).</td>
<td>0.38</td>
<td>.27</td>
<td>.07</td>
<td>0.48</td>
</tr>
<tr>
<td>19. If I performed poorly on the assessment tests, my professors would be disappointed.</td>
<td>1.17</td>
<td>.73</td>
<td>.53</td>
<td>1.20</td>
</tr>
<tr>
<td>STUDENTS</td>
<td></td>
<td></td>
<td></td>
<td>.507</td>
</tr>
<tr>
<td>20. Fellow students urged me to try my best on the assessment tests.</td>
<td>0.33</td>
<td>.19</td>
<td>.04</td>
<td>0.56</td>
</tr>
<tr>
<td>21. Fellow students discouraged me from taking the assessment tests seriously (R).</td>
<td>0.92</td>
<td>.58</td>
<td>.34</td>
<td>0.90</td>
</tr>
<tr>
<td>22. Fellow students speak negatively about the assessment tests at the University (R).</td>
<td>1.21</td>
<td>.74</td>
<td>.54</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Note. Factor variances were set to unity.
Table 11

Correlation Residuals for the Three-factor Model for the SAIAT-HE-revised (10 items)

<table>
<thead>
<tr>
<th>Item</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.029</td>
<td>.029</td>
<td>.038</td>
<td>.134</td>
<td>.026</td>
<td>.060</td>
<td>.050</td>
<td>.013</td>
<td>.010</td>
<td>.001</td>
</tr>
<tr>
<td>3</td>
<td>-.018</td>
<td>-.045</td>
<td>.005</td>
<td>-.040</td>
<td>-.022</td>
<td>.047</td>
<td>.048</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.019</td>
<td>-.036</td>
<td>-.048</td>
<td>.035</td>
<td>.014</td>
<td>-.008</td>
<td>.046</td>
<td>.052</td>
<td>.052</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.036</td>
<td>.170</td>
<td>-.044</td>
<td>-.003</td>
<td>-.004</td>
<td>-.063</td>
<td>-.118</td>
<td>-.036</td>
<td>-.014</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-.068</td>
<td>.056</td>
<td>.053</td>
<td>.002</td>
<td>-.002</td>
<td>.011</td>
<td>.053</td>
<td>-.009</td>
<td>-.002</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-.055</td>
<td>.043</td>
<td>.020</td>
<td>.004</td>
<td>-.001</td>
<td>-.003</td>
<td>.024</td>
<td>-.023</td>
<td>-.027</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-.094</td>
<td>.044</td>
<td>-.031</td>
<td>-.065</td>
<td>.008</td>
<td>-.002</td>
<td>.027</td>
<td>-.018</td>
<td>.018</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.046</td>
<td>-.046</td>
<td>.053</td>
<td>-.089</td>
<td>.072</td>
<td>.075</td>
<td>.057</td>
<td>.084</td>
<td>-.013</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.049</td>
<td>-.050</td>
<td>.010</td>
<td>-.038</td>
<td>.022</td>
<td>-.002</td>
<td>.001</td>
<td>.148</td>
<td>-.028</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.009</td>
<td>-.040</td>
<td>-.017</td>
<td>-.011</td>
<td>-.005</td>
<td>-.019</td>
<td>-.036</td>
<td>-.015</td>
<td>-.022</td>
<td></td>
</tr>
</tbody>
</table>

