

Running head: MODELING MOTIVATION

**Modeling Motivation Over the Course of a Testing Period: A Mixture Modeling Approach**

Allison R. Brown, Carol L. Barry, S. Jeanne Horst, Sara J. Finney, and Jason P. Kopp

*James Madison University*

Key Words: test-taking motivation, low-stakes testing, accountability, mixture modeling, expectancy-value theory

### **Abstract**

This study was conducted to explore the existence of “types” of test-takers in a low-stakes testing context. Mixture modeling results supported three distinct classes or types of test-takers characterized by different patterns of test-taking motivation over the course of five tests. Classes 1 and 2 had varying levels of motivation whereas Class 3 had a steady level of motivation across the five tests. Additionally, examination of the relationships between class membership and external variables suggested that the classes could be differentiated by achievement goals, personality, and ability. Expectancy-value theory is used to aid in the interpretation of the results. Implications of these results and directions for future research are discussed with an emphasis on how these results impact assessment practice as the push for educational accountability continues to grow.

### **Modeling Motivation Over the Course of a Testing Period: A Mixture Modeling Approach**

Imagine a group of students who spend approximately three hours taking a series of five tests for institutional accountability purposes (e.g., No Child Left Behind Act, 2002; U. S. Department of Education, 2006). These students complete two relatively easy non-cognitive tests, then complete a very demanding cognitive test that measures math and science ability, and finish by completing two additional non-cognitive tests. In this situation, *cognitive* tests refer to those that measure knowledge, skills, or abilities and items are scored as right or wrong. *Non-cognitive* tests refer to those that measure attitudes or affect (e.g., agreeableness, achievement goal orientation) and items have no correct answer (Kyllonen, 2005). Although the students are required to complete the tests, their performance on these tests has no personal consequences for them (i.e., tests are low-stakes). Thus, not *all* of the students are necessarily expending a great deal of effort across *all* tests given the lack of consequences to the individual. That is, the amount of effort each examinee puts forth on the tests may differ from person to person and from test to test. For example, consider the following three students. First, we have John, who gives a lot of effort on the first two non-cognitive measures, finds the cognitive test too demanding and puts forth less effort, and then returns to a high level of effort on the final two non-cognitive tests. Annie, on the other hand, does not fluctuate much in her effort, giving a moderate level of effort on all five tests. Finally, Carla puts forth a high level of effort on the first test, but steadily declines in effort on each of the subsequent tests. These three students represent three different hypothetical types of test takers, each characterized by different patterns of test-taking motivation throughout the testing session (see Figure 1 for a graphical depiction of these patterns of test-taking motivation).

The current study seeks to explore the existence of distinct test-taking types during low-stakes testing endeavors. Before discussing the specifics of our study, we first provide a broad overview of the importance of reporting test-taking motivation and understanding its impact on the validity of inferences we make from test scores. We then discuss the importance of exploring test-taking types and how important differences in motivation can be masked when presenting only aggregate results when distinct subpopulations of examinees exist.

### **The Importance of Collecting, Reporting, and Studying Test-Taking Motivation**

Test-taking motivation can be defined as the extent to which an examinee gives his or her “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). Test-taking motivation is likely to be high when examinees complete tests for which there are direct personal consequences associated with their scores (e.g., achievement tests, admissions tests, placement tests). These situations are termed *high-stakes*. Alternatively, test-taking motivation is likely to be more variable when examinees complete tests for which there are little to no personal consequences associated with their scores. Situations in which this is the case are termed *low-stakes*. Understanding students’ test-taking motivation is extremely important, especially in low-stakes contexts where it is likely to vary, because it affects whether the test scores truly reflect the knowledge, skills, and abilities of the test takers. That is, the validity of the inferences made from test scores rely in part on examinees putting forth the effort needed to demonstrate their competence (Ewell, 1991; O’Neil, Sugrue, & Baker, 1995; Sundre & Wise, 2003; Wainer, 1993; Wise & DeMars, 2005; Wise et al., 2006). Given this, the most recent version of the Standards For Educational and Psychological Testing (AERA, APA, & NCME, 2004) recommend that test-taking motivation be collected, reported, and used to aid in the interpretation of test scores.

The study of test-taking motivation is imperative given the pervasive nature of low-stakes testing. With the enactment of the No Child Left Behind Act (2002) and a similar initiative in higher education (U. S. Department of Education, 2006), there is an increasing emphasis on accountability in education. Specifically, tests are administered to students and scores are used to make decisions about program or school effectiveness (i.e., the “value-added” regarding student learning and development). Additionally, resource and budget allocation is often tied to this accountability data, making these assessments high-stakes to the schools, programs, and others who use the information. Despite this, many of these tests are of low- or no-stakes for the students completing them. To the extent that students do not put forth their best effort, the scores on these tests may underestimate actual levels of proficiency, which could have dire implications for educational institutions (Wise & DeMars, 2005). That is, scores obtained on low-stakes tests may not represent what students know, but rather “what students will demonstrate with minimal effort” (O’Neil, Sugrue, & Baker, 1995/1996, p.135). Thus, understanding test-taking motivation is essential to determining the extent to which policy makers, educators, and parents can make valid inferences and sound decisions from the results of low-stakes tests.

### **The Importance of Exploring the Existence of Test-taking Types**

An additional need for studying test-taking motivation involves the fact that assessment professionals and other stakeholders tend to make assumptions about the test-taking motivation of examinees. For example, one may assume that examinees give a high degree of effort regardless of the stakes associated with the test and that test scores are valid representations of ability. In this case, one might mistakenly attribute low test scores to program ineffectiveness (i.e., no value-added) when in fact the low scores are due to low motivation. Alternatively, one

may assume that students fail to provide effort in low-stakes contexts and use this assumption as an excuse to ignore low test scores that indicate program ineffectiveness or to advocate ending low-stakes testing programs altogether.

