

Exploring Change in Test-Taking Motivation

Carol L. Barry and Sara J. Finney

James Madison University

Paper presented at 2009 annual meeting of the Northeastern Educational Research Association

Correspondence regarding this paper should be addressed to: Carol L. Barry, Center for Assessment and Research Studies--MSC 6806, James Madison University, Harrisonburg, VA, 22807. E-mail: [barrycl@jmu.edu](mailto:barrycl@jmu.edu).

### Abstract

In this study, we expanded upon the work of Brown et al., (2009) by exploring the existence of types of test-takers characterized by qualitatively different patterns of test-taking effort across the course of a three-hour low-stakes testing session. Mixture modeling results did not indicate there were types of test-takers for this sample of upperclass examinees. Latent growth modeling results indicated that change in effort across the testing session was well-explained by a quadratic growth form. There was significant variability in effort for each of tests as well as in rates of change. The inclusion of external variables indicated that whether an examinee attended the regular testing session versus a makeup session, conscientiousness, and agreeableness explained variability in effort for the various tests, whereas only agreeableness was related to rates of change in effort. Implications of these results and directions for future research are discussed.

### Exploring Change in Test-Taking Motivation

In the United States, low-stakes testing is extremely prevalent given the accountability demands that are placed on all levels of education (National Assessment of Educational Progress; No Child Left Behind Act, 2002; U. S. Department of Education, 2006).

Unfortunately, in low-stakes contexts, examinees often realize that the tests have no impact on their grades, may feel the tests are unrelated to what they are learning in class, or may distrust the tests because they feel the tests only serve political purposes (Paris, Lawton, Turner, & Roth, 1991). When this is the case, a byproduct of low-stakes testing is that test-taking motivation may be low. Test-taking motivation is defined as the extent to which examinees 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” (Wise & DeMars, 2005, p. 2). In high-stakes situations, it is likely that most examinees have a high-degree of motivation to perform well, as the scores they receive on the test directly impact their lives. However, in low-stakes situations, test-taking motivation is likely to be much more variable among examinees, with some putting forth a high degree of effort and with others putting forth very little effort.

#### *Implications of Low Test-Taking Motivation*

Unfortunately, there are several negative implications for low motivation on tests. First, the scores obtained on these tests may underestimate the examinees’ actual levels of proficiency (Mislevy, 1995; Wainer, 1993; Wise & DeMars, 2005). Rather than representing what students actually know and can do, the test scores may instead represent only “what students will demonstrate with minimal effort” (O’Neil, Sugrue, & Baker, 1995/1996, p. 135). In fact, in a number of empirical studies, researchers have demonstrated a positive relationship between test-taking motivation and test performance (e.g., Arvey, Strickland, Drauden, & Martin, 1990;

Brown & Walberg, 1993; Sundre & Kitsantas, 2004; Wise & DeMars, 2005; Wolf & Smith, 1995). This impacts the validity of the inferences made about student learning as well as the soundness of decisions regarding the effectiveness of educational programming.

Second, low motivation impacts the properties of test and item statistics. Because knowledge of test and item statistics is necessary for understanding the way a given test functions and how it should be scored, low motivation impacts this aspect of validity. For example, item difficulties estimated under low-stakes conditions may not represent item difficulties estimated in high-stakes conditions (DeMars, 2000). Further, Wise (2006) found that including examinees with low-motivation artificially increased reliability coefficients and artificially decreased the correlation coefficients used for external validity evidence of scores. In short, low motivation negatively impacts not only what we know about students, but also what we know about the properties of the tests they complete.

### *Perspectives for Understanding Test-Taking Motivation*

There are two main perspectives that are used throughout the literature to understand test-taking motivation: 1) Expectancy-Value (EV) theory, which comes from the literature on achievement motivation; and 2) Fatigue, which comes more from the applied literature on testing. These perspectives offer helpful frameworks for interpreting the literature on test-taking motivation and the factors that impact it. Fully understanding the factors that impact motivation can aid testing and assessment practitioners in creating tests and testing programs that maximize test-taking motivation.

#### *Expectancy-Value Theory*

In general, EV theory states that the effort individuals give to a particular task is influenced by their expectancies regarding their performance on the task and the degree to which

they value the task (Atkinson, 1957; Eccles et al., 1983; Eccles & Wigfield, 2002; Wigfield & Eccles, 1992, 2000). The most prevalent conceptualization of this theory is that of Eccles and colleagues (Eccles et al., 1983; Eccles & Wigfield, 2002; Wigfield & Eccles, 1992, 2000), who posit that expectancies consist of two components: (1) *expectancies for success*, which is defined as individuals' beliefs about how well they will perform upcoming tasks, either in the immediate or long term future; and (2) *ability beliefs*, which is defined as individuals' beliefs about their current competence in a given domain. In addition, values are posited to consist of four components: (1) *attainment value*, which is defined as the importance of doing well on a given task to the individual; (2) *intrinsic value*, which is defined as the enjoyment that the individual gets from doing the task or activity; (3) *utility value*, which is defined as how the task or activity relates to future goals and plans; and (4) *cost*, which is defined as the negative aspects of engaging in a given task and includes judgments of how the task limits access to other activities, how much effort is needed to successfully complete the task, and the emotional states associated with the task.

When considering the research on test-taking motivation, it has been empirically demonstrated that test item difficulty or the degree of mental taxation is related to test-taking motivation, with more difficult tasks generally eliciting lower motivation in low-stakes conditions (Bovaird, 2002; Wise, 2006; Wolf et al., 1995). These findings can easily be interpreted through the lens of EV theory in that more difficult tasks should be associated with higher cost, lower expectancies for success, and, therefore, lower motivation. Somewhat related to this is that test-taking motivation has also been shown to relate to the type of test the examinee completes. Examinees tend to be less motivated on constructed-response tests (i.e., examinees must *provide* an answer) than on selected-response tests (i.e., examinees must *select* the correct

answer from a list; DeMars, 2000; Sundre & Kitsantas, 2004). Again, EV theory provides a useful framework for interpreting these findings in that constructed-response items tend to measure higher levels of cognitive complexity and require more effort than selected-response items, thus having a high degree of cost and lower expectancies for success.

### *Fatigue*

The second perspective often used for understanding test-taking motivation is that of fatigue, which can be defined as the extent to which examinees become tired or bored and fail to attend to later items on a test or later tests in a testing session. The fatigue perspective comes from the literature on test design as well as observations of examinee behavior. In general, the longer the examinee is required to perform either on a single test or on a set of tests, the greater the decrease in motivation. Even in low-stakes testing situations, examinees may initially be curious about the test and give effort when responding, but may eventually lose interest and begin guessing as the test progresses (Cao & Stokes, 2008). Research findings that align with fatigue include studies in which researchers show that, within a given test, the location of a given item is related to motivation. Specifically, examinees tend to have lower test-taking motivation for items near the end of the test than for items near the beginning of the test (Bovaird, 2002; Cao & Stokes, 2008; Wise, 2006).

### *Do Types of Test-Takers Exist?*

When reading the empirical studies of test-taking motivation, one will find that the large majority of the work focuses on motivation for a group of examinees as a whole. That is, test-taking motivation is often examined for the entire sample, and the results are reported at an aggregate level. Furthermore, assessment practitioners often report and discuss test-taking effort for an entire group of students. For example, assessment practitioners might make statements

like “motivation for this group of examinees was low” or “motivation was not an issue.” One limitation of this type of reporting is that it ignores the possible existence of different *types* of examinees. It seems quite likely that examinees would vary from one another in the degree of effort they give to test items, tests, or set of tests, especially in low-stakes conditions, leading to the existence of different types of test-takers.

The possible existence of different test-taking types is supported by the theoretical perspectives that are often used to understand test-taking motivation. For example, using the fatigue perspective, one would expect examinees to be less motivated as the test or testing session progresses. However, it seems reasonable that some examinees may become fatigued sooner than others, whereas others might not become fatigued at all (e.g., some examinees may enjoy being cognitively stimulated and not decrease in motivation). Further, EV theory posits that examinees will display levels of test-taking motivation based on 1) their expectancies to successfully perform on the item or test, 2) their ability beliefs for the domain, as well as 3) the extent to which they value completing the item or test. Given that examinees differ in their academic strengths and weaknesses as well as their values, it seems very reasonable that they will have differential patterns of motivation across test items or across a testing session.

Most of the studies that have examined the existence of types of test-takers have done so at the item level. That is, they have used mixture IRT models to classify examinees into either a solution-behavior class (i.e., motivated) or a rapid-guessing behavior class (i.e., low motivation) on the basis of each individual’s response times to the test items. The results of these studies indicate the mixture IRT models do provide good fit to the data, thereby providing evidence for the existence of test-taking types based on different patterns of item-level motivation (Bovaird, 2002; Cao & Stokes, 2008; Meyer, 2008).

To our knowledge, only one group of researchers has examined the existence of different types of test-takers across the course of an entire testing session (Brown, Barry, Horst, Finney, & Kopp, 2009). These authors used mixture modeling (MM) to examine test-taking motivation for freshman students across the course of a three-hour, low-stakes testing session, during which examinees completed two non-cognitive tests, a very difficult cognitive test of quantitative and scientific reasoning, and then two additional non-cognitive tests, providing a self-report measure of test-taking effort for each test. Three types, or classes, of examinees characterized by different patterns and levels of motivation across the five tests were uncovered. The first class consisted of examinees with extremely high motivation for the first two non-cognitive tests, a substantial drop in motivation for the difficult, cognitive test, and then very high motivation for the final two non-cognitive tests (See Figure 1). The second class consisted of examinees that had the same pattern of motivation scores across the five tests, but motivation was lower overall in magnitude. Thus, classes 1 and 2 represent quantitatively ordered classes in that they had the same pattern of effort across the testing session, but they differed by a matter of degree. If these ordered classes were the only ones to emerge, it would suggest that qualitatively distinct test taking types do not exist, and thus, the aggregate data could be used to estimate the average change in motivation across the testing session; if this were the case, researchers could then employ a technique such as latent growth modeling to model this average change in motivation. However, the third class reported moderate levels of motivation for all five tests; that is, examinees in class 3 did not report lower motivation for the cognitive test. Thus, the third class represented a qualitatively distinct type of test-taker because the pattern of effort scores across the five tests differed from the other two classes.

Further evidence of the distinction between these classes was gathered when examining whether the classes could be differentiated by external variables. It was found that examinees in the first class (i.e., the “high dippers”) were higher than the other two classes on mastery approach goal orientation (i.e., seek to gain competence) and performance approach goal orientation (i.e., seek to demonstrate competence relative to others), agreeableness, conscientiousness, and openness. That is, individuals in class 1 had more favorable levels of the external variables than did the other two classes. The third class (i.e., the “flat liners”) was differentiated from the other two classes (i.e., the classes that dipped in motivation for the difficult, cognitive test) in that examinees in this class had higher levels of math ability than the other two classes.