Note. Values from the mid-career calibration student sample ($N = 857$) are below the diagonal; values from the first-year student sample ($N = 856$) are above the diagonal. Positive values indicate that the relationship between two items was underestimated by the model; negative values indicate that the relationship between two items was overestimated by the model. Values greater than $|.10|$ are bolded. None of the correlation residuals from the mid-career validation sample ($N = 850$) exceeded $|.05|$. 
<table>
<thead>
<tr>
<th>Items</th>
<th>First-year N = 856</th>
<th>Mid-career calibration N = 857</th>
<th>Mid-career validation N = 850</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VALIDITY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Assessment test results are not accurate (R)</td>
<td>0.86</td>
<td>.74</td>
<td>.55</td>
</tr>
<tr>
<td>3. Assessment test results accurately reflect basic skills and knowledge of a subject</td>
<td>0.45</td>
<td>.38</td>
<td>.15</td>
</tr>
<tr>
<td>4. Assessment tests are not valid (R)</td>
<td>1.00</td>
<td>.87</td>
<td>.75</td>
</tr>
<tr>
<td>5. Assessment test scores don’t reflect my true ability (R)</td>
<td>0.58</td>
<td>.48</td>
<td>.23</td>
</tr>
<tr>
<td><strong>PURPOSE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. I don’t understand how assessment tests are related to my education (R)</td>
<td>0.9</td>
<td>.80</td>
<td>.63</td>
</tr>
<tr>
<td>7. I don’t understand the need for assessment tests (R)</td>
<td>1.00</td>
<td>.87</td>
<td>.76</td>
</tr>
<tr>
<td>9. I understand the purpose of assessment tests</td>
<td>0.61</td>
<td>.58</td>
<td>.33</td>
</tr>
<tr>
<td><strong>DISILLUSIONMENT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. The more assessment tests I complete, the more I dislike assessment tests</td>
<td>0.56</td>
<td>.50</td>
<td>.25</td>
</tr>
<tr>
<td>11. There is too much assessment testing</td>
<td>0.73</td>
<td>.65</td>
<td>.42</td>
</tr>
<tr>
<td>12. Assessment tests are a waste of my time</td>
<td>1.00</td>
<td>.86</td>
<td>.74</td>
</tr>
</tbody>
</table>

Note. The following referent indicators were used to set the metric of each factor: Validity item 4, Purpose item 7, Disillusionment item 12.
Table 13

*Tests of Invariance of SAIAT-HE-revised Across First-year and Mid-career Validation Student Samples*

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$ ML</th>
<th>df</th>
<th>$\Delta \chi^2$ ML</th>
<th>$\Delta df$</th>
<th>p-value</th>
<th>CFI</th>
<th>$\Delta$CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural Invariance</td>
<td>309.58*</td>
<td>64</td>
<td></td>
<td></td>
<td>.98</td>
<td>.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric Invariance</td>
<td>324.07*</td>
<td>71</td>
<td>14.49</td>
<td>7</td>
<td>.043</td>
<td>.98</td>
<td>&lt; .01</td>
<td>.056</td>
</tr>
<tr>
<td>Scalar Invariance</td>
<td>403.51*</td>
<td>78</td>
<td>79.44</td>
<td>14</td>
<td>&lt; .01</td>
<td>.98</td>
<td>&lt; .01</td>
<td>.063</td>
</tr>
</tbody>
</table>

*Note.* The following invariant indicators were used to set the metric of each factor: Validity item 4, Purpose item 7, and Disillusionment item 12. The $\chi^2$ for the configural model is simply the sum of the $\chi^2$ values from fitting the three-factor model to the first-year and mid-career validation samples (values presented in Table 8). ML = Maximum Likelihood; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error Approximation.

*p < .01*
Table 14

Mean Differences of Observed and Latent SALAT-HE-revised Factor Scores across First-year and Mid-career Students

<table>
<thead>
<tr>
<th>Validity</th>
<th>Purpose</th>
<th>Disillusionment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent mean difference</td>
<td>-0.630</td>
<td>-0.310</td>
</tr>
<tr>
<td>Effect size</td>
<td>-0.505</td>
<td>-0.221</td>
</tr>
<tr>
<td>Observed mean difference</td>
<td>-0.501</td>
<td>-0.256</td>
</tr>
<tr>
<td>Effect size</td>
<td>-0.447</td>
<td>-0.193</td>
</tr>
</tbody>
</table>

*Note. The latent mean for first-year students was fixed to zero; the latent mean for mid-career students represents the difference between the groups. All differences are significant at p < .05. First-year student sample N = 856. Mid-career student sample N = 850. Unstandardized estimates (latent and observed mean differences, observed means) range from 1 to 7.*
Table 15

Correlations among the SAIAT-K-12 and SAIAT-HE-revised Subscales for Matched Incoming First-year Student Sample