Importantly, when making these kinds of assumptions, testing and assessment professionals often make broad statements about *all* students (e.g., *all* students try, *all* students don't try). However, for a series of low-stakes tests, it seems more likely there would be different levels of test-taking motivation not only across examinees but also across different tests (i.e., the sample of examinees is a mixture of several "Johns", "Annies" and "Carlas"; Figure 1). That is, for some tests some examinees exhibit test-taking effort and test scores are valid representations of their ability, whereas for other examinees the scores simply represent guessing or thoughtless responding. In fact, Wise and DeMars (2005) described different types of test-takers that demonstrate varying levels of test-taking motivation, both across items within a given test (e.g., may try on easy items but not on difficult items) and across different tests within a testing session (e.g., may try on all tests).

Understanding whether types of test-takers exist becomes important when one attempts to use motivation information to aid in the interpretation of test scores, as suggested by the Standards (AERA, APA, & NCME, 2004). Specifically, simply reporting an average item-level or test-level motivation score assumes that the average motivation score is representative of *all* the examinees in the sample. Alternatively, if motivation differs across examinees and varies across tests or test items, simply using the average may mask differences in motivation that could lead to different test score interpretations and testing practices. For example, examine the dashed line in Figure 1, which represents the average test taking motivation for each of the five tests. Notice that this line is most similar in pattern and magnitude to our hypothetical test-taker

Annie—those examinees who have a moderate effort across all tests. Although Annie can be adequately represented by this average motivation trajectory, John and Carla’s patterns would go unnoticed. In fact, if the average motivation scores were interpreted, practitioners would mistakenly infer that the “Carlas” put forth effort on the last two tests, when clearly they did not. That is, in reality we would expect more valid test scores from the “Johns” than the “Carlas” on the last two tests, with an easily made argument that the Carlas’ test scores should not even be reported. In addition to complicating the interpretation of test scores, reporting the average test-taking motivation may hinder innovations to testing practice. That is, if the three test-taking types represented in Figure 1 emerged, the number of examinees representing each type would be of interest. For example, if it was consistently found in low-stakes conditions that the majority of examinees were “Carlas”, testing practitioners would have serious conversations regarding the length of testing sessions. On the other hand, if the majority of examinees were “Johns”, discussion concerning the appropriate amount of cognitive demand, not testing time, would be of primary importance. Moreover, profiling the types of test-takers regarding ability, personality, and affect may lead to motivation-enhancing strategies that are targeted to different types of examinees. For these reasons, it is important to investigate whether there are types of examinees that can be characterized by different levels of test-taking motivation or whether aggregate motivation scores are appropriate.

Although Wise and Demars (2005) merely discussed the possibility that these types may exist and did not test for these types, other researchers have explicitly studied whether types of test-takers could be identified based on their levels of test-taking motivation (Bovaird, 2002; Cao & Stokes, 2008; Meyer, 2008; Wise, 2006). The researchers used mixture item response theory (IRT) models to examine the existence of types of examinees on the basis of item-level

motivation, which was operationalized using item response time. The results of these studies indicated the mixture IRT models fit the data well, thereby providing support for the existence of two types of examinees for each individual item: (1) those that are motivated, and (2) those that are not. The identification of types of test-takers using these models allow for more accurate item parameter estimates and examinee ability estimates (e.g., Cao & Stokes, 2008), which subsequently results in a more accurate understanding of what examinees know and can do.

Although the existing studies examining test-taking types evaluate item-level motivation, it seems reasonable that types of test-takers might also exist across a set of tests, such as those illustrated in Figure 1. Rather than assuming test-taking motivation is constant across all examinees and all tests, we believe that there may be different types of test-takers characterized by different profiles of test-taking motivation. Our university has administered low-stakes assessments to students twice a year for over 20 years. As proctors during these assessments, we have observed a wide variety of behaviors indicative of test-taking motivation. It is not uncommon to scan the room during any given test and observe examinees that are fully attending to the test, examinees that are skimming the questions and answering as they see fit, and examinees that are sleeping. Furthermore, the frequency of these behaviors appears to change throughout the course of the testing session.

Fortunately, methods exist that allow for the identification of types of test-takers on the basis of test-level motivation. These methods, such as cluster analysis and mixture modeling, are referred to as person-centered (PC) approaches because the focus is on identifying types of people that share similar values on a set of variables. That is, in PC approaches, the focus is on the pattern or profile of scores on a set of variables rather than on examining the relationships among those variables, which is the focus of variable-centered (VC) approaches (e.g.,

correlation, regression, factor-analysis). Thus, PC approaches allow researchers to view individuals in a holistic sense and emphasize *types* rather than *dimensions* (Magnusson, 1998; Robins, John, & Caspi, 1998). If investigations utilizing PC approaches yield results that support the existence of distinct types of examinees, it would be beneficial to illustrate the patterns of motivation demonstrated by each type as well as what differentiates them from one another.

### **Using Expectancy-Value Theory to Understand the Existence of Test-Taking Types**

Motivation theories provide theoretical rationale for differences in test-taking effort. One especially pertinent theory that has been applied to test-taking motivation (Wise & DeMars, 2005; Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995) is expectancy-value theory. In general, expectancy-value theory posits that the amount of effort individuals put forth on a particular task is a product of their expectancies regarding how well they will do on the task, as well as how much they value the activity (Atkinson, 1957; Eccles et al., 1983; Eccles & Wigfield, 2002; Wigfield, 1994; Wigfield & Eccles, 1992, 2000). Although all expectancy-value theorists agree on these two determinants of motivation, there are subtle differences in the exact specification of expectancies and values. Eccles and colleagues (Eccles et al., 1983; Eccles & Wigfield, 2002; Wigfield & Eccles, 1992, 2000) posit that expectancies consist of two aspects: (1) *expectancies for success*, which refer to one's assessment of the likelihood of performing well on the activity; and (2) *ability beliefs*, which refer to one's perception of his or her current level of aptitude for the content domain covered by the activity. Values consist of four aspects: (1) *attainment value*, which refers to how important performing well on the task is to the individual; (2) *intrinsic value*, which refers to the satisfaction or pleasure one receives from the activity; (3) *utility value*, which is concerned with the usefulness of the activity for future plans; and (4) *cost*, which refers to the drawbacks of participating in the activity and includes both

emotional states (e.g., anxiety, fear or success of failure) and the amount of effort needed to succeed at the task (i.e., the cognitive demand of the task).