Interestingly, the authors noted that contrary to their prediction, the results of this study did not appear to align with the fatigue perspective in that there was no class that had a steady decline in motivation across the course of the testing session (Brown et al., 2009). The EV perspective, on the other hand, was quite useful for interpreting the results, especially for understanding the different patterns of effort associated with the third class (i.e., examinees with no drop in motivation and higher levels of quantitative ability). Given the higher math ability of those in Class 3, these individuals may have not exhibited a decrease in motivation for the quantitative cognitive test either because 1) the higher ability resulted in higher expectancies, 2) the higher ability resulted in less of a cost associated with the test and thus higher value, or 3) some combination of the two.

#### *Need for Additional Studies of Test-Taking Types*

If average test-taking motivation is modeled for a series of tests, understanding the average form of growth (e.g., liner, nonlinear) could be useful in that it would allow researchers

to better understand the average change in motivation and examine external variables that would predict variability in this growth. This would allow testing programs and assessment professionals to understand *why* examinees have different initial levels of motivation (e.g., high versus low) as well as *why* examinees change at a greater rate than others (e.g., more rapid declines in motivation). Understanding these issues is especially useful given that this could allow testing programs or assessment professionals to design interventions or marketing strategies in the hopes of increasing motivation. However, examining the average test-taking motivation for an entire group overall, researchers may be ignoring the existence of qualitatively distinct test-taking types that are characterized by different forms of growth and differences in competence and personality. Although one study has examined this issue across a large-scale low-stakes testing session (Brown et al., 2009), additional research is needed for several reasons.

First, the sample utilized in the one existing study consisted of only freshmen students who completed the testing prior to the start of their college experience (weekend before classes started). Given that expectancies and values are formed in a social context (Eccles et al., 1983), the test-taking motivation for low-stakes tests could be quite different for incoming students compared to students who have been on campus for several semesters. Thus, it is necessary to examine whether the results from the freshman sample generalize to students who have been on campus for several semesters and have prior experience with these low-stakes tests. Perhaps incoming freshmen students mistakenly believe the tests *do* have consequences for them whereas upperclassman students realize that these tests do not impact them personally; as a result, this might lead upperclassman students to have different values and therefore exert different levels of motivation than freshman students. Alternatively, although there may be a “culture of assessment” among the faculty and administrators, it is possible that there is a “counter-culture

of assessment” among the students in that the longer students are on campus, the more they may devalue the low-stakes that tests they are required to complete. For this reason, it is possible that students who have been on campus for several semesters view these assessments as less important or relevant than do freshman students, thereby leading to lower levels of value for these tests and, thus, lower motivation.

Second, in the single existing study, the researchers only examined one configuration of tests (i.e., noncognitive, noncognitive, cognitive, noncognitive, noncognitive). It would be interesting to explore the pattern of motivation scores for a different configuration of tests and whether this would result in the identification of different types of test-takers. Given the lack of study of test-taking types, one would take an exploratory approach to answering questions such as this. However, one could hypothesize that by simply moving the cognitively demanding test, the profile of motivation would change, with motivation being lower for the cognitive test for at least one class relative to the noncognitive test. Regardless, given the limited study thus far, this needs to be explicitly examined.

### *Research Questions*

Understanding test-taking motivation is critical to making valid inferences from test scores, and for this reason, Standard 15.4 of the Standards for Educational and Psychological Testing recommends that test programs report information on examinee motivation (AERA, APA, & NCME, 2004). However, much of the existing literature is limited in that it focuses on motivation at an aggregate level (i.e., average level of motivation for whole sample). Thus, the current study seeks to address some of the limitations of the current literature on test-taking motivation by exploring the existence of test-taking types characterized by differing levels of and change in motivation across the course of a large-scale testing session.

*Question 1.* Are there distinct types of test-takers? If there *are* types of examinees for whom patterns of test-taking motivation are different, what variables differentiate the types/classes of examinees from one another?

*Question 2.* If there are *not* types of examinees for whom patterns of test-taking effort are different, what is the overall average trajectory or form of growth (i.e., linear, nonlinear)? What variables predict the variability in the change in motivation?

### *Method*

#### *Participants and Procedures*

The sample used for the current study consists of upperclass students who completed a three-hour testing session during the Spring 2009 semester. Every university student is required to participate in two university-wide assessment days. The first assessment day, for entering students, occurs during the week before fall classes begin, and the second occurs during the spring semester when students have accumulated 45-70 credit hours (i.e., typically sophomore or junior status). During these assessment days, students must complete a series of general education and developmental or attitudinal measures. Trained proctors administer the instruments and read instructions aloud before students begin responding. If students do not attend the scheduled assessment day, they are contacted and required to complete the assessments during a makeup session.

Approximately 3,320 students attended either the scheduled Spring 2009 assessment day or one of three makeup sessions. Only subset of this total sample completed the set of measures used in the current study ( $N = 683$ ), and of this group 462 examinees attended the assessment day and 221 attended a makeup session. These participants completed a set of five measures, each consisting of approximately 60 items. The first was a cognitive test (i.e., measures knowledge,

skills, or ability and has items that are scored as right or wrong) designed to assess quantitative and scientific reasoning. The remaining four tests were non-cognitive tests that measured attitudes or affect. After completing each of the five measures, participants responded to a set of items about their test-taking motivation for the test they just completed. It is these motivation scores that were modeled to address the research questions. Finally, after completing all five measures, participants completed an additional measure designed to assess math efficacy expectations. Demographic information about the sample is included in Table 1.

### *Measures*

*Test-taking motivation.* The Student Opinion Scale (SOS: Thelk, Sundre, Horst, & Finney, in press; Sundre & Moore, 2002) was used as a measure of test-taking motivation for each of the five tests (see Appendix A). This instrument consists of two subscales. The Effort subscale consists of five items that measure the degree to which examinees put forth effort on a given test, whereas the Importance subscale consists of five items that measure the degree to which examinees view a given test as important. Examinees responded to items using a scale of 1 (*Strongly Disagree*) to 5 (*Strongly Agree*), and items were averaged for each subscale to create two subscale scores, with higher values (i.e., closer to 5) indicating higher degrees of effort and importance. Only scores from the Effort subscale were used as a measure of motivation.

### *External Variables*

Data from a variety of additional measures were used either to gather external validity evidence for any classes that emerged (i.e., explain differences between the types of test takers) or to predict individual variability in change in motivation for the entire group of test takers if no classes emerged (i.e., explain differences in starting level of motivation or in rates of change in motivation across the five tests). These measures include, but are not limited to, the cognitive test

and some of the non-cognitive tests that participants completed during the assessment day testing session. Example items for each of the noncognitive measures listed below are included in Appendix B.

*Quantitative ability.* Quantitative ability was measured using two measures. The first measure of quantitative ability was the Natural World test (NW-9: Sundre, Thelk, & Wigtil, 2008), which is a 66-item cognitive test designed to assess students' quantitative and scientific reasoning skills. Each item was dichotomously scored, and the scored responses were summed to create a total NW-9 score. The NW-9 was the cognitive general education test that was administered first during the testing session. It is important to note that this is a very demanding test of quantitative and scientific reasoning.

SAT math scores, obtained from university records, served as a second measure of quantitative ability. This second measure of ability was gathered because scores on the cognitive test and motivation for that test cannot be completely disentangled. That is, for students with low test-taking motivation for the cognitive test, the scores on that test may not serve as accurate representations of their ability level.

*Math efficacy expectations.* Math efficacy expectations were measured using the Sources of Middle School Mathematics Self-Efficacy Scale (MSES: Usher & Pajares, 2009). This 24-item instrument consists of four 6-item subscales representing the four sources of self-efficacy described by self-efficacy theorists: Mastery Experience, Vicarious Experience, Social Persuasions, and Physiological State. Although this measure was developed for use with middle school students, an examination of the items suggested that it would also be applicable to high school and college students. Items were averaged for each of the four subscales to create four subscale scores and these scores served as proxies for expectancies.

*Need for cognition.* Need for cognition was measured using the Need for Cognition Scale (NCS: Cacioppo, Petty, & Kao, 1984). This is an 18-item instrument designed to measure an individual's need to be cognitively stimulated. Examinees responded to a series of statements using a scale of 1 (*Extremely Uncharacteristic*) to 5 (*Extremely Characteristics*). Responses to each of the 18 items were averaged to create a total score, with higher values representing a higher degree of need for cognition. Examinees with a higher need to be cognitively stimulated should report higher levels of motivation both initially and throughout the testing session.

*Achievement goal orientation.* Students' achievement goals were measured by the Attitudes Towards Learning questionnaire (ATL: Finney, Pieper, & Barron, 2004; Pieper, 2004). This is a 16-item instrument, consisting of five subscales: mastery approach (MAP), performance approach (PAP), mastery avoidance (MAV), performance avoidance (PAV), and work avoidance (WAV). Participants respond to a series of statements using a scale from 1 (*Not at all true of me*) to 7 (*Completely true of me*). Items were summed to create five subscale scores, with higher scores indicating higher levels of each achievement goal. Examinees with higher MAP and PAP should report higher levels of test-taking motivation than those with lower scores. Further, examinees with higher MAV and PAV (i.e., those who tend to be threatened by poor performance), and WAV (i.e., those who avoid doing work) should report lower test-taking motivation than those with lower scores.

*Personality.* Participants' personality characteristics were measured by the Big Five Inventory (BFI: John & Srivastava, 1999). The BFI consists of 44 items that represent five dimensions of personality. Thus, the measures consist of five subscales: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Participants responded to a series of statements using a scale from 1 (*Disagree Strongly*) to 5 (*Agree Strongly*), and

responses were averaged to create five subscale scores, which higher scores indicating higher levels of each personality dimension. Examinees who are more agreeable (i.e., compliant), conscientious (i.e., achievement oriented), and open to experience (i.e., curious) should report higher levels of test-taking motivation.

## *Results*

### *Descriptive Statistics*

Descriptive statistics for all variables are reported in Table 2 and a plot of mean effort scores is shown in Figure 2. Note that effort scores were the lowest for the first test (i.e., cognitive test) and were the highest for the fourth test (i.e., noncognitive test containing measures of personality, well-being, worry, and sense of identity). When examining these correlations (Table 2), effort scores for each of the five tests were moderately- to highly-correlated; however, of these, the correlations between effort for the first test (cognitive) and effort for the remaining four tests (noncognitive) were lowest. Also, relative to the size of the other correlations, the correlations between effort scores and the personality variables (i.e., agreeableness and conscientiousness) and achievement goal variables (i.e., MAP, and WAV) were relatively strong. Interestingly, the correlation between SAT math scores and effort scores for the cognitive was very small. This was somewhat surprising given that Brown et al. (2009) found SAT math scores to differentiate between examinees of high effort versus low effort. The small correlation observed here may indicate that SAT math score may not help understand test-taking motivation on the cognitive test as well as it did previously (Brown et al., 2009). Finally, although expectancy-value theory would suggest that math efficacy expectations should be highly correlated with motivation (i.e., effort), for these data these correlations were actually quite small.