<table>
<thead>
<tr>
<th></th>
<th>Validity K-12</th>
<th>Purpose K-12</th>
<th>Disill K-12</th>
<th>Parents K-12</th>
<th>Validity HE</th>
<th>Purpose HE</th>
<th>Disill HE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity K-12</td>
<td>.84</td>
<td>.52</td>
<td>-.60</td>
<td>.28</td>
<td>.33</td>
<td>.19</td>
<td>-.27</td>
</tr>
<tr>
<td>Purpose K-12</td>
<td>.78</td>
<td>.74</td>
<td>-.38</td>
<td>.23</td>
<td>.22</td>
<td>.31</td>
<td>-.26</td>
</tr>
<tr>
<td>Disill K-12</td>
<td>-.74</td>
<td>-.59</td>
<td>.83</td>
<td>-.16</td>
<td>-.23</td>
<td>-.17</td>
<td>.33</td>
</tr>
<tr>
<td>Parents K-12</td>
<td>.46</td>
<td>.34</td>
<td>.29</td>
<td>.68</td>
<td>.15</td>
<td>.16</td>
<td>-.18</td>
</tr>
<tr>
<td>Validity HE</td>
<td>.26</td>
<td>.31</td>
<td>-.22</td>
<td>.20</td>
<td>.72</td>
<td>.47</td>
<td>-.54</td>
</tr>
<tr>
<td>Purpose HE</td>
<td>.36</td>
<td>.41</td>
<td>-.19</td>
<td>.22</td>
<td>.63</td>
<td>.79</td>
<td>-.51</td>
</tr>
<tr>
<td>Disill HE</td>
<td>-.32</td>
<td>-.37</td>
<td>.34</td>
<td>-.26</td>
<td>-.75</td>
<td>-.71</td>
<td>.75</td>
</tr>
</tbody>
</table>

\[M\] 3.67  4.57  5.22  5.10  4.08  5.07  4.39
\[SD\] 1.26  1.30  1.44  1.08  1.03  1.18  1.19
\[95\% CI\] [3.57, 3.76] [4.47, 4.66] [5.11, 5.33] [5.02, 5.18] [3.99, 4.15] [4.98, 5.16] [4.30, 4.48]

Skewness 0.15 -0.18 -0.64 -0.32 -0.07 -0.41 -0.09
Kurtosis -0.32 -0.28 0.05 -0.19 0.22 -0.05 0.09

*Note.* Disil - Disillusionment. \(M\) - Mean; \(SD\) - Standard Deviation; CI = Confidence Interval. SAIAT-K-12 and SAIAT-HE-revised subscale scores range from 1 to 7. In order to obtain correlations among latent factors, a seven-factor interrelated model was fit to the data, yielding adequate model-data fit: \(\chi^2_{SB}(278) = 864.08;\) Robust CFI = .95; RMSEA = .058; SRMR = .056. Correlations among latent factors are above the diagonal; correlations among observed subscales are below the diagonal. Cronbach's alphas for each subscale are listed on the diagonal. Correlations among the corresponding SAIAT-K-12 and SAIAT-HE-revised dimensions are bolded. All correlations are significant at \(p < .05. N = 678.\)
### Table 16

Descriptive Statistics and Correlations among the Variables Used in Path Analysis for the Matched Incoming and First-year Student Sample