Several alternative conceptualizations of the expectancy-value framework provide different treatments of cost, perhaps as a result of the lack of attention to cost in the literature (Wigfield & Eccles, 2000). One alternative conceptualization has been to treat the emotional states aspect of cost as a separate entity (Pintrich, 1988; Pintrich, 2004; Pintrich & De Groot, 1999). A second alternative conceptualization has been to integrate the cognitive demand of the task into the expectancy component of the model (Wolf & Smith, 1995; Wolf et al., 1995). Those endorsing this second alternative conceptualization argue that individuals take into account the effort required to successfully complete the task when calculating their expectancies for success. A third alternative conceptualization has been to examine both emotional states and cognitive demand as separate components of expectancy-value theory, wherein motivation is determined by expectancy for success, task value, cognitive demand, and emotional reactions to the test (Wise & DeMars, 2005). Regardless of the specific conceptualization, expectancy-value theory provides a useful framework for understanding test-taking motivation and could prove useful for differentiating between types of test takers. Moreover, low-stakes testing situations provide a unique context in which to study the cost aspect of expectancy-value theory, as cognitive demand is likely to have a large impact on effort put forth on tests that have no personal consequence. Given expectancy-value theory, our observations during low-stakes testing endeavors, and test-taking types uncovered at the item-level, we hypothesized that specific types of test-takers would emerge across a large-scale low-stakes testing session.

### **Expected Test-Taking Types**

When assessing test-taking motivation over the course of the testing session described in the opening scenario of this paper (see Figure 1), we would expect different profiles of test-taking motivation which would represent different types or subpopulations of test-takers. First, we expect to observe a pattern of test-taking motivation similar to that of John, where motivation is high on the first two non-cognitive tests, decreases for the cognitive test (a difficult quantitative and scientific reasoning test), and returns to a high level for the final two non-cognitive tests. This drop in motivation for the cognitive test could be explained by considering the cost associated with cognitively demanding tasks. That is, in a low-stakes setting some students may have higher motivation for tests that are not especially difficult but may have much lower motivation for tests that are especially cognitively demanding. This has been empirically-supported at the item-level (Bovaird, 2002; Wise, 2006; Wolf et al., 1995). However, this phenomenon has not been studied at the test-level.

A second pattern of test-taking motivation we might observe is similar to that exhibited by Annie, wherein motivation does not fluctuate throughout the testing session but rather stays at a moderate level. Expectancy-value theory also provides a nice explanation for this type of test-taker. We would expect these test takers to be competent at math and science, to be interested in the subject, and to expect to do well on this cognitive test. Thus, despite the low-stakes nature of the context, we would expect motivation to be higher on the cognitive test for these examinees than for those like John due to 1) higher expectancies for success, 2) the fact that this test is less cognitively demanding for these examinees, or, most likely, 3) some combination of the two. This hypothesis would be empirically supported by finding that the “Annies” have higher math ability (e.g., higher Math SAT) than the “Johns”.

Finally, we might observe a pattern of test-taking motivation similar to Carla's, wherein motivation begins high but steadily declines throughout the course of the testing session. For these test-takers, levels of test-taking motivation may not be impacted by expectancies or the cognitive demand of the tests, but rather the extent to which the examinees become fatigued throughout the testing session. In general, the longer examinees are required to perform, the larger the decrease in motivation (Mislevy, 1995). Research has demonstrated that, within a given test, examinees may become fatigued and thus, not respond to items near the end of the test with appropriate levels of motivation (Cao & Stokes, 2008; Wise, 2006). It is likely that this pattern found at the item-level would also be observed across a set of tests, especially if the testing session is several hours in length.

### **Purpose of the Current Research**

As the emphasis on accountability increases, so does the use of low-stakes data to make high-stakes decisions about program effectiveness. If we are to make valid inferences about what students know and can do, it is crucial to understand the test-taking motivation of examinees in these low-stakes contexts. However, we believe that a single statement regarding the pattern of test-taking motivation across a testing session does not accurately represent the motivation exhibited by all examinees. Given this, the purpose of the current research was four-fold: (1) to investigate the existence of different types of test-takers with regard to test-taking motivation; (2) to explore whether the types that emerge could be differentiated by a set of external variables (e.g., achievement goals, personality, ability); (3) to evaluate the utility of expectancy-value theory for differentiating between test-taking types; and (4) to examine the extent to which a cognitively demanding test impacts test-taking motivation.

### **Methods**

## Participants and Procedures

Data were collected from incoming college freshmen who completed a three-hour testing session during a university-wide assessment day at a mid-sized southeastern university. All university students are required to participate in two assessment days for the purpose of institutional accountability mandates. The first occurs the week before fall classes begin (i.e., first semester), and the second occurs during the spring semester after students complete 45-70 credit hours (typically sophomore or junior status). The data for the current study were collected during the Fall 2008 assessment day (i.e., only freshman data were analyzed). On these assessment days, students complete a range of general education and developmental measures, with trained proctors administering the instruments and reading instructions aloud before students begin responding. Importantly, the students completing tests on these assessment days are told that these tests are of low-stakes to them personally. That is, the students are told that the test scores do not impact their grades or academic standing, but rather scores are used to assess the effectiveness of the university's academic and student affairs programming. Further, students receive no feedback regarding their scores on the tests. Thus, the students should realize that there are no consequences for them regarding their performance on the tests.

A subset of the incoming freshmen completed the set of five tests used in the current study ( $N = 887$ ; see Table 1). The first and second tests were non-cognitive, the third was a cognitive test designed to assess quantitative and scientific reasoning, and the fourth and fifth test were non-cognitive; all tests consisted of approximately 60 items. After completing each test, examinees were instructed to respond to a set of items assessing their test-taking effort for the test they just completed. It is these five test-specific effort scores that were modeled to identify types of test takers.