*Research Questions 1: Are there distinct types of test takers?*

*MM Specification.* The goal of MM is to identify types, subgroups, or *classes* of individuals who have similar values on a set of variables. In these situations the observed data are thought to be sampled from a population composed of a *mixture* of distributions. That is, the population is thought to be composed of a number of subpopulations or classes of people. MM is a latent variable technique in that individuals' values on a latent categorical variable (i.e., class membership) drive their responses to observed variables; thus, these observed variables are called *indicators* of class membership. MM makes it possible to estimate unique parameters (e.g., means, variances, and covariances) for each identified class but also allows for some of the parameters to be constrained to be equal across classes. Thus, models can vary not only in the number of classes, but also in parameterization. Additionally, MM provide estimates for the mixing proportion, or the weight of each class in the population (i.e., proportion of sample belonging to each class), as well as posterior probabilities (i.e., the probability of belonging in each class for each person).

For the current analyses, the five SOS Effort scores were used as indicators for the latent categorical variable of test-taking type. A series of MMs that varied in both parameterization (Model A through Model D) and number of classes were estimated (i.e., 1-, 2-, and 3-class models). In Model A, means for the five Effort scores were estimated freely for each class, but the variances were constrained to be equal across classes and the covariances were constrained to be zero. Model A forces local independence and was the most parsimonious model tested. In Model B, the Effort means and variances were estimated freely for each class, and covariances were again constrained to be zero (i.e., forcing local independence). In Model C, means and variances were freely estimated for each class and covariances were estimated but constrained to

be equal across classes. Finally, in Model D, the means, variances, and covariances were estimated separately for each class; Model D was the most complex parameterization. In addition, for all models (regardless of parameterization),  $k - 1$  ( $k =$  number of classes) mixing proportions were estimated, representing the proportion of the total sample that is in each class.

Model-data fit was evaluated using the Akaike Information Criterion (AIC: Akaike, 1987), the Bayesian Information Criterion (BIC; Schwarz, 1978), the Sample Size Adjusted BIC (SSA-BIC: Sclove, 1987), and the Lo Mendell Rubin likelihood ratio test (LMR: Lo, Mendell, Rubin, 2001). The AIC, BIC, and SSA-BIC were used to compare models that differed in both parameterization and the number of classes with smaller values indicating better relative fit. Alternatively, the LMR test was used to compare models of the same parameterization but that differed in the number of classes; a non-significant LMR test indicates the model with fewer classes should be championed. The SSA-BIC has been shown to function well and was given the most weight in deciding which model to champion (Tofighi & Enders, 2007). Additionally, posterior probabilities, the classification table, and entropy statistic were evaluated to assess each model's ability to classify individuals. The class means were plotted to assess if the classes differed quantitatively (i.e., class means have the same pattern and differed only by a matter of degree) or qualitatively (i.e., class means have different patterns).

*MM Results.* The MM were estimated using Mplus, version 5.2. Model fit is reported in Table 3. Importantly, the best log-likelihood value was not replicated for the two-class model C and three-class model D, and the three-class model C failed to converge to an admissible solution; thus, fit indices for these models are not reported.

Of the models tested, the two-class model D had the best relative fit. However, the second class in this solution consisted of only 7% of the total sample. Additionally, preliminary

examination of the external variables indicated that the two classes were not differentiated by any of the external variables. Taken together, this suggests that the classes in this solution do not represent meaningful subgroups, and, for this reason, this model was not championed. Given the three-class model B also had good relative fit, we also explored this solution. Examination of the class means indicated that the three classes had the same pattern of means across the five test administrations and thus differed only by degree. More specifically, the three classes had the exact same pattern of means as that shown in Figure 2; the first class had lower means, the second class had means approximately equal to those shown in Figure 2, and the third class had higher means. Additionally, examination of the external variables indicated a similar pattern. That is, individuals in class one (i.e., lowest effort) had the lowest levels of favorable external variables (e.g., Mastery Approach, Agreeableness, Conscientiousness) and the highest levels of unfavorable external variables (e.g., Work Avoidance); individuals in class three (i.e., highest effort) had the highest levels of favorable external variables and the lowest levels of unfavorable external variables. Taken together, these results indicate that these three classes represent quantitatively ordered groups rather than qualitatively distinct types of test-takers. That is, it appears there is one pattern of test-taking effort across the five tests and individuals vary in their levels, with some reporting higher effort or lower effort across the five tests. The three-class solution seems to be an artificial categorization of this continuum. This suggests that there are *not* types of test-takers. Given these results, the focus of the study shifted from describing the types of test-takers to exploring how motivation changes across the course of the testing session for the entire sample.

*Research Question 2: What is the Overall Trajectory of Test-Taking Motivation?*

*Latent growth modeling (LGM) specification.* Given the MM results did not indicate that there were types of examinees, LGM was used to evaluate whether a growth structure could be used to describe the trajectory of motivation scores for the entire sample. The benefit of applying a growth structure to the motivation scores is that it allows one to model the average starting motivation (i.e., motivation for the first test), the average rate of change in motivation (e.g., steady decline in motivation, a slight increase and then a gradual decrease), as well as whether there is variability in both the initial values and rates of change (i.e., do individuals vary in their initial motivation and change in motivation?). Further, the relationship between initial values and rates of change can be examined (e.g., do individuals who have higher effort scores for the first test have less change in their effort scores throughout the testing session?).

Two alternative LGMs were fit to the data to examine various forms of growth (e.g., linear, nonlinear). In these models, both the number of factors and the loadings on those factors described the form of growth. In the models tested in this study, we specified the form of growth *a priori* by fixing the factor loadings on the factors to set values; specifically, we tested models for linear growth (Figure 3) and for quadratic growth (Figure 4).

For both models, factor means and variances are estimated, indicating average initial motivation levels and rates of change as well as the amount of individual variability in those initial levels and rates of change. For the linear growth model, two factor means are estimated. The intercept mean indicates the average Effort score for the test administration coded 0 (e.g., initial Effort), whereas the linear slope mean indicates the average linear change in Effort scores across the five test administrations. For the quadratic growth model, three factor means are estimated. The intercept mean indicates the average Effort score for the test administration coded 0, the linear slope mean indicates the average change in effort between the test administrations

coded 0 and 1, and the quadratic slope mean indicates the average change in that linear slope across the course of the testing session. It is important to note that although the intercept is often set to the initial administration (i.e., effort for the first test), it can be set to any administration. In this case, the intercept mean is interpreted as the average effort score for that test administration and intercept variability is interpreted as the amount of variability in effort scores for that test administration.

Given the data were multivariate nonnormal (i.e., Mardia's normalized multivariate kurtosis value exceeded the recommended cutoff; Finney & DiStefano, 2006), the Satorra-Bentler adjusted chi-square, fit indices, and robust standard errors were used. Thus, model-data fit was evaluated using the Satorra-Bentler adjusted  $\chi^2$  statistic in conjunction with several other global fit indices: the Satorra-Bentler comparative fit index ( $CFI_{S-B}$ ), the Satorra-Bentler root mean square error of approximation ( $RMSEA_{S-B}$ ), and the standardized root mean square residual (SRMR). When data are nonnormal, it has been suggested that  $CFI_{S-B}$  values greater than .95,  $RMSEA_{S-B}$  values smaller than .05, and SRMR values smaller than .07 are indicative of adequate model-data fit (Yu & Muthen, 2002). Additionally, local fit was evaluated by examining the standardized mean and covariance residuals, with residuals with absolute values of 3 or greater indicating areas of local misfit (Raycov & Marcoulides, 2000).

Fit indices for these models are reported in Table 4. The fit indices for the linear model suggested that it did not fit the data globally (Table 4). Further, there were numerous standardized residuals for both the means and the covariances that were larger in absolute value than 3, suggesting there were substantial areas of local misfit. For the quadratic model, although the  $RMSEA_{S-B}$  for the quadratic model was slightly higher than the recommended cut off, the other fit indices for this model indicated adequate fit (see Table 4). Most standardized covariance

residuals for this quadratic model were all very small (i.e., less than 3 in absolute value); interestingly, the standardized residuals for the mean effort scores were somewhat large for the third and fourth test administrations (i.e., -9 and 11, respectively), indicating that was a bit of a discrepancy between the observed and model-implied means for these two tests (see Figure 2). However, given that the remainder of the standardized residuals were very small and given that the CFI and SRMR fell within the recommended cutoffs, we felt the fit of this model was adequate enough to interpret the parameter estimates. Parameter estimates for this model (i.e., *Unconditional Quadratic Model 1*) are reported in Table 5. The mean for the intercept factor indicated that, on average, effort for the first test was approximately 3.6; given the scale ranges from 1 to 5, this indicates that individuals put forth moderate levels of effort for this difficult, cognitive test. The mean for the linear factor was positive, indicating that on average, effort increased between the first (i.e., cognitive) and second (i.e., noncognitive) tests by approximately 0.19 points. However, the quadratic factor mean was negative, indicating that as the testing session progressed, the change in effort became less positive (i.e., curved downward). This pattern can be seen when examining the plot of effort means (Figure 2). All factor variances were significant, indicating significant individual variability in levels of effort for this first test and in rates of change. Further, the correlation between the linear and quadratic factors was negative and significant. Given the positive mean for the linear factor and the negative mean for the quadratic factor, this indicates that greater linear change is associated with less quadratic change. That is, individuals that have greater increases in effort between the first and second tests have less of a subsequent decrease in effort across the remaining tests.

Given the significant factor variances, we would incorporate external predictors to account for this variability in initial levels and rates of change. However, before doing so, we

wanted to estimate the variability in effort scores for each of the remaining four tests in order to evaluate whether these scores could also be predicted. Fortunately, LGM provides an elegant way of estimating this in that the intercept can be set to any of the five test administrations. By setting the intercept to the first, second, third, fourth, and fifth test administrations, one can then estimate the intercept mean and variance as the mean and variability in effort for each test. The main point in moving the intercepts was to not to estimate the intercept means, as these means are simply the model-implied means for this model (see Tables 5 and 6). Rather, the advantage of moving the intercepts was to estimate the intercept *variability*, which allowed us to examine whether there was significant variability in effort for each of the five test administrations. As expected given the model-implied means, the intercept means indicated that overall, effort was higher for these four noncognitive tests than for the cognitive test (i.e., the first test). Most importantly, there was significant variability in effort for all tests in the testing session (Table 6).

*Conditional LGM results.* External variables were incorporated into the model as predictors of the intercept, linear, and quadratic factors. All continuous predictors were grand mean centered to aid in interpretability, and given the number of predictor variables included in the model, the alpha level was adjusted to 0.01 to control for Type I errors. We were also interested in determining what predicted variability in effort for each of the five tests in the session. Thus, we also tested several models wherein we set the intercept to the first, second, third, fourth, and fifth test administration. The list of potential predictors was quite extensive, so an effort was made to include only those that were relevant. Expectancy-value theory states that one's expectancies for the task should impact motivation. Thus, the four subscales representing math efficacy expectations (i.e., Mastery Experiences, Vicarious Experiences, Social Persuasions, and Physiological State) were initially included. This allowed us to determine

whether effort on the difficult, cognitive test of quantitative and scientific reasoning (i.e., the intercept) was related to math expectancies. A series of preliminary models were fit to the data to test the utility of all external variables as predictors of intercepts or rates of change, but in the interest of space, only the final model will be presented in this paper. As foreshadowed by their low bivariate correlations with effort scores, need for cognition, NW9 scores, SAT math scores, PAP, MAV, PAV, extraversion, neuroticism, and openness were not significant predictors in a statistical sense, a practical sense, or both. Although MAP and WAV had moderate correlations with effort scores, when included as predictors (either alone, or when included with other variables), these variables were either not statistically significant or not practically significant predictors. MAP and WAV were moderately correlated with one another ( $r = -.512$ ) and thus may have lost their predictive utility due to multicollinearity. Interestingly, none of the four expectancy subscales predicted variability in either initial levels of effort or rates of change in either a statistically or practically significant manner. Given the relatively low correlations between these variables and effort scores (Table 2), this is not entirely unexpected. Based on the results of the preliminary models, the final conditional model (*Conditional Quadratic Model*) included only makeup status (i.e., attended the assessment day versus attended a makeup session), agreeableness, and conscientiousness as predictors of the growth factors (see Figure 5 for a graphical depiction of this model).