<table>
<thead>
<tr>
<th></th>
<th>Validity HE</th>
<th>Purpose HE</th>
<th>Disil HE</th>
<th>Validity K-12</th>
<th>Purpose K-12</th>
<th>Disil K-12</th>
<th>Parents K-12</th>
<th>Importance</th>
<th>Effort</th>
<th>SAT math</th>
<th>SAT verbal</th>
<th>GLEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity HE</td>
<td>.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose HE</td>
<td>.470</td>
<td>.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disil HE</td>
<td>-.541</td>
<td>-.505</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity K-12</td>
<td>.328</td>
<td>.187</td>
<td>-.269</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose K-12</td>
<td>.218</td>
<td>.307</td>
<td>-.258</td>
<td>.519</td>
<td>.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disil K-12</td>
<td>-.227</td>
<td>-.167</td>
<td>.329</td>
<td>-.596</td>
<td>-.384</td>
<td>.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parents K-12</td>
<td>.147</td>
<td>.157</td>
<td>-.180</td>
<td>.275</td>
<td>.228</td>
<td>-.159</td>
<td>.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importance</td>
<td>.318</td>
<td>.341</td>
<td>-.453</td>
<td>.169</td>
<td>.187</td>
<td>-.171</td>
<td>.229</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>.252</td>
<td>.281</td>
<td>-.276</td>
<td>.033</td>
<td>.090</td>
<td>-.032</td>
<td>.148</td>
<td>.394</td>
<td>.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT math</td>
<td>.066</td>
<td>-.020</td>
<td>-.023</td>
<td>.203</td>
<td>.178</td>
<td>-.138</td>
<td>.013</td>
<td>-.040</td>
<td>.018</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>SAT verbal</td>
<td>.161</td>
<td>.151</td>
<td>-.009</td>
<td>.103</td>
<td>.197</td>
<td>-.019</td>
<td>.019</td>
<td>.031</td>
<td>.142</td>
<td>.416</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>GLEX total</td>
<td>.176</td>
<td>.151</td>
<td>-.056</td>
<td>.104</td>
<td>.168</td>
<td>-.039</td>
<td>.023</td>
<td>.143</td>
<td>.288</td>
<td>.386</td>
<td>.645</td>
<td>.74</td>
</tr>
<tr>
<td>Number of items</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Mean</td>
<td>4.08</td>
<td>5.07</td>
<td>4.39</td>
<td>3.67</td>
<td>4.57</td>
<td>5.22</td>
<td>5.10</td>
<td>3.18</td>
<td>3.80</td>
<td>579.56</td>
<td>573.95</td>
<td>21.00</td>
</tr>
<tr>
<td>SD</td>
<td>1.03</td>
<td>1.18</td>
<td>1.19</td>
<td>1.26</td>
<td>1.30</td>
<td>1.44</td>
<td>1.08</td>
<td>0.75</td>
<td>0.71</td>
<td>66.70</td>
<td>68.83</td>
<td>4.82</td>
</tr>
<tr>
<td>95% CI</td>
<td>[4.00, 4.16]</td>
<td>[4.98, 5.16]</td>
<td>[4.30, 4.48]</td>
<td>[3.57, 3.76]</td>
<td>[4.47, 4.67]</td>
<td>[5.11, 5.33]</td>
<td>[5.02, 5.18]</td>
<td>[3.12, 3.24]</td>
<td>[3.75, 3.85]</td>
<td>[574.54, 584.58]</td>
<td>[568.77, 579.13]</td>
<td>[20.64, 21.36]</td>
</tr>
<tr>
<td>Skew</td>
<td>-0.07</td>
<td>-0.41</td>
<td>-0.09</td>
<td>0.15</td>
<td>-0.18</td>
<td>-0.64</td>
<td>-0.32</td>
<td>-0.30</td>
<td>-0.55</td>
<td>-0.12</td>
<td>0.22</td>
<td>-0.27</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.22</td>
<td>-0.05</td>
<td>0.09</td>
<td>-0.32</td>
<td>-0.28</td>
<td>0.05</td>
<td>-0.19</td>
<td>0.27</td>
<td>0.63</td>
<td>0.01</td>
<td>0.11</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

**Note.** K-12 notes subscales in the SAIAT-K12; HE notes subscales in SAIAT-HE-revised. Disil - Disillusionment. SAIAT-K-12 and SAIAT-HE subscales range from 1 to 7; Importance and Effort subscales range from 1 to 5; SAT math ranges from 360 to 790; SAT verbal ranges from 410 to 800; GLEX ranges from 6 to 32. Cronbach's alphas for each measure are on the diagonal. *N* = 678.
### Table 17

**Descriptive Statistics and Correlations among the Variables Used in Path Analysis for the Mid-career A-Day 1 Student Sample**