## Instruments

**Test-Taking Motivation.** Examinee test-taking motivation was measured using the Student Opinion Scale (SOS), which consists of two subscales and has been extensively studied (Sundre & Moore, 2002; Thelk, Sundre, Horst, & Finney, in press). The Effort subscale consists of five items that measure the degree to which examinees put forth effort on a given test (e.g., “I gave my best effort on this test.”). The Importance subscale consists of five items that measure the degree to which examinees view a given test as important (e.g., “Doing well on this test was important to me.”). Examinees are asked to respond to a series of statements using a scale of 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Items were averaged for each subscale to create two subscale scores, with higher values indicating higher degrees of effort and importance. Only scores from the Effort subscale were used to uncover types of test-takers.

**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 of the 66 items was dichotomously scored, and then scored responses were summed to create a total NW-9 score. The NW-9 was the cognitive test that was administered as the third test in the testing session. Unfortunately, scores on the NW-9 may be confounded by the amount of effort that the examinees put forth on this test, and thus may include construct-irrelevant variance. For this reason, SAT math scores, obtained from university records, were used as a second measure of quantitative ability.

Quantitative ability was deemed an important external variable given that the cognitively demanding test examinees completed was a test of quantitative and scientific reasoning. Recall that ability is one component of expectancy under Eccles’ conceptualization of expectancy-value

theory. That is, higher ability should be associated with higher expectancy, and subsequently, higher motivation.

**Need for Cognition.** Need for cognition can be defined as one's need to be cognitively stimulated. This was measured using the 18-item Need for Cognition Scale (NCS: Cacioppo, Petty, & Kao, 1984). Participants read a series of statements and indicated how characteristic each statement is of them using a scale of 1 (*Extremely Uncharacteristic*) to 5 (*Extremely Characteristics*). Responses to each of the 18 items were summed to create a total score, with higher values representing a higher degree of need for cognition. Need for cognition was deemed an important external variable because individuals with higher levels of this construct will likely try harder on low-stakes tests simply because they enjoy the cognitive stimulation that completing these tests offers.

**Achievement Goal Orientation.** Students' achievement goals were measured by the Attitudes Towards Learning questionnaire (ATL: Finney, Pieper, & Barron, 2004; Pieper, 2004), which consists of five subscales: mastery approach (MAP), performance approach (PAP), mastery avoidance (MAV), performance avoidance (PAV), and work avoidance (WAV). MAP is defined as the goal of developing competence (e.g., "I want to learn as much as possible this semester."). PAP is defined as demonstrating competence relative to others (e.g., "My goal this semester is to get better grades than most of the other students."). MAV is defined as avoiding misunderstanding (e.g., "I'm afraid that I may not understand the content of my classes as thoroughly as I'd like."). PAV is defined as avoiding inferiority relative to others (e.g., "I just want to avoid doing poorly compared to other students this semester."). Finally, WAV is defined as avoiding as much work as possible (e.g., "I want to do as little work as possible this semester"). Participants responded 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.

Achievement goals were collected to gather validity evidence for the classes because one's underlying goal orientations may be related to the amount of effort that one puts forth in a low-stakes context. In particular, individuals with high levels of approach orientations (i.e., MAP and PAP) will likely report moderate to high effort given that these individuals either enjoy learning for learning's sake (i.e., MAP: likely to enjoy taking tests) or want to demonstrate competence relative to others (i.e., PAP: likely to try hard on the test to prove they are best). Alternatively, individuals with high levels avoidance orientations (i.e., MAV, PAV, and WAV) will likely report low effort, given that these individuals are likely to avoid situations that will make them feel or appear incompetent (i.e., MAV or PAV), or are likely to avoid hard work in general (i.e., WAV).

**Personality.** Participants' personality characteristics were assessed 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 (i.e., outgoing), Agreeableness (i.e., compliant), Conscientiousness (i.e., detail oriented), Neuroticism (i.e., tense and worried), and Openness (i.e., curious). Participants respond to a series of statements using a scale from 1 (*Disagree Strongly*) to 5 (*Agree Strongly*). Items were summed to create five subscale scores, with higher scores indicating higher levels of each personality dimension.

The Big Five were collected to help us understand if test-taking types differed due to broad personality traits. For example, examinees that put forth high effort on low-stakes tests (e.g., John) may be higher on conscientiousness than are examinees who put forth moderate to

low effort on these types of tests (e.g., Annie). A similar pattern might emerge for agreeableness because examinees are being asked to complete tests for which there are no direct consequences to them. Regardless, although the use of personality characteristics to explore differences between types of test-takers is largely exploratory, we believe these traits could possibly offer useful insight as to the individuals who fall within each class.

### **Data Analysis**

**Mixture Modeling.** Mixture modeling was used to evaluate whether there exist different types of test-takers on the basis of the set of five effort scores. In mixture modeling, the observed data are assumed to have been sampled from a population that consists of a *mixture* of distributions. Thus, the overall population is thought to be composed of a number of subpopulations, or classes, of people. Mixture modeling is considered a latent variable technique, wherein an individual's value on the latent categorical variable (i.e., class membership) is thought to drive his or her responses to the observed variables (in our case, test effort scores). Unlike cluster analysis, another PC technique, mixture modeling allows researchers to freely estimate or constrain parameters (e.g., means, variances, and covariances) across classes. Further, mixture modeling provides information about the proportion of the sample belonging to each class (i.e., mixing proportions) and the probability of belonging to each class for each person (i.e., posterior probabilities).

In the current analyses, we estimated a series of mixture models varying in number of classes and parameterizations, which is typical given the exploratory nature of mixture modeling (Pastor, Barron, Miller, & Davis, 2007). Specifically, we tested one- through four-class solutions with four different parameterizations (Models A through D). In Model A, means were estimated within each class, variances were fixed to be equal across classes, and covariances

were fixed to be zero for all classes (forcing local independence for the five effort scores). By local independence, we mean that after controlling for class membership, the five Effort scores are uncorrelated; thus, this model assumes that the observed correlations between Effort scores are only due to the fact that there is a mixture of classes represented in the data (Bauer & Curran, 2004). Given this, Model A was the most restricted and parsimonious parameterization. In Model B, means and variances were freely estimated within each class, and covariances were again fixed to be zero for all classes (again forcing local independence). For Model C, means and variances were again freely estimated within each class, and covariances were estimated but fixed to be equal across classes. By estimating within-class covariances, the assumption of local independence is relaxed. In other words, Effort scores are allowed to correlate even after controlling for class membership; thus, class membership moderates the relationships between effort scores (Bauer & Curran, 2004). Finally, for Model D, means, variances, and covariances were all freely estimated for each class; Model D was the least restricted parameterization. In addition, for each model,  $k - 1$  mixing proportions were estimated ( $k =$  number of classes), which indicate the proportion of the total sample representing each class.