Fit indices for the final conditional quadratic model are presented in Table 4. The fit indices indicated good global fit of the model, and all but two standardized mean and covariance residuals were below 3 in absolute value. Similar to the unconditional model, the standardized residuals for the mean effort scores were somewhat large for the third and fourth test administrations (i.e., -8 and 10, respectively), indicating that the model did not reproduce these

means particularly well. This is not surprising given the model-implied means for the conditional model are identical to those for the unconditional model (see Tables 5 and 7). Despite this, we again considered the model to have adequate fit because the remainder of the standardized residuals were very small and because the global fit indices fell within the recommended cutoffs. Thus, we examined the parameter estimates to assess whether the external variables predicted effort for the first test and/or rates of change in effort (Table 7). For this model, both makeup status and conscientiousness significantly predicted the intercept. Specifically, when controlling for agreeableness and conscientiousness, individuals that attended a makeup session had a mean effort score on the Cognitive test .156 points lower than those that attended the scheduled assessment day. Although this difference is statistically significant, recall that effort is on a 1-5 scale. Thus, this difference between these groups in effort for this test is quite small from a practical standpoint. Additionally, when controlling for makeup status and agreeableness, examinees with higher levels of conscientiousness put forth more effort on this test than those with lower levels of conscientiousness. Finally, agreeableness significantly predicted both linear and quadratic factors. To aid in the interpretation of these relationships, change in effort scores is plotted for several different levels of agreeableness (Figure 6). After controlling for makeup status, and conscientiousness, individuals with higher agreeableness scores tended have a greater increase in effort between the first test and fourth test, where effort peaked, and tended to have a greater subsequent decrease in effort between the fourth and fifth test than less agreeable examinees.

Parameter estimates describing the relationships between the predictor variables and effort scores for all five test administrations are reported in Table 8. By again setting the intercept to each test administration and including the makeup status, conscientiousness, and

agreeableness, we were able to examine whether these variables could account for variability in effort for these five tests. Although for each of these models the relationships between the external variables and all three factors were estimated, the relationships of primary interest are those between the external variables and intercept (i.e., effort); thus, the values reported in Table 8 reflect only this. When examining these coefficients, especially in relation to those for the model where the intercept was set to the first (i.e., cognitive) test, several things are of interest. First, for all tests, examinees of average conscientiousness and average agreeableness but who attended a make-up session (i.e.,  $\text{makeup} = 1$ ) reported significantly lower effort than those who attended the scheduled assessment day. Interestingly, the difference in effort scores was the smallest for the first, cognitive test. Second, conscientiousness was significantly and positively related to effort for all five tests in the testing session. That is, for examinees of average agreeableness who attended the scheduled assessment day, more conscientious individuals reported higher effort for all five tests than less conscientious individuals. Third, after controlling for the other variables in the model, agreeableness did not predict effort for the first test, but *did* predict effort for the remaining four tests. Specifically, for examinees of average conscientiousness who attended the scheduled assessment day, more agreeable examinees tended to report higher effort for these tests than less agreeable examinees. Finally, overall, the relationships between the predictors and effort scores varied to some extent across the five tests. They appear to be the weakest for the first test, possibly suggesting that there are other factors related to how much effort examinees give on this difficult, cognitive test. The relationships appeared to be the strongest for the last test in the session, perhaps indicating that as individuals become tired or bored (i.e., fatigued), these variables become more salient in explaining variability in effort.

### *Summary of Results*

In summary, for this data, there did *not* appear to be types or classes of test-takers characterized by different patterns of effort scores across the five tests. Rather, there appeared to be a single pattern of effort scores for the entire sample. This pattern of effort scores followed a quadratic growth form, with an initial increase in motivation between the first and second test, a gradual slowing of this increase between the second and fourth test, and then a decrease in motivation from the fourth to the fifth test. There was significant variability in this change over time, and examinees higher in the personality trait of agreeableness tended to have a greater initial increase and a greater subsequent slowing. Additionally, makeup status and conscientiousness significantly predicted effort scores for each of the five tests, with examinees who attended a makeup session reporting lower effort than those who attended the scheduled assessment day and with examinees higher in the personality trait of conscientiousness reporting greater effort than those lower in conscientiousness. Finally, although agreeableness did not predict variability in effort for the first test, it did predict for the remaining four tests, with more agreeable examinees reporting higher motivation than less agreeable examinees.

### *Discussion*

It is important to consider that examinees may vary in their levels of test-taking motivation and these levels may vary across the course of a testing session. Thus, the purpose of the current study was to build upon previous work (Brown et al., 2009) by exploring the possible existence of types of test-takers, to examine whether and how test-taking motivation changed across different tests administered across the course of a testing session, and to explain any existing variability in that change in motivation.

### *Do Types of Test-Takers Exist?*

The mixture modeling results of the current study indicated that there did *not* appear to be types of test-takers characterized by different patterns of test-taking motivation. More specifically, the classes that were identified were quantitatively ordered in that they all had the same pattern of effort across the five tests, but simply differed by degree (i.e., class 1 had the lowest effort means, class 3 had the highest effort means, and class 2 was in the middle). Thus, for these data, there appears to be a single pattern of effort across the course of a testing session and variability around that effort. This finding was especially surprising because previous study *did* find qualitatively distinct types of test-takers (Brown et al., 2009). However, the current study differed from the Brown et al (2009) in several ways. The first, and perhaps most tenable, reason for the differing results involves the samples utilized. Whereas the previous study examined test-taking motivation for freshman students (Brown et al., 2009), the current study utilized a sample of upperclass students. It is possible that the longer students remain on campus, the more similar they become with regard to what they are willing to expend effort on. Under expectancy-value theory, motivation for a given task is a product of individuals' expectancies for success and the extent to which they value the task. Both of these components are formed in social contexts (Eccles et al., 1983). Thus, it is possible that the expectancies and values of upperclass students are more similar than are those of freshman students. In particular, as students spend time on campus and experience the same university culture, the things they are willing to expend effort on may become more similar. Still, students may differ in the *degree* to which they expect to do well or value certain things and, subsequently, the *degree* to which they are motivated to perform. Regardless, this is one possible explanation for why we did not observe types of test-takers with different patterns of motivation but rather found evidence of a single pattern of motivation with individuals varying in their levels of motivation.

A second difference between these studies that may have led to the discrepant findings is that in the current study, we examined a different configuration of tests than was examined in the Brown et al. (2009) study. Specifically, in the current study, the difficult cognitive test of quantitative and scientific reasoning was administered first in the testing session, whereas in the previous study this test was administered in the middle of the testing session. Although the placement of this cognitive test would likely have impacted the pattern of motivation, it is unclear whether this would impact whether or not *types* of examinees were observed. Thus, we feel this is a less tenable explanation for the differing results regarding types of test-takers.

*How does test-taking motivation change over the course of the testing session?*

Given that there did not appear to be types of test-takers for this sample, the focus of the current study then shifted to examining change in motivation across the course of the testing session. The LGM results indicated that the pattern of effort scores for the aggregate group was best explained by a quadratic growth form. Specifically, effort increased in magnitude from the first to second test. However, this increase became less positive across the remaining four tests, eventually becoming negative from the fourth to the fifth test. Similar to the findings of the Brown et al. (2009) study, effort was the lowest for the difficult cognitive test. Taken together, this may suggest that in low-stakes contexts, examinees are willing to put forth effort on relatively easy tests, but are less willing to put forth effort on more difficult tests. It is also interesting to note that in this study effort was the highest for the fourth test administered in the testing session and effort scores for the remaining three noncognitive tests (i.e., tests 2, 3, and 5) were approximately equal to one another. It is not entirely clear why motivation would be higher for this noncognitive test relative to the others. The fourth test contained measures of personality, well-being, worry, and sense of identity, so one possibility is that examinees found these

measures to be more personally interesting or relevant and therefore put forth more effort on this test. Despite this change in effort scores across the five tests, it is still important to note that overall this change over time was relatively small. The largest difference in effort was between 3.61 (i.e., for the cognitive test administered first) and 3.92 (i.e., for the noncognitive test administered fourth). However, effort is reported on a 1-5 scale. Thus, although the change in effort over time was statistically significant, one may question whether the differences are practically meaningful. That is, is a 0.3 point difference in reported effort a cause for concern? Does a difference this small impact the extent to which assessment professionals can trust the scores?

In examining the change in overall effort means, it is especially interesting to compare the results of the current study to those of the Brown et al. (2009) study. First, similar to the current findings, for the Brown et al. (2009) study, the largest difference in effort found was that between the cognitive test and the noncognitive test containing measures of personality, well-being, worry, and sense of identity; these were the third and fourth tests administered in the Brown et al. (2009) study. However, for the Brown et al. study, the difference in effort for these tests was approximately 0.5 points (both for the overall sample and for the two classes that had decreases in motivation) compared to approximately 0.3 points for the current study. Thus, the difference observed in the current study was smaller than that observed previously. These findings suggest that upperclass students fluctuate less in their effort across the various tests than do freshman students.

Second, the effort given to the cognitive test by the upperclass students in the current study appears to be slightly higher than that of the freshman students in the previous study (Brown et al., 2009). This finding is interesting but not particularly surprising given that these

upperclass students have had more opportunity to learn the material on this test. According to expectancy-value theory, better knowledge of the content domain for this test should lead to higher ability beliefs, higher expectancies for success and, consequently, higher motivation for this test (Eccles et al., 1983; Eccles & Wigfield, 2002). This is what was observed here.

Third, the effort put forth by upperclass students for all four noncognitive tests was slightly lower than that given by freshman students. However, even for these noncognitive tests, the upperclass students used in the current study still reported fairly high levels of effort. Thus upperclass students may still be willing to put forth moderate to high levels of effort on these low-stakes tests and may contradict beliefs held by assessment professionals that there is a “counter-culture of assessment” among students.

#### *Expectancy-Value's Utility in Explaining Test-Taking Motivation*

Given that expectancy-value theory is often used to frame the discussion of examinee motivation, we were especially interested in examining how quantitative or math expectancy related to test-taking motivation (i.e., effort) for the cognitive test of quantitative and scientific reasoning. The results indicated that math efficacy expectations were not related to effort for this test. This was somewhat surprising because, given expectancy-value theory, we would expect individuals with higher math expectancies to put forth more effort on this quantitative test. This was not supported.