<table>
<thead>
<tr>
<th></th>
<th>Validity HE</th>
<th>Purpose HE</th>
<th>Disil HE</th>
<th>Importance</th>
<th>Effort</th>
<th>SAT math</th>
<th>SAT verbal</th>
<th>GLEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity HE</td>
<td>-.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose HE</td>
<td>.416</td>
<td>.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disil HE</td>
<td>-.548</td>
<td>-.497</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importance</td>
<td>.372</td>
<td>.376</td>
<td>-.436</td>
<td>.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>.330</td>
<td>.352</td>
<td>-.348</td>
<td>.430</td>
<td>.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT math</td>
<td>.066</td>
<td>-.006</td>
<td>.033</td>
<td>-.071</td>
<td>.122</td>
<td></td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>SAT verbal</td>
<td>.062</td>
<td>.010</td>
<td>.091</td>
<td>-.072</td>
<td>.118</td>
<td>.429</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>GLEX total</td>
<td>.132</td>
<td>.123</td>
<td>-.023</td>
<td>.159</td>
<td>.339</td>
<td>.352</td>
<td>.565</td>
<td>.77</td>
</tr>
<tr>
<td>Number of items</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>3.42</td>
<td>4.87</td>
<td>4.88</td>
<td>2.66</td>
<td>3.72</td>
<td>577.65</td>
<td>565.86</td>
<td>22.81</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>1.16</td>
<td>1.39</td>
<td>1.25</td>
<td>0.89</td>
<td>0.81</td>
<td>68.84</td>
<td>70.05</td>
<td>4.81</td>
</tr>
<tr>
<td><strong>95% CI</strong></td>
<td>[3.34, 3.50]</td>
<td>[4.77, 4.97]</td>
<td>[4.79, 4.97]</td>
<td>[2.60, 2.72]</td>
<td>[3.66, 3.78]</td>
<td>[572.75, 582.55]</td>
<td>[560.87, 570.85]</td>
<td>[22.47, 23.15]</td>
</tr>
<tr>
<td><strong>Skew</strong></td>
<td>0.00</td>
<td>-0.62</td>
<td>-0.14</td>
<td>-0.01</td>
<td>-0.59</td>
<td>0.15</td>
<td>0.08</td>
<td>-0.48</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>-0.22</td>
<td>0.08</td>
<td>-0.47</td>
<td>-0.36</td>
<td>0.49</td>
<td>0.09</td>
<td>0.06</td>
<td>-0.11</td>
</tr>
</tbody>
</table>

**Note.** K-12 notes subscales in the SAIAT-K-12; HE notes subscales in SAIAT-HE-revised. Disil - Disillusionment. SAIAT-K-12 and SAIAT-HE subscales range from 1 to 7; Importance and Effort subscales range from 1 to 5; SAT math ranges from 350 to 800; SAT verbal ranges from 320 to 800; GLEX ranges from 7 to 32. Cronbach's alphas for each measure are on the diagonal. N = 758.
Table 18

*Fit Indices for Competing Single-Indicator Latent Variable Models*

<table>
<thead>
<tr>
<th>Model</th>
<th>χ²</th>
<th>df</th>
<th>p -value</th>
<th>CFI</th>
<th>RMSEA</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: Fully-mediated model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matched Incoming and First-year Student Sample</td>
<td>26.39</td>
<td>9</td>
<td>&lt;. 01</td>
<td>.99</td>
<td>.054</td>
<td>.034</td>
</tr>
<tr>
<td>Mid-career A-Day 1 Student Sample</td>
<td>82.84</td>
<td>9</td>
<td>&lt;. 01</td>
<td>.96</td>
<td>.100</td>
<td>.055</td>
</tr>
<tr>
<td>Model 2: Partially-mediated model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matched Incoming and First-year Student Sample</td>
<td>26.00</td>
<td>6</td>
<td>&lt;. 01</td>
<td>.99</td>
<td>.071</td>
<td>.034</td>
</tr>
<tr>
<td>Mid-career A-Day 1 Student Sample</td>
<td>80.89</td>
<td>6</td>
<td>&lt;. 01</td>
<td>.96</td>
<td>.130</td>
<td>.055</td>
</tr>
<tr>
<td>Model 3: Partially-mediated model with SAIAT-K-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matched Incoming and First-year Student Sample</td>
<td>85.71</td>
<td>27</td>
<td>&lt;. 01</td>
<td>.98</td>
<td>.058</td>
<td>.043</td>
</tr>
<tr>
<td>Model 4: Modified fully-mediated model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matched Incoming and First-year Student Sample</td>
<td>15.03</td>
<td>10</td>
<td>.131</td>
<td>.99</td>
<td>.027</td>
<td>.024</td>
</tr>
<tr>
<td>Mid-career A-Day 1 Student Sample</td>
<td>52.83</td>
<td>10</td>
<td>&lt;. 01</td>
<td>.98</td>
<td>.076</td>
<td>.038</td>
</tr>
<tr>
<td>Model 5: Modified fully-mediated model with SAIAT-K-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matched Incoming and First-year Student Sample</td>
<td>75.52</td>
<td>31</td>
<td>&lt;. 01</td>
<td>.98</td>
<td>.046</td>
<td>.040</td>
</tr>
</tbody>
</table>