Model-data fit was examined using several fit indices. There are four commonly used relative fit indices, all of which are based upon the *LL*: 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). That is, these fit indices compare the fit of competing models to one another. The AIC, BIC, and SSA-BIC are used to compare models that differ in both the number of classes and the parameterization, with smaller values indicating better relative fit. The LMR test is used to compare models that differ in the number of classes, but that have the same

parameterization; a non-significant  $p$ -value associated with this test statistic indicates that the model with  $k$  classes does *not* fit significantly better than a model with  $k - 1$  classes. In this case, the model with fewer classes would be favored. The SSA-BIC will be given the most weight in championing a solution as it has been shown to function well in terms of identifying the correct number of classes (Tofighi & Enders, 2007).

Multiple maxima of the likelihood often exist in mixture modeling, thus the championed solution may have converged on a local maximum rather than a global maximum (Muthén & Muthén, 1998-2006). A solution is not stable and thus should not be interpreted if it converged on a local maximum. To assess the stability of a solution, the model is estimated multiple times using different start values, and the solutions are examined to make sure they converged on the same log-likelihood value and yielded equivalent parameter estimates. When the log-likelihood values replicate and parameter estimates are the same across different start values, the solution can be deemed stable. The use of multiple start values (5,000) was employed in the current study to ensure that the solutions were stable. Any solutions that were not stable were noted and not interpreted.

**Gathering Validity Evidence for Classes.** Nonnormality of observed data, misspecification of a model's parameterization (e.g., forcing local independence), and nonlinear relationships among observed variables all can cause spurious classes to emerge in mixture modeling (Bauer & Curran, 2004). Thus, gathering validity evidence for classes is a crucial step in mixture modeling analyses, and provides evidence that the classes extracted represent *meaningful subpopulations* rather than simply patterns in the data. To gather validity evidence for these classes, one might examine whether the classes can be differentiated using theoretically-related criteria. There are several methods available to do this.<sup>1</sup> The method

chosen for the current set of analyses, implemented using the AUXILIARY command in Mplus (Muthén & Muthén, 1998-2007), derives class membership based only of the five effort scores, and uses the posterior probabilities to compute class means for each of the external variables. Differences between these class means are then tested for statistical significance.

## **Results**

### **Descriptive Statistics**

Table 2 contains descriptive statistics for each of the variables and Table 3 contains correlations between effort scores at each of the five time points. As expected, correlations between effort scores at time 3 (i.e., effort for the cognitive test) were lower than correlations between the other time points (i.e., effort for the non-cognitive measures), with the exception of the correlation between the first and last measure. Figure 2 graphically depicts the pattern of average effort scores (i.e., aggregate) across the five time points. Note that the average effort score following the cognitive test (Effort 3) is lower than the average effort scores for the non-cognitive tests.

### **Mixture Modeling**

In order to examine the possibility of underlying subpopulations of test-takers, one-, two-, three-, and four-class models with four different parameterizations (models A, B, C, and D) were estimated via robust maximum likelihood estimation (MLR) using Mplus version 5.2. Table 4 contains the fit indices. Notice that the log-likelihood for four-class model D was not replicated; hence, the fit indices are not reported.

Of the models tested, the three-class model D (i.e., means, variances and covariances freely estimated within and across classes) fit the best. Results of the LMR test suggested that the three-class model D did *not* fit significantly better than the two-class model D. However, the

AIC, BIC, and SSABIC were the lowest for the three-class model D. The classification table (Table 5) and the entropy statistic (0.890) for the three-class model D solution suggested high “classification utility” (Pastor et al., 2007, p. 20), in comparison to that of the two-class model D solution (0.699). Further, when plotting the means for each class (see Figures 3 and 4), it appeared that the third class was characterized by a distinct pattern of means and may represent a qualitatively distinct type of examinee. On the other hand, the two-class solution represents two quantitatively ordered classes. That is, rather than representing distinct patterns of effort scores, these two classes represent gradients or degrees of the same pattern of effort scores (i.e., high, high, low, high, high). However, note that for the three-class solution there are distinct patterns of effort exhibited across the testing session. That is, Class 3 appears to have a qualitatively distinct pattern of test-taking effort (i.e., steady and moderate) compared to Classes 1 and 2 (i.e., high, high, low, high, high). Means, standard deviations, and correlations for the five Effort scores for each of the three classes are presented in Table 6. It is also interesting to note how much the correlations for the five Effort scores vary across the three classes. Specifically, although these correlations are low to moderate for Classes 1 and 2, the correlations are quite high for Class 3. Given the consistency in Effort scores for this Class 3, this is reasonable and indicates that individuals within this class tended to stay in the same rank order across the five tests. These differences in correlations across the classes also highlight the value in relaxing the assumption of local independence and allowing these relationships to vary across classes.

A closer examination of Figure 4 reveals that Class 1 reported higher effort than the other classes across all five time points, with the exception of practically no difference between Class 1 and Class 3 for the cognitive test (effort 3). Average effort scores for Class 1 also fluctuated more between tests than effort scores for the other two classes. At first glance, it appears that

effort scores for Class 2 were consistently the lowest of the three groups. However, closer inspection revealed that Classes 2 and 3 were quite similar in test-taking effort except for the cognitive test (effort 3), where Class 2 had lower effort than Class 3. That is, similar to Class 1, Class 2 showed a decrease in effort for the cognitive test; however, effort scores associated with the non-cognitive tests were similar to Class 3. Class 3 demonstrated an altogether different pattern of effort across the testing session. Effort scores for Class 3 were steady across the testing session, ranging from 3.986 for the first measure to 3.923 for the last measure. Importantly, there was no drop in effort for the cognitive test as was found for the other two classes.