There are several possible reasons for this finding. First, the measure of expectancy we used in this study measured *sources* of math self-efficacy rather than directly measuring math self-efficacy. Second, although we would expect math self-efficacy to be related to effort on a quantitative and scientific reasoning test, it is possible that a closer match in specificity is needed to observe this relationship. For example a measure of math and science expectancy might be

needed to observe a relationship between expectancy and motivation. Finally, it is possible that the non-significant relationship between expectancy and motivation observed in the current study is indicative of a larger issue explained by expectancy-value theory. It may be that expectancy loses its utility to predict motivation when the situation is low- rather than high-stakes. This is because expectancy-value theory posits that motivation is the product of expectancies *and* values (Eccles et al., 1983; Eccles & Wigfield, 2002). The very definition of low-stakes, however, implies that value is decreased in these contexts. Specifically, if there are no consequences to the examinee, there is very little value in performing well. In this situation, even if expectancy is extremely high, the fact that there is very little value in performing well may result in low motivation. It may be that when value is low, expectancy does not predict motivation. Unfortunately, we did not include a measure of value in the current analyses and were thus unable to more fully explore this possibility.

If expectancies are not useful for explaining test-taking motivation in low-stakes contexts, one might wonder what is. In the current study we found that the personality variables of conscientiousness and agreeableness both significantly predicted test-taking for the various tests and that agreeableness significantly predicted change in motivation across the course of the testing session. Interestingly, the Brown et al. (2009) study found that the types of test-takers uncovered could be differentiated by these same personality characteristics. Taken together, this may suggest that trait-like characteristics such as these can be used in addition to expectancies and values to explain varying levels of motivation in low-stakes contexts.

#### *Implications for Low-Stakes Testing and Assessment Practice*

The results of the current study have several important implications for low-stakes testing contexts. First, the failure to find types of test-takers may suggest that for some populations (e.g.,

upperclass students) examinees have similar patterns of motivation across a testing session. Despite this, it still may not be appropriate to report and interpret test-taking motivation at an aggregate level, as there *was* significant variability in levels of motivation for each of the tests administered. Reporting the average motivation score for a test still may obscure the fact that some examinees expend a high degree of effort whereas others expend very little effort. Therefore, it may still be prudent for assessment professionals to report test scores as a function of different levels of test-taking motivation in order to best be able to interpret these test scores.

Second, across all five tests administered within this testing session, effort was the lowest for the difficult, cognitive test of scientific and quantitative reasoning. This result aligns with that of the previous study (Brown et al., 2009), and suggests examinees are less willing to give effort on difficult tests than easy tests. In other words, examinees appear to be willing to put forth effort in low-stakes contexts, but are more willing on tests that are not exceedingly difficult; at some point, if the test is too difficult, effort decreases. Unfortunately, assessment professionals are often interested in assessing higher-order skills and abilities, but the tests that measure these constructs tend to be more difficult and mentally taxing than are tests of more basic skills and attitudes. Thus, assessment professionals may have find a balance between assessing the constructs they are truly interested in and the potentially lower motivation that will result from these tests.

Third, although this study found a slight decrease between the fourth and fifth test, effort for the last test appeared to be approximately the same as for the other noncognitive measures. That is, there was not a substantial decrease in motivation for this test relative to all other tests in the session. This result aligns with previous findings (Brown et al., 2009), and suggests that, even over a three-hour testing session, fatigue does not impact test-taking motivation. Thus, our

results do not support recommendations made by some to shorten the length of the testing session in order to increase motivation.

### *Limitations of the Current Study and Directions for Future Research*

There are several limitations to the current study that we should note. First, the sample used in the current study was relatively small. Although the sample size was adequately large for the LGMs tested in the current study, it is possible that the MM results would have been different had the sample been larger. That is, it is possible that there *are* types of test-takers for this population but that we were unable to identify them due to small sample sizes. Future research should continue to explore the existence of test-taking types using adequately large samples.

Second, with regard to the mixture modeling results, the current study differed quite a bit from the Brown et al. (2009) study. We did not find evidence of types of examinees whereas previous study did (Brown et al., 2009), but there were two main differences between these studies: the configuration of the tests and the age of the examinees included in the sample. Although we feel it is more probable that the different results were a function of the sample used, it is not possible to determine what led to the discrepant findings. Thus, future research should more fully explore this either by administering the current configuration of tests to a sample of first-year students, by administering the configuration used in the previous study to upperclass students, or both.

Third, although we did include a measure of expectancy, we did not incorporate a measure of value in the current analyses. Thus, we were unable to fully examine the relationships between expectancies, values, and motivation (for the cognitive test in particular). Given that expectancies were not related to motivation for the cognitive test, it is especially important that

future studies include measures of value if we are to fully understand the utility of expectancy-value theory for explaining motivation in low-stakes contexts.

There are also several additional avenues for future research. First, given that effort has consistently been shown to be lower for the more difficult, cognitive test, researchers should explore *why* this is the case. Although this test consisted of approximately the same number of items as the four noncognitive tests, this test took nearly four times as long to complete as did the noncognitive tests. Thus, it is possible that examinees become fatigued during this test and therefore report lower motivation for it. Alternatively, it is possible motivation is lower for this test because the items are much more cognitively demanding. Second, the results of this study suggested that personality characteristics may be more useful than motivational variables (e.g., expectancies, goal orientation) in explaining motivation in low-stakes context. Thus, future research should further explore the utility of personality and other trait-like characteristics in explaining motivation in low-stakes contexts.

### *Conclusion*

In conclusion, we did not find evidence for types of test-takers characterized by differing patterns of test-taking effort across the course of a low-stakes testing session. Rather, the pattern of effort appeared to be similar for the entire sample of examinees and could best be described by a quadratic growth form wherein effort started low, increased over the second through fourth tests, and then slightly decreased between the fourth and fifth test. However, there was significantly variability in effort for all five tests. Thus, reporting aggregate motivation data may still result in biased conclusions about the skills and abilities of students. Additionally, the fact that personality characteristics were related to motivation but expectancies were not may suggest stable, trait-like characteristics should be considered in addition to expectancies and values when

understanding motivation in low-stakes contexts. Further, because effort appears to be related to these stable personality characteristics, interventions designed to increase motivation may not be effective.

Assessment and accountability continue to be driving forces within the K-12 and Higher Education arenas, subsequently leading to an abundance of low-stakes testing. Given that performance on these low-stakes tests is related to test-taking motivation (e.g., Wise & DeMars, 2005; Wolf & Smith, 1995), it is important that assessment practitioners fully understand test-taking motivation in these contexts. Such an understanding is crucial if assessment professionals are to feel confident in the validity of inferences made about student learning and program effectiveness.

## References

- AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.
- Atkinson, J. W. (1957). Motivational determinants of risk taking behavior. *Psychological Review*, *64*, 201-252.
- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*(4), 695-716.
- Bauer, D.J., & Curran, P.J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3-29.
- Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research*, *86*, 133-136.
- Bovaird, J. A. (2002). New applications in testing: Using response time to increase the construct validity of a latent trait estimate. (Doctoral dissertation, University of Kansas, 2002). *Dissertation Abstracts International*, *64*, 998.
- Brown, A. R., Barry, C. L., Horst, S. J., Finney, S. J., Kopp, J. P. (2009). *Identifying types of test takers in low-stakes conditions: A mixture modeling approach*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.
- Cacioppo, J.T., Petty, R.E., & Kao, C.F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306-307.
- Cao, J., & Stokes, S. L., (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209-230.

- DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77.
- Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B. Kaczala, C. M. Meece, J. L., & Midgely, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75-146). San Francisco, CA: W. H. Freeman.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.
- Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501-519.
- Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 269-314), Greenwich, CT: Information Age Publishing.
- Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 62*, 365-382.
- John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John (Eds.), *Handbook of Personality: Theory and Research* (2<sup>nd</sup> ed, pp. 102-138).
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.
- Meyer, J. P. (2008, March). *A mixture rasch model with item response time components*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

- Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*, 419-437.
- Muthén, B. O. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change* (pp. 291– 322). Washington, DC: American Psychological Association.
- Muthén, L.K., & Muthén, B.O. (1998-2007). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén.
- No Child Left Behind Act of 2001. Pub. L. 107-110. 8 Jan. 2002. Stat. 115.1425.
- O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment, 3*(2), 135.
- Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A Developmental Perspective on Standardized Achievement Testing. *Educational Researcher, 20*(5), 12-20.
- Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology, 32*, 8-47.
- Pieper, S. L., (2004). Refining and extending the 2 x 2 achievement goal framework: Another look at work-avoidance. (Doctoral dissertation, James Madison University, 2003). *Dissertation Abstracts International, 64*, 4436.
- Raycov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.

- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- Sclove, L.S. (1987). Application of model selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.
- Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee Motivation. *Assessment Update*, 14, 8-9.
- Sundre, D. L., & Thelk, A., & Wigtil, C. (2008). *The Natural World Test, Version 9: A measure of quantitative and scientific reasoning, Test Manual*. Harrisonburg, VA. James Madison University, Center for Assessment and Research Studies.
- Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, 29(1), 6-26.
- Thelk, A., Sundre, D.L., Horst, J. S., & Finney, S. J. (in press). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education*.
- Tofighi, D. & Enders, C.K. (2007). Identifying the correct number of classes in growth mixture modeling. In G. R. Hancock (Ed.), *Mixture Models in Latent Variable Research* (pp. 317-341). Information Age: Greenwich, CT.
- U.S. Department of Education (2006). *A Test of Leadership: Charting the Future of U.S. Higher Education* (ED Pubs No. ED002591P). Washington, D.C.
- Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, 34, 89-101.

- Wainer, H. W. (1993). Measurement Problems. *Journal of Educational Measurement*, 30(1) 1-21.
- Wigfield, A. (1994). Expectancy-value theory of achievement motivation: a developmental perspective. *Educational Psychology Review*, 6, 49-78.
- Wigfield, A., & Eccles, J. A. (1992). The development of achievement task values: a theoretical analysis. *Developmental Review*, 12, 265-310.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68-81.
- Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, 19(2), 95-114.
- Wolf & Smith, (1995). The consequence of consequence: Motivation, anxiety, and test performance.
- Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education*, 8(4), 341-351.
- Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans.

Table 1.  
*Demographic Characteristics of Sample (N = 663)*

---

Age	
Mean (SD)	20.23 (0.77)
Gender	
% Male	44.4
% Female	55.6
Ethnicity	
% White	82.4
% Black	3.2
% Asian	3.8
% Hispanic	2.8
% Other	7.8

---

Table 2.