*Note.* A single-indicator latent variable approach was used in order to partial out measurement error. In accordance with Brown's (2006) guidelines, the error variance of each observed subscale was fixed to the error variance calculated as (1 – rxx) * (Varx), where rxx equals Cronbach's alpha. The observed subscale scores were used as single indicators of the corresponding latent variables, with the paths from observed to latent variables fixed to one. Model 2 is nested within Model 1. Matched Incoming and First-year Student Sample N = 768; Mid-career A-Day 1 Student Sample N = 758.

*p < .05*
Table 19

*Effects Decomposition for Model 4: Modified Fully-mediated Model in Matched Incoming and First-year Student Sample*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance</th>
<th>Effort</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unst.</td>
<td>SE</td>
<td>St.</td>
</tr>
<tr>
<td>SAT math</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>SAT verbal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Validity HE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-.024</td>
<td>.067</td>
<td>-.030</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-.024</td>
<td>.067</td>
<td>-.030</td>
</tr>
<tr>
<td>Purpose HE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>.063</td>
<td>.059</td>
<td>.098</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>.063</td>
<td>.059</td>
<td>.098</td>
</tr>
<tr>
<td>Disillusionment HE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-.355*</td>
<td>.059</td>
<td>-.538</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-.355*</td>
<td>.059</td>
<td>-.538</td>
</tr>
<tr>
<td>Importance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Effort</td>
<td>Total Indirect Effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Effort</td>
<td>Total Effect</td>
<td>.389*</td>
<td>.047</td>
</tr>
</tbody>
</table>

**Note.** Unst. = unstandardized; SE = standard error; St. = standardized. $N = 678$.  
*p < .05*
### Table 20

**Effects Decomposition for Model 4: Modified Fully-mediated Model in Mid-career Student Sample**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance</th>
<th>Effort</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unst.</td>
<td>SE</td>
<td>St.</td>
</tr>
<tr>
<td>SAT math</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>SAT verbal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Validity HE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>.078</td>
<td>.064</td>
<td>.095</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>.078</td>
<td>.064</td>
<td>.095</td>
</tr>
<tr>
<td>Purpose HE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>.098</td>
<td>.042</td>
<td>.147</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>.098*</td>
<td>.042</td>
<td>.147</td>
</tr>
<tr>
<td>Disillusionment HE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-.300*</td>
<td>.069</td>
<td>-.398*</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-.302*</td>
<td>.069</td>
<td>-.398*</td>
</tr>
<tr>
<td>Importance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Effort</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
</tbody>
</table>

*Note. Unst. = unstandardized; SE = standard error; St. = standardized. N = 758.
*p < .05
### Table 21

**Effects Decomposition for Model 5: Modified Fully-mediated Model with SAIAT-K-12 in Matched Incoming and First-year Student Sample**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance</th>
<th>Effort</th>
<th>Performance</th>
<th>Validity HE</th>
<th>Purpose HE</th>
<th>Disillusionment HE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.010*</td>
<td>.003</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.043*</td>
<td>.002</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.043*</td>
<td>.002</td>
</tr>
<tr>
<td>SAT verbal</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.314*</td>
<td>.033</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Validity K-12</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.328*</td>
<td>.040</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Purpose K-12</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.330*</td>
<td>.032</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Disillusionment K-12</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.333*</td>
<td>.055</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Validity HE</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.391*</td>
<td>.047</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Purpose HE</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>.664*</td>
<td>.118</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Disillusionment HE</td>
<td>Direct Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>1.699*</td>
<td>.227</td>
</tr>
<tr>
<td>Total Indirect Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total Effect</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
</tbody>
</table>