In summary, Class 1 consisted of approximately 8% of the total sample and had a pattern of means indicating that these individuals had extremely high effort for the first two tests (i.e., non-cognitive), dropped substantially in effort for the third test (i.e., cognitive), returned to a very high effort level for the fourth test (i.e., non-cognitive), and then had a slight decrease in effort for the final test (i.e., non-cognitive). Given the patterns of means, Class 1 was named the “high dippers” class. Class 2 was the largest class, consisting of approximately 71% of the total sample, and had a pattern of means similar to but lower than Class 1. That is, individuals in Class 2 had moderately high effort for the first two tests, dropped in effort for the third test, and then returned to moderately high effort for the final two tests. Thus, Class 2 was named the “low dippers” class. Interestingly, Class 3, consisting of approximately 21% of the total sample, displayed a different pattern of test-taking motivation. Individuals in this class had a moderately-high level of effort that remained relatively constant across the five tests. For this reason, Class 3 was named the “consistently-motivated” class.

### **Gathering Validity Evidence for Classes**

To gather evidence that the classes represented meaningful subgroups, differences among the three classes on theoretically meaningful external criteria were explored (see Table 7). Because the means for the external criteria are on different scales, average scores are represented graphically as z-scores. It is important to note that the z-scores in Figure 5 are *relative* comparisons among the three groups, rather than the *absolute* magnitude of means for each of the external criteria, which can be found in Table 7. Below, each class is profiled holistically using the external variables, followed by a description of the variables that help distinguish between classes (each external variables is examined to assess how it discriminated between classes).

**Profiling each class holistically.** When examining the external variables for each class (see Figure 5), individuals in Class 1 (i.e., the “high dippers”) tended to be highest on MAP, PAP, Agreeableness, Conscientiousness, and Openness, and tended to have the lowest levels of WAV compared to the other two classes. Class 2, on the other hand (i.e., the “low dippers”), had moderate levels on all of the external variables compared to the other two classes. Finally, Class 3 (i.e., the “consistently-motivated class”) also had relatively moderate levels on all of the external variables except for the measures of quantitative ability. In particular, individuals in Class 3 were characterized by slightly elevated SAT Math scores and NW-9 scores compared to the other two classes.

**Assessing utility of each external variable.** When examining each external variable, Need for Cognition, MAV, PAV, Extraversion, and Neuroticism did not distinguish between the three classes (see Table 7). MAP, PAP, WAV, Agreeableness, Conscientiousness, and Openness distinguished Class 1 (i.e., “high dippers”) from both Classes 2 (i.e., “low dippers”) and 3 (i.e., “consistently-motivated”). Specifically, Class 1 was significantly higher on MAP, PAP,

Agreeableness, Conscientiousness, and Openness and significantly lower on WAV than *both* Classes 2 and 3 (Table 7). Classes 2 and 3 were not significantly different on these variables. Interestingly though, SAT math and NW-9 scores did distinguish between Classes 2 and 3. Specifically, Class 3 was significantly higher on SAT math and NW-9 than was Class 2, suggesting that Class 3 is characterized by higher quantitative ability than Class 2. Further, although Class 3 had significantly higher SAT math scores than did Class 1, these two classes did not differ in their NW-9 scores. The fact that these two classes had similar NW-9 scores but not SAT math scores may be due to the fact that examinees in Class 3 reported lower effort on the NW-9 test than did examinees in Class 1 and, therefore, their scores on this test may underestimate their true ability level.

In summary, given the results of the mixture modeling analyses of the five effort scores, we championed a three-class solution in which the three classes differed in their means, variances, and covariances for the five effort scores. Classes 1 and 2 (consisting of approximately 8% and 71% of the total sample, respectively) appeared to have the same pattern of effort scores but differed by a matter of degree. That is, for these classes, individuals had moderate to high effort for the first two and last two tests (i.e., non-cognitive) and dropped substantially in effort for the third test (i.e., cognitive). On the other hand, Class 3 (consisting of approximately 21% of the total sample) appeared to represent a qualitatively distinct type of test-taker, having moderately-high effort across all five tests (i.e., no drop in motivation for the cognitive test). External validity evidence for the classes indicated that Class 1 was differentiated from the other two classes by more favorable levels of most external attitudinal and personality variables, whereas Class 3 was characterized by higher levels of quantitative ability (as measured by SAT math scores) than the other two classes. Thus, both the pattern of

means and the external validity evidence for the three classes suggest that these classes represent *meaningful* types or subpopulations of test-takers.

### **Discussion**

Decisions about program effectiveness and student learning are frequently made from data collected in low-stakes testing situations. Given that low motivation in these testing situations may obscure the true levels of the construct under investigation, there is an urgent need to understand the test-taking motivation of examinees in low-stakes contexts. It is likely that individuals vary in their test-taking motivation, which may reflect the existence of examinee types characterized by different levels and patterns of test-taking motivation. Thus, the purpose of the current study was to explore the possible existence of types of test-takers, gather validity evidence for the types that were uncovered, and explore the utility of expectancy-value theory or other motivation theories to explain the differences between types.

#### **Using Expectancy-Value Theory to Understand Patterns of Test-Taking Motivation**

Expectancy-value theory provides a useful framework for interpreting the results of the current analyses. Specifically, for Classes 1 and 2, we observed a decrease in test-taking motivation (i.e., effort scores) for the cognitive test measuring quantitative and scientific reasoning skills. This test was quite difficult and cognitively demanding, in contrast to all four of the other tests, which were less demanding tests of attitudes and behaviors. Thus, the amount of mental taxation required to successfully answer these cognitively demanding math and science questions was much greater than that required to answer the items on the non-cognitive tests. In other words, there was a higher cost associated with the cognitive test, which could have led to the lower levels of test-taking motivation.