*Correlations with Effort, Means, Standard Deviations, and Scale Reliabilities for Overall Sample*

Measure	Mean	SD	$\alpha$	1	2	3	4	5
1. Effort 1	3.61	0.72	0.81	---				
2. Effort 2	3.76	0.77	0.82	0.46	---			
3. Effort 3	3.77	0.75	0.79	0.47	0.64	---		
4. Effort 4	3.92	0.78	0.83	0.46	0.68	0.75	---	
5. Effort 5	3.79	0.81	0.82	0.45	0.61	0.71	0.80	---
6. Mastery Experience	3.85	1.29	0.92	0.08	0.04	0.07	0.06	0.05
7. Vicarious Experiences	3.69	1.14	0.95	0.14	0.11	0.13	0.12	0.11
8. Social Persuasions	3.56	1.48	0.87	0.05	-0.02	0.02	0.00	0.01
9. Physiological State	4.16	1.29	0.93	0.02	0.02	0.08	0.05	0.02
10. NW-9	48.15	7.97	0.83	0.30	0.10	0.12	0.12	0.11
11. SAT Math	577.13	69.14	---	0.08	-0.07	-0.07	-0.07	-0.08
12. Need for Cognition	3.37	0.59	0.88	0.15	0.13	0.07	0.14	0.09
13. Mastery Approach	5.38	1.12	0.84	0.29	0.27	0.32	0.34	0.31
14. Performance Approach	5.04	1.50	0.91	0.14	0.09	0.16	0.16	0.16
15. Mastery Avoidance	3.98	1.26	0.79	0.08	0.06	0.04	0.09	0.04
16. Performance Avoidance	4.47	1.34	0.69	0.05	0.04	0.06	0.06	0.06
17. Work Avoidance	3.11	1.23	0.83	-0.20	-0.21	-0.21	-0.23	-0.22
18. Extraversion	3.56	0.75	0.84	-0.01	0.09	0.08	0.10	0.12
19. Agreeableness	3.88	0.62	0.81	0.16	0.33	0.35	0.37	0.34
20. Conscientiousness	3.63	0.61	0.80	0.23	0.27	0.31	0.30	0.31
21. Neuroticism	2.77	0.74	0.82	0.10	0.05	0.04	0.07	0.04
22. Openness	3.73	0.60	0.79	0.10	0.17	0.08	0.15	0.10

Table 3.  
*Fit for Various Mixture Models (N = 683)*

	AIC	BIC	SSABIC	LMR	Entropy	LL	# of free parameters
1-Class	7,870.84	7,916.10	7,884.35	N/A	N/A	-3,925.42	10
1-Class C	5,891.51	5,982.04	5,918.54	N/A	N/A	-2,925.76	20
2-Class A	6,742.41	6,814.83	6,764.03	$p < 0.001$	0.839	-3,355.21	16
2-Class B	6,713.74	6,808.80	6,742.12	$p < 0.001$	0.823	-3,335.87	21
2-Class C <sup>1</sup>							
2-Class D	5,709.29	5,894.87	5,764.69	0.0873	0.941	-2,813.64	41
3-Class A	6,211.61	6,311.19	6,241.34	0.0011	0.875	-3,083.81	22
3-Class B	5,914.65	6,059.50	5,957.89	0.1026	0.898	-2,925.33	32
3-Class C <sup>2</sup>							
3-Class D <sup>3</sup>							

<sup>1</sup> The best loglikelihood did not replicate when using 10,000 start values.

<sup>2</sup> This model failed to converge to an admissible solution.

<sup>3</sup> The best loglikelihood did not replicate.

Table 4.  
*Fit Indices for Unconditional and Conditional Latent Growth Models*

Model	$\chi^2$	df	CFI <sub>S-B</sub>	RMSEA <sub>S-B</sub>	SRMR
<i>Unconditional Models</i>					
Linear	98.62	10	0.93	0.11	0.08
Quadratic Model	38.46	6	0.98	0.09	0.04
<i>Conditional Models</i>					
Conditional Quadratic Model	43.94	12	0.98	0.06	0.03

Table 5.  
*Parameter estimates for unconditional quadratic LGM*

Factor means, variances, and correlations					
Factor	Factor Mean	Factor Variance	Factor Correlations		
			1.	2.	3.
1. Intercept	3.591**	0.220**	---		
2. Linear	0.190**	0.062*	0.291	---	
3. Quadratic	-0.034**	0.003*	-0.218	-0.878**	---

Observed and model-implied means, and time-specific error variances				
	Observed Means	Model-Implied Means	Error Variances	Standardized Error Variances
Effort 1	3.606	3.591	0.306**	0.582**
Effort 2	3.756	3.747	0.244**	0.434**
Effort 3	3.767	3.835	0.165**	0.282**
Effort 4	3.916	3.855	0.119**	0.197**
Effort 5	3.787	3.807	0.106**	0.162**

\* Indicates significance at the  $p = 0.05$  level

\*\* Indicates significance at the  $p = 0.01$  level

Table 6.  
*Estimated Means and Variances for Effort for Each Test Administration*

---

Factor	Factor Mean	Factor Variance
Effort 1	3.591**	0.220**
Effort 2	3.747**	0.318**
Effort 3	3.835**	0.420**
Effort 4	3.855**	0.486**
Effort 5	3.807**	0.547**

---

\*\* Indicates significance at the  $p = 0.01$  level

Table 7.  
*Parameter estimates for Conditional Quadratic Model*

Factor	Factor means, variances, and correlations					Relationships with Predictors		
	Factor Mean	Factor Variance	Factor Correlations			Makeup	Conscien	Agreeableness
			1.	2.	3.			
1. Intercept	3.640**	0.182**	---			-0.156**	0.211**	0.116
2. Linear	0.199**	0.051*	0.192	---		-0.031	-0.007	0.172**
3. Quadratic	-0.035**	0.003	-0.142	-0.862**	---	0.004	0.005	-0.030**

Observed and model-implied means, and time-specific error variances				
	Observed Means	Model-Implied Means	Error Variances	Standardized Error Variances
Effort 1	3.606	3.591	0.305**	0.580**
Effort 2	3.756	3.747	0.242**	0.432**
Effort 3	3.767	3.835	0.165**	0.283**
Effort 4	3.916	3.854	0.120**	0.198**
Effort 5	3.787	3.806	0.103**	0.158**

*Note.* The factor means in the conditional model represent the initial effort score and rates of change for individuals of average agreeableness and conscientiousness, who attended the scheduled assessment day. For this reason, the factor means in this table differ slightly from those reported for the unconditional model

\*Indicates significance at the  $p = 0.05$  level

\*\* Indicates significance at the  $p = 0.01$  level

Table 8.  
*Prediction of Effort for each Test Administration*

Test Administration	Relationships with Predictors		
	Makeup	Conscien	Agreeableness
Effort 1	-0.156**	0.211**	0.116
Effort 2	-0.182**	0.208**	0.257**
Effort 3	-0.200**	0.215**	0.388**
Effort 4	-0.209**	0.231**	0.258**
Effort 5	-0.209**	0.256**	0.317**