**Note.** Unst. = unstandardized; SE = standard error; St. = standardized. \( N = 678. \)

* \( p < .05 \)
Table 22

*Descriptive Statistics and Correlations among the SAIAT-HE-revised Subscales and Conscientiousness in the Mid-career A-Day 2 Student Sample*

<table>
<thead>
<tr>
<th></th>
<th>Validity</th>
<th>Purpose</th>
<th>Disil</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose</td>
<td>.480</td>
<td>.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disil</td>
<td>-.659</td>
<td>-.508</td>
<td>.76</td>
<td></td>
</tr>
<tr>
<td>Cons</td>
<td>.073</td>
<td>.063</td>
<td>-.141</td>
<td>.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Number of items</th>
<th>Mean</th>
<th>SD</th>
<th>95% CI</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>3.51</td>
<td>1.12</td>
<td>[3.38, 3.64]</td>
<td>-0.21</td>
<td>-0.36</td>
</tr>
<tr>
<td>Purpose</td>
<td>3</td>
<td>4.61</td>
<td>1.41</td>
<td>[4.45, 4.76]</td>
<td>-0.49</td>
<td>-0.14</td>
</tr>
<tr>
<td>Disil</td>
<td>3</td>
<td>4.98</td>
<td>1.23</td>
<td>[4.84, 5.12]</td>
<td>-0.17</td>
<td>-0.32</td>
</tr>
<tr>
<td>Cons</td>
<td>9</td>
<td>3.75</td>
<td>0.61</td>
<td>[3.68, 3.82]</td>
<td>-0.05</td>
<td>-0.77</td>
</tr>
</tbody>
</table>

*Note.* Disil - Disillusionment. Cons - Conscientiousness. SAIAT-HE-revised subscales range from 1 to 7; Conscientiousness subscale ranges from 1 to 5. Cronbach's alphas for each measure are on the diagonal. $N = 300$. 
Table 23

Descriptive Statistics and Correlations among the SAIAT-HE-revised Subscales and Conscientiousness in the Mid-career Make-up Student Sample

<table>
<thead>
<tr>
<th></th>
<th>Validity</th>
<th>Purpose</th>
<th>Disil</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose</td>
<td>.576</td>
<td>.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disil</td>
<td>-.660</td>
<td>-.503</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>Cons</td>
<td>.119</td>
<td>.147</td>
<td>-.122</td>
<td>.79</td>
</tr>
<tr>
<td>Number of items</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Mean</td>
<td>3.28</td>
<td>4.15</td>
<td>5.26</td>
<td>3.47</td>
</tr>
<tr>
<td>SD</td>
<td>1.3</td>
<td>1.65</td>
<td>1.45</td>
<td>1.3</td>
</tr>
<tr>
<td>95% CI</td>
<td>[3.08,</td>
<td>[3.90,</td>
<td>[5.04,</td>
<td>[3.38,</td>
</tr>
<tr>
<td>Skew</td>
<td>3.48]</td>
<td>4.40]</td>
<td>5.48]</td>
<td>3.57]</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.09</td>
<td>-0.53</td>
<td>0.05</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

Note. Disil - Disillusionment. Cons - Conscientiousness. SAIAT-HE-revised subscales range from 1 to 7; Conscientiousness subscale ranges from 1 to 5. Cronbach's alphas for each measure are on the diagonal. N = 169.
Table 24

*Mean Differences on the SAIAT-HE-revised Subscales across Compliant and Non-compliant Students Before and After Adjusting for Conscientiousness*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unadjusted</th>
<th>Adjusted for Conscientiousness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>M (SD)</td>
</tr>
<tr>
<td>Validity</td>
<td>3.51 (1.12)</td>
<td>3.28 (1.30)</td>
</tr>
<tr>
<td>Purpose</td>
<td>4.61 (1.41)</td>
<td>4.15 (1.65)</td>
</tr>
<tr>
<td>Disil</td>
<td>4.98 (1.23)</td>
<td>5.26 (1.45)</td>
</tr>
<tr>
<td>Cons</td>
<td>3.75 (0.61)</td>
<td>3.47 (1.30)</td>
</tr>
</tbody>
</table>

*Note.* Disil - Disillusionment. Cons - Conscientiousness. Mid-career A-Day 2 (Compliant) N = 300. Mid-career Make-up (Non-compliant) N = 169. *p < .05
Figure 1. Schematic of partially-mediated (top) and fully-mediated (bottom) path models.