Expectancy-value theory also provides a useful framework for interpreting the differences between Class 3 and Classes 1 and 2. Specifically, Class 3 did *not* decrease in motivation for the cognitive test whereas Classes 1 and 2 *did decrease* in motivation for this test. Recall that Class 3 had higher quantitative ability (i.e., SAT math scores) than did Classes 1 and 2. Thus, what appeared to distinguish between examinees that decreased in motivation and those that did not was quantitative ability (i.e., SAT math). From the expectancy-value framework, Class 3 individuals may have exhibited more motivation for this test because they had higher ability. This could lead to higher expectancies for success, lower perceived cost, higher value, or some combination of the three. More specifically, individuals that have high ability would likely be more confident that they can succeed on a given task (i.e., higher expectancy); specifically, the higher individuals' quantitative ability, the more they expect to succeed on a test of quantitative and scientific reasoning. Alternatively, it could be that individuals with higher ability have to put forth less effort to succeed (i.e., less cost); specifically, the higher the quantitative ability, the easier a test of quantitative and scientific reasoning. A third explanation is that individuals having high ability may view the task as more important (i.e., higher value) than individuals with lower ability; specifically, individuals with higher quantitative ability are more likely to value a test of quantitative and scientific reasoning. It is most likely, however, that some combination of these three factors resulted in the higher motivation for the cognitive test for Class 3.

The external variable values and levels of effort for Class 1 are congruent with several recent research studies that considered personality and goal orientation in tandem with student motivation (Ackerman & Kanfer, 2009; Liem, Lau, & Nie, 2008; Yeo & Neal, 2008). Specifically, the high conscientiousness, high effort, and lower ability of Class 1 are congruent

with findings that students who are lower in ability and higher in conscientiousness, report higher effort than students with higher ability and lower levels of conscientiousness (Yeo & Neal, 2008). Class 1 characteristics of high MAP and high effort also align with the findings that students with high mastery goal orientation were less likely to disengage from a task (Liem et al., 2008). Openness to experience, conscientiousness, and MAP has also been related to lower levels of subjective fatigue throughout lengthy low-stakes testing sessions (Ackerman & Kanfer, 2009). Therefore, we felt that the external validity evidence, coupled with motivation theory, provided support for distinction among the test-taking classes.

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

The results of this study have several important implications for the use of low-stakes contexts to gather test scores. First, contrary to our expectation, we did not observe a group that displayed an overall and steady decline in test-taking motivation. Had we observed this pattern, it might have suggested that the length of the testing session influences fatigue which, in turn, impacts test-taking motivation. The fact that we did not observe this pattern, even over a three-hour testing session, implies that examinee fatigue may not be as big of a problem as some would suggest. Therefore, recommendations to shorten the length of a testing session in order to increase motivation are not supported by our results. This is congruent with recent research on fatigue, where performance actually increased over longer testing sessions, despite subjective reports of fatigue by some participants (Ackerman & Kanfer, 2009).

Second, the majority of examinees in our sample (78 %) reported lower motivation for the more cognitively-demanding test. Interestingly, the two classes with lower effort associated with the cognitive test were those that had lower math ability. This implies that for students with low to moderate levels of ability, the cognitive demand of a test and its associated cost can lead

to decreased motivation in a low-stakes setting. This suggests that matching examinee ability and interest with cognitive test content may result in increased effort when tests of are no consequence to the students. However, this type of matching may not be possible when the goal is to gauge the competency of a heterogeneous group of examinees with varying ability and interest levels.

Third, Wise and DeMars' (2005) review of the impact of test-taking motivation on performance indicated that motivated students consistently outperform less motivated students. One result of this is that low motivation may obscure the presence of high ability and vice versa. Interestingly, it appears that this effect also occurred in the current study. Despite the fact that students in Class 3 had the significantly higher entering SAT math scores than students in Class 1, the means for these two classes on the cognitive test were statistically the same. This is likely due to the fact that Class 3 had lower motivation for the cognitive test than did Class 1. Had Class 3 put forth more effort on the cognitive test, the scores may have reflected the difference in quantitative ability that was indicated by the SAT math scores. A similar phenomenon was found when comparing Class 1 and Class 2. Class 1 and Class 2 did not statistically significantly differ on entering math ability (SAT math) or on the cognitive test. However, note the relative pattern of test scores in Figure 5, in which Class 2 scored approximately 0.2 standard deviations higher on math ability than Class 1. Yet Class 1 scored approximately 0.2 standard deviations higher on the cognitive test than Class 2, possibly due to their increased effort. Unfortunately, scores on the cognitive test appear to reflect differences in motivation levels rather than variation in math ability. This inaccurate representation of math competence by the cognitive test has serious implications for assessing value-added, particularly if assessment results are reported for different groups (e.g., those who have completed a program/treatment versus those that have not)

and group membership is confounded by motivation. In these situations, the test scores may lead assessment professionals to conclude that there are not group differences in knowledge or ability when there actually are, or that there are group differences in knowledge or ability when there actually are not. Either situation results in invalid inferences made about what students know and can do, and ultimately program effectiveness.

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

There are several limitations of the current study that should be noted. First, although SAT math scores and NW-9 scores served as measures of math ability, we did not have a direct measure of expectancy for success for the cognitively-demanding test to aid in our understanding of test-taking motivation. Thus, future research should obtain measures of expectancy for success, such as self-efficacy. Although ability and expectancy have been shown to be positively related (Eccles et al., 1983; Wigfield & Eccles, 2000), obtaining direct measures of expectancy for success would allow researchers to fully examine Eccles' expectancy component by measuring not only whether examinees have the ability to perform well on a test, but also whether they expect to perform well. This could be especially pertinent in low-stakes testing contexts, wherein examinees may believe they have the ability to perform well, but the cost is too great for them to put forth effort and thus they do not expect to perform well. In these situations, it may be that expectancies for success (i.e., self-efficacy) help differentiate between examinees that are motivated and those that are not.

Second, Eccles' value component was not examined in the current study. Researching class differences in terms of the perceived importance (i.e., value) of each test could provide additional validity evidence for the three classes. From the perspective of Expectancy-Value theory, one would expect that greater valuation of a test would coincide with increased effort.