\*\* Indicates significance at the  $p = 0.01$  level

Figure 1. Results from Brown et al. (2009) study

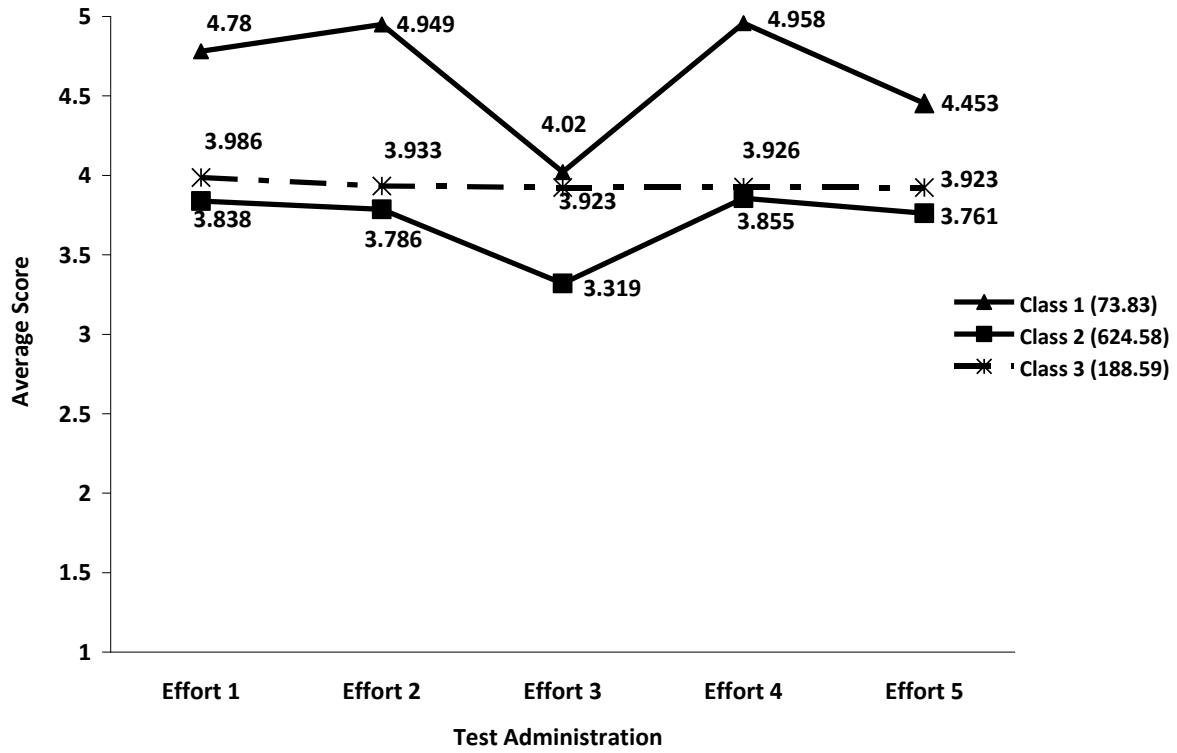


Figure 2. Observed and model-implied effort means for each test administration

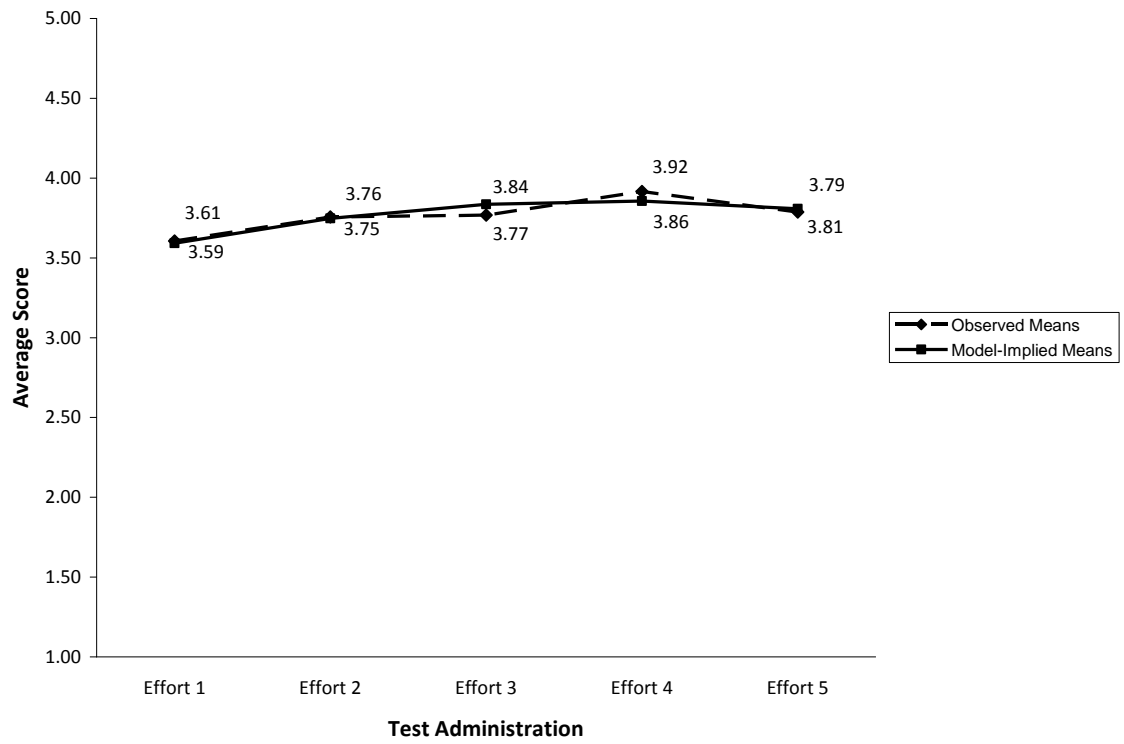


Figure 3. Unconditional linear latent growth model

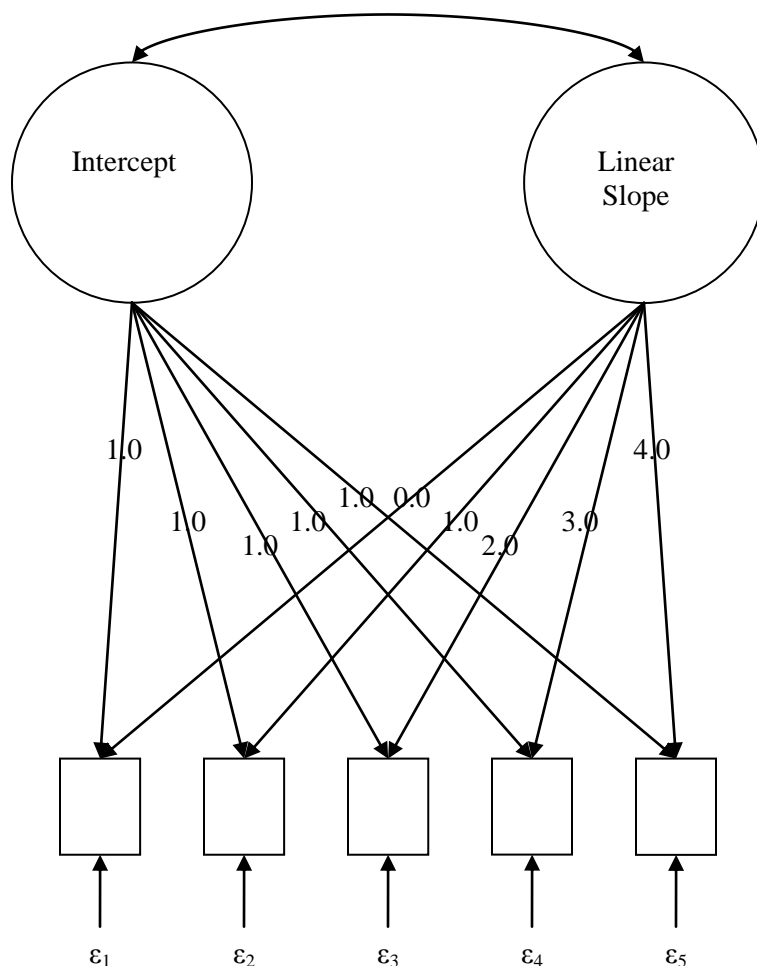


Figure 4. Unconditional quadratic latent growth model

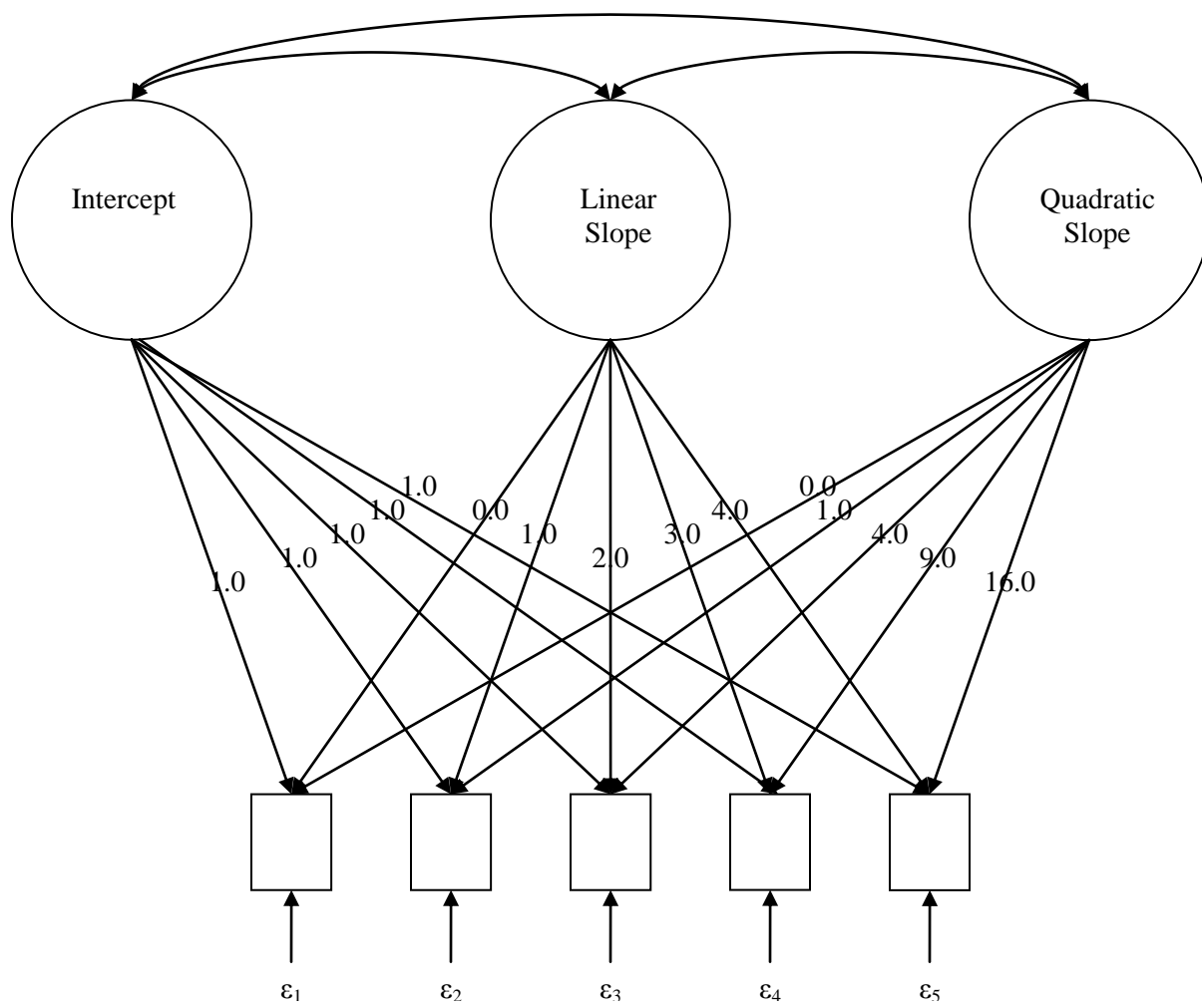
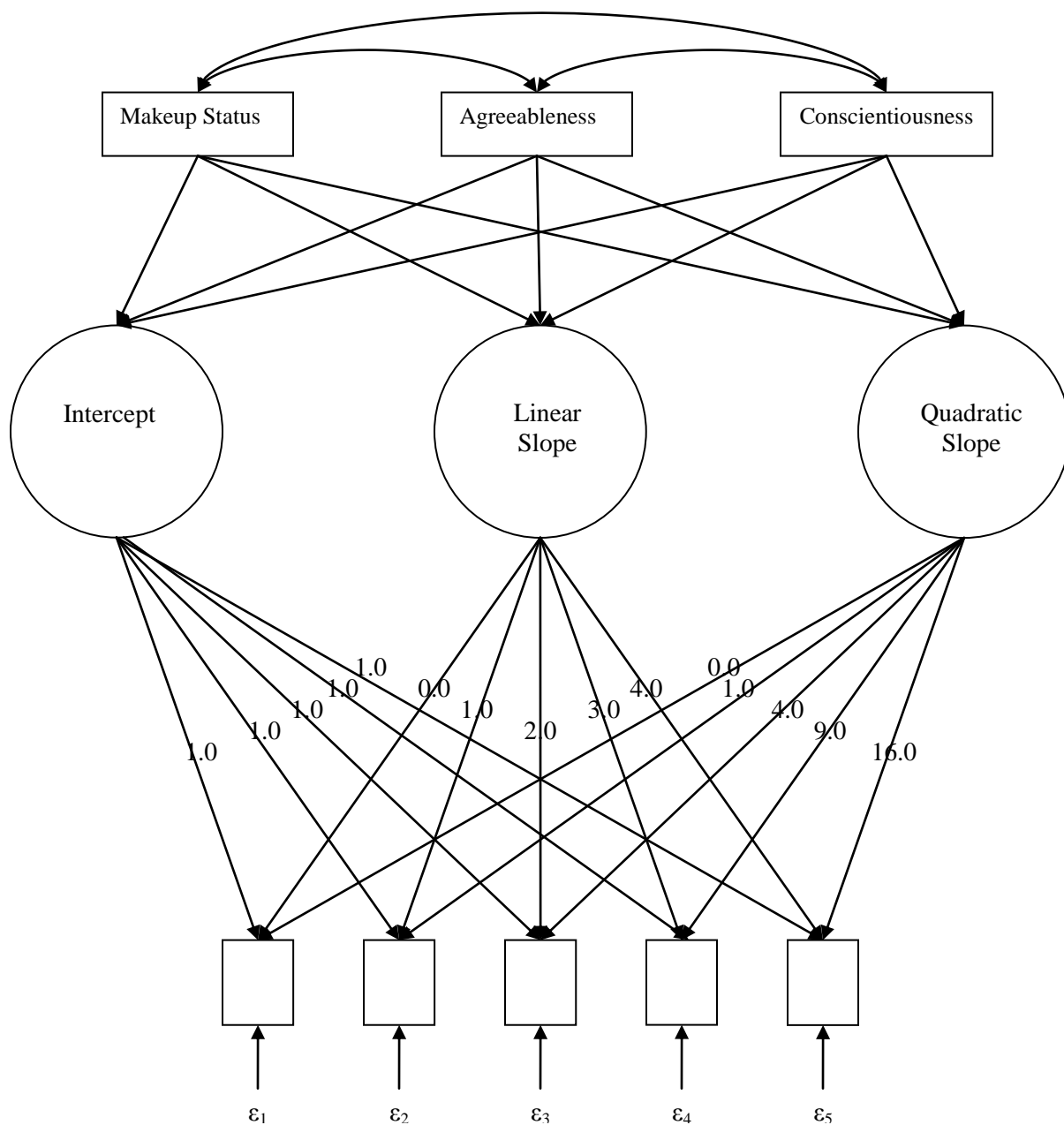
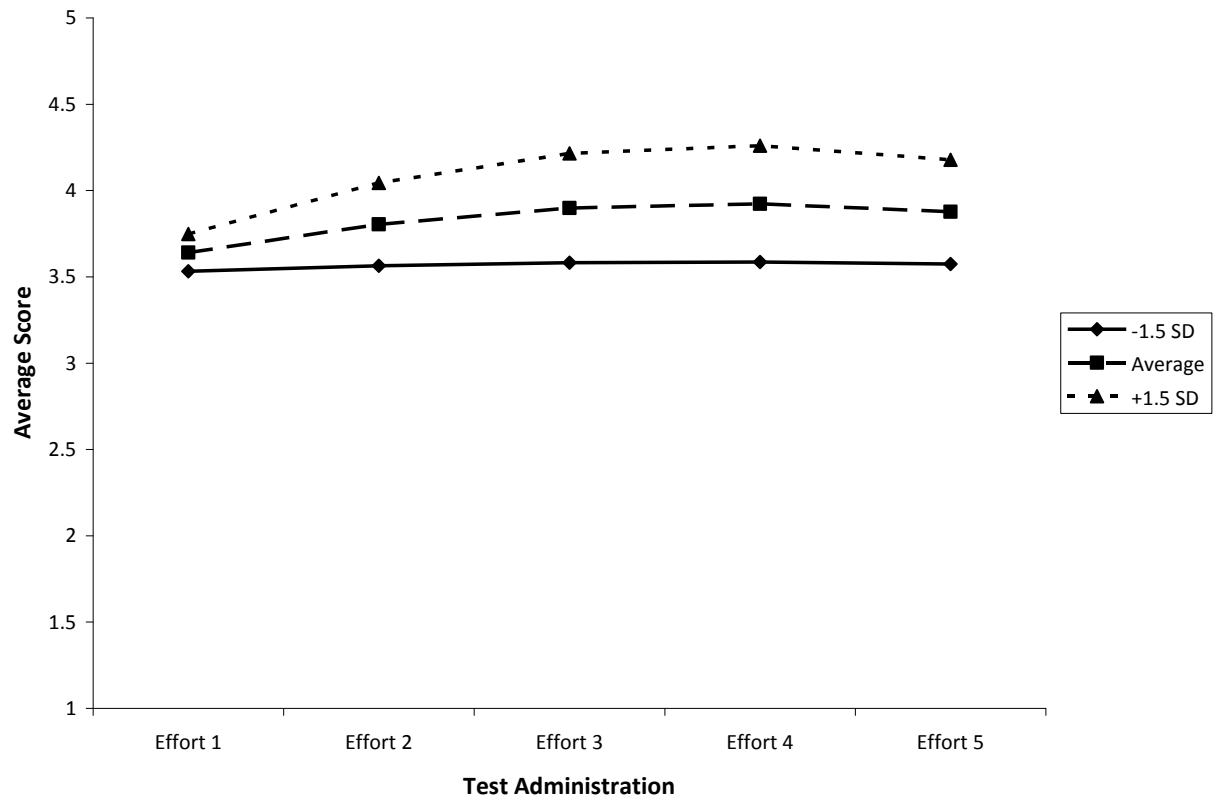