Note. In a partially-mediated (top) model X and Y are related directly and also indirectly via M. In a fully-mediated model (bottom) X and Y are related only directly via M.
Figure 2. Model 1: Fully-mediated model.

Note. This model was fit to Matched incoming and first-year \((N = 678)\) and Mid-career A-Day 1 \((N = 758)\) student samples. Disill = Disillusionment; HE = Higher Education; Perform = Performance. All exogenous variables were allowed to correlate; correlation values for two samples are presented in Tables 16 and 17, respectively. The metrics of latent variables were set by fixing direct paths from observed single indicators to latent variables to 1. Ovals represent disturbance terms – percent of variance in the factor not explained by the variables affecting it. Model-data fit information for two samples is presented in Table 18. Standardized parameter estimates for the First-year sample are presented first; standardized parameter estimates for Mid-career sample are presented in parentheses.
Figure 3. Model 2: Partially-mediated model.

Note. This model was fit to Matched incoming and first-year \( (N = 678) \) and Mid-career A-Day 1 \( (N = 758) \) student samples. Disill = Disillusionment; HE = Higher Education. Perform = Performance. All exogenous variables were allowed to correlate; correlation values for two samples are presented in Tables 16 and 17, respectively. The metrics of latent variables were set by fixing direct paths from observed single indicators to latent variables to 1. Ovals represent disturbance terms – percent of variance in the factor not explained by the variables affecting it. Model-data fit information for two samples is presented in Table 18.
Figure 4. Model 3: Partially-mediated model with SAIAT-K-12.

Note. This model was fit to Matched incoming and first-year (N = 678) student sample. Disill = Disillusionment; HE = Higher Education. Perform = Performance. All exogenous variables were allowed to correlate; correlation values are presented in Tables 16. In order to account for interrelationships between Validity HE, Purpose HE, and Disill HE factors, the error covariances of these three factors were allowed to correlate. The metrics of latent variables were set by fixing direct paths from observed single indicators to latent variables to 1. Ovals represent disturbance terms – percent of variance in the factor not explained by the variables affecting it. Model-data fit information is presented in Table 18.
Figure 5. Model 4: Modified fully-mediated model.

Note. This model was fit to Matched incoming and first-year (N = 678) and Mid-career A-Day 1 (N = 758) student samples. Disill = Disillusionment; HE = Higher Education; Perform = Performance. Model 4 differs from Model 1 in one added direct path (bolded) and two omitted direct paths (from SAT math and SAT verbal to Importance). All exogenous variables were allowed to correlate; correlation values for two samples are presented in Tables 16 and 17, respectively. The metrics of latent variables were set by fixing direct paths from observed single indicators to latent variables to 1. Ovals represent disturbance terms – percent of variance in the factor not explained by the variables affecting it. Model-data fit information is presented in Table 18. Standardized parameter estimates for the First-year sample are presented first; standardized parameter estimates for Mid-career sample are presented in parentheses. See Tables 19 and 20 for all direct, indirect, and total effects. * p < .05.
Figure 6. Model 5: Modified fully-mediated model with SAIAT-K-12.

Note. This model was fit to Matched incoming and first-year (N = 678) student sample. Disill = Disillusionment; HE = Higher Education. Perform = Performance. All exogenous variables were allowed to correlate; correlation values are presented in Tables 16. In order to account for interrelationships between Validity HE, Purpose HE, and Disill HE factors, the error covariances of these three factors were allowed to correlate. The metrics of latent variables were set by fixing direct paths from observed single indicators to latent variables to 1. Ovals represent disturbance terms – percent of variance in the factor not explained by the variables affecting it. Model-data fit information is presented in Table 18. Model 5 differs from Model 4 in the inclusion of the SAIAT-K-12 variables. Standardized parameter estimates for direct effects are presented in the figure. See Table 21 for all direct, indirect, and total effects.

* p < .05.