Figure 5. Unconditional quadratic latent growth model



*Note.* The correlations between the three growth factors have been removed from this figure, but were estimated in the model.

Figure 6. Plot of effort scores as a function of agreeableness.



## Appendix A

**Student Opinion Scale**

Please think about the test that you **just completed**.

Mark the answer that best represents how you feel about each of the statements below. You ended your last test at a particular number. Please enter your response to the first item below (Item A) in the next available number on the Scantron that you used for the test you just completed.

Remember, when responding, think about the test you just completed.

1 = Strongly Disagree

2 = Disagree

3 = Neutral

4 = Agree

5 = Strongly Agree

- A. Doing well on this test was important to me.
- B. I engaged in good effort throughout this test.
- C. I am not curious about how I did on this test relative to others.
- D. I am not concerned about the score I receive on this test.
- E. This was an important test to me.
- F. I gave my best effort on this test.
- G. While taking this test, I could have worked harder on it.
- H. I would like to know how well I did on this test.
- I. I did not give this test my full attention while completing it.
- J. While taking this test, I was able to persist to completion of the task.

Appendix B  
List of Noncognitive Measures

Test 2 - Noncognitive

Need for Cognition Scale (NCS-2) - 59 items. Scale consists of two separate subtests:

Subtest	Subscale Name	Sample Item	Scale Range
(1) Need for Cognition	<i>No subscales</i>	"I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought."	1-5 ("Extremely uncharacteristic of me" to "Extremely characteristic of me")
(2) Civic Responsibility Behavior Questionnaire	(a) Civic Behaviors	"I follow news or current events regarding issues that affect my personal rights."	1-5 ("Never or Rarely" to "Frequently")
	(b) Political Behaviors	"I vote in National political elections."	1-5 ("Never or Rarely" to "Frequently")
	(c) Social Behaviors	"How often do you participate in community events outside the university?"	1-5 ("Never or Rarely" to "Frequently")
	(d) Civic Efficacy	"I feel confident that my role in the political system is meaningful."	1-5 ("Strongly Disagree" to "Strongly Agree")
	(e) Civic Motivation	"I feel compelled to exercise my civil liberties (i.e. freedom of speech)."	1-5 ("Strongly Disagree" to "Strongly Agree")
	(f) Values	"Individual citizens have the ability to make an impact in political races by acting on their personal values."	1-5 ("Strongly Disagree" to "Strongly Agree")

Test 3 - Noncognitive

Attitudes Toward Learning (ATL-9) - 66 items. Consists of four separate subtests.

Subtest	Subscale Name	Sample Item	Scale Range
(1) ATL - Achievement Goals	(a) Mastery approach	"I want to learn as much as possible this semester."	1-7 ("Not at all true of me" to "Very true of me")
	(b) Mastery avoidance	"I'm afraid that I may not understand the content of my classes as thoroughly as I'd like."	1-7 ("Not at all true of me" to "Very true of me")
	(c) Performance approach	"My goal this semester is to get better grades than most of the other students."	1-7 ("Not at all true of me" to "Very true of me")
	(d) Performance avoidance	"I just want to avoid doing poorly compared to other students this semester."	1-7 ("Not at all true of me" to "Very true of me")
	(e) Work avoidance	"I want to do as little work as possible this semester."	1-7 ("Not at all true of me" to "Very true of me")
(2) Dweck's Theories of Intelligence	<i>No subscales</i>	"Your intelligence is something about you that you can't change very much."	1-6 ("Strongly Disagree" to "Strongly Agree")
(3) Perceived Cohesion	(a) Sense of belonging	"I see myself as part of the JMU community."	1-9 ("Strongly Disagree" to "Strongly Agree")
	(b) Feelings of morale	"I am happy to be at JMU."	1-9 ("Strongly Disagree" to "Strongly Agree")
(4) University Mattering Scale	(a) Awareness	"The people of the JMU community pay attention to me."	1-6 ("Strongly Disagree" to "Strongly Agree")
	(b) Importance	"I have noticed that people at JMU will take the time to help me."	1-6 ("Strongly Disagree" to "Strongly Agree")

(c) Reliance	"When people at JMU need help, they come to me."	1-6 ("Strongly Disagree" to "Strongly Agree")
(d) Ego Extension	"My successes are a source of pride to the people of the JMU community."	1-6 ("Strongly Disagree" to "Strongly Agree")

---

Test 4 - Noncognitive

General Attitudes Packet (GAP-3) - 66 items. Consists of four separate subtests.

Subtest	Subscale Name	Sample Item	Scale Range
(1) Big Five Inventory	(a) Extraversion	"I see myself as someone who is talkative."	1-5 ("Disagree Strongly" to "Agree Strongly")
	(b) Agreeableness	"I see myself as someone who is helpful and unselfish with others."	1-5 ("Disagree Strongly" to "Agree Strongly")
	(c) Conscientiousness	"I see myself as someone who does a thorough job."	1-5 ("Disagree Strongly" to "Agree Strongly")
	(d) Neuroticism	"I see myself as someone who worries a lot."	1-5 ("Disagree Strongly" to "Agree Strongly")
	(e) Openness	"I see myself as someone who is ingenious, a deep thinker ."	1-5 ("Disagree Strongly" to "Agree Strongly")
(2) Scale of Psychological Well Being	(a) Self Acceptance	"When I look at the story of my life, I am pleased with how things have turned out."	A-F ("Strongly Disagree" to "Strongly Agree")
(3) Student Worry Questionnaire	(a) Worrisome Thinking	"I worry a lot about many daily life events and situations."	1-5 ("Almost Never" to "Almost Always")
(4) Sense of Identity	<i>No subscales</i>	"I have a definite sense of purpose in life."	1-5 ("Strongly Disagree" to "Strongly Agree")

---

Test 5 - Noncognitive

Conformity/Authority (CA-1) - 64 items. Consists of three separate subtests.

---

(1) Hong Reactance Scale	(a) Reactance to Compliance	"Regulations trigger a sense of resistance in me. "	1-5 ("Strongly Disagree" to "Strongly Agree")
	(b) Emotional Response to Restricted Choice	"The thought of being dependent on others aggravates me."	1-5 ("Strongly Disagree" to "Strongly Agree")
	(c) Resisting Influence from Others	"I am content only when I am acting of my own free will."	1-5 ("Strongly Disagree" to "Strongly Agree")
	(d) Reactance to Advice and Recommendations	"I consider advice from others to be an intrusion."	1-5 ("Strongly Disagree" to "Strongly Agree")
(2) Help-Seeking (future)	(a) Instrumental Help-Seeking	"The purpose of seeking help would be just to get me started so that I could figure out the rest on my own."	1-7 ("Not at all true of me" to "Completely true of me")
	(b) Executive Help-Seeking	"Getting help in my courses would be a way of avoiding doing some of the work."	1-7 ("Not at all true of me" to "Completely true of me")
	(c) Help-Seeking Threat	"I would feel like a failure if I need help in my courses."	1-7 ("Not at all true of me" to "Completely true of me")
	(d) Help-Seeking Avoidance	"I would rather I do worse on an assignment I couldn't finish, than ask for help."	1-7 ("Not at all true of me" to "Completely true of me")
	(e) Formal v. Informal Sources of Help	"If I seek help in my courses, I would ask my professors rather than another student."	1-7 ("Not at all true of me" to "Completely true of me")
(3) Conformity Scale	(a) Peer Pressure	"My friends could push me into doing just about anything."	1-7 ("Not at all true of me" to "Very true of me")
	(b) Popularity	"I've gone to parties, just to be part of the crowd."	1-7 ("Not at all true of me" to "Very true of me")
	(c) Conformity	"I usually do what I am told."	1-7 ("Not at all true of me" to "Very true of me")

---