Thus, future research should measure how important students perceive each test to be (i.e., how much they value each test) by administering the Importance subscale of the SOS. Doing so could illuminate additional qualitative differences among the classes, and could potentially clarify how the perceived importance of a test influences test-taking motivation.

Third, the current study examined a testing situation in which a single, cognitively demanding quantitative and scientific reasoning test was administered. Similar studies using cognitively demanding tests of other domains would help to determine how expectancy, prior knowledge, and interest impact test-taking motivation. Although we do not think the general findings and conclusions of our study would change due to the subject matter of the test, this has yet to be formally tested.

Finally, in the current study, the cognitively demanding test was placed in the middle of the testing session. It may be helpful to examine how the placement of demanding tests within the testing session impacts test-taking motivation and the types of test-takers that emerge. For example, if the cognitive test were moved to the beginning of the testing session, what would the overall pattern of motivation scores look like? Would the dip in motivation for the cognitive test still occur for some test-takers? Would different types of test-takers emerge? It is unclear, at this point, how placement of the cognitive test would impact these results.

### **Conclusion**

In conclusion, we found evidence for three classes of test-takers that differ in their levels and patterns of test-taking effort across the course of a three-hour low-stakes testing session. The qualitatively distinct patterns of effort, as well as the fact that these types were differentiated by external criteria, suggest that these classes represent meaningful subgroups or types of test-takers. Given these findings, we believe that examining and reporting aggregate motivation data

may result not only in biased conclusions about test-taking motivation, but also biased conclusions about what students know and can do. Additionally, these classes appeared to differ from one another on relatively stable traits, such as achievement goals, personality dimensions, and ability. This may suggest that interventions and treatments designed to increase motivation in low-stakes contexts may be less effective than hoped. An alternative approach could be to report test scores separately for the different types of test-takers and interpret these scores taking into account the motivation of each of these groups. We feel such an approach would align with and extend the recommendations made by the Standards For Educational and Psychological Testing (AERA, APA, & NCME, 2004)

Regardless, given that accountability and low-stakes testing within K-12 and Higher Education will only continue to increase in the coming years, it is imperative to understand the motivation of examinees in these low-stakes testing contexts, especially if aggregate motivation scores do not reflect the motivation of all types of examinees. Doing so will help assessment practitioners understand the extent to which they can trust test scores and, consequently, will result in more valid inferences regarding student learning and program effectiveness.

## References

- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects of performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*, 163-181. doi:10.1037/a0015719
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.
- Asparouhouv, T. & Muthén, B.O. (2007). *Wald test of mean equality for potential latent class predictors in mixture modeling*. Retrieved January 25, 2008, from Mplus web site: <http://www.statmodel.com/download/MeanTest1.pdf>
- Atkinson, J. W. (1957). Motivational determinants of risk taking behavior. *Psychological Review, 64*, 201-252.
- 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. DOI: 10.1037/1082-989X.9.1.3
- 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.
- Clark, S. L. & Muthén, B. (2009, April). *Relating latent class analysis results to variables not included in the analysis*. Paper presented the annual meeting of the American Educational Research Association, San Diego, CA.
- 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. doi: 10.1007/s11336-007-9045-9

- 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. doi:10.1146/annurev.psych.53.100901.135153
- Ewell, P. T. (1991). To Capture the Ineffable: New Forms of Assessment in Higher Education. *Review of Research in Education*, *17*, 75-125. <http://www.jstor.org/stable/1167330>
- 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. DOI: 10.1177/0013164403258465
- 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).
- Kyllonen, P. (2005, September). The case for noncognitive assessments. *R & D Connections* (pp. 1-7). Retrieved September 1, 2009, from [http://www.ets.org/Media/Research/pdf/RD\\_Connections3.pdf](http://www.ets.org/Media/Research/pdf/RD_Connections3.pdf)
- Liem, A. D., Lau, S., & Nie, Y. (2008). The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology*, *33*, 486-512. doi:10.1016/j.cedpsych.2007.08.001
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767-778. Retrieved from <http://www.jstor.org/stable/2673445>

- Magnusson, D. (1998). The logic and implications of a person-oriented approach. In R. R. Cairns, L. R., Bergman, & J. Kagan (Eds.), *Methods and models for studying the individual: Essays in honor of Marian Radke-Yarrow* (pp. 33 – 64). Thousand Oaks, CA: Sage.
- 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. Retrieved from <http://www.jstor.org/stable/1164436>
- Muthén, L.K., & Muthén, B.O. (1998-2007). *Mplus User's Guide*. Fifth Edition. Los Angeles, CA: Muthén & Muthén
- 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. doi: 10.1207/s15326977ea0302\_2
- 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. doi:10.1016/j.cedpsych.2006.10.003
- 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.

- Pintrich, P. R. (1988). A process-oriented view of student motivation and cognition. In J. S. Stark & L. Mets (Eds.), *Improving teach and learning through research. New directions for institutional research*, 57 (pp. 55-70). San Francisco: Jossey Bass.
- Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review*, 16(4), 385-407. doi: 10.1007/s10648-004-0006-x
- Pintrich, P. R., & De Groot, E. V. (1999). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33-40. doi: 10.1037/0022-0663.82.1.33
- Robins, R. W., John, O. P., & Caspi, A. (1998). The typological approach to studying personality. In R. R. Cairns, L. R., Bergman, & J. Kagan (Eds.), *Methods and models for studying the individual: Essays in honor of Marian Radke-Yarrow* (pp. 135 – 160). Thousand Oaks, CA: Sage.
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464. Retrieved from <http://www.jstor.org/stable/2958889>
- 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., & Wise, S. L. (2003). Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests. *Paper presented at the National Council on Measurement in Education, Chicago, IL.*
- 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.*
- Tofghi, 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.
- Wainer, H. W. (1993). Measurement Problems. *Journal of Educational Measurement, 30*(1) 1-21. doi:10.1111/j.1745-3984.1993.tb00419.x
- Wigfield, A. (1994). Expectancy-value theory of achievement motivation: a developmental perspective. *Educational Psychology Review, 6*, 49-78. doi: 10.1007/BF02209024
- Wigfield, A., & Eccles, J. A. (1992). The development of achievement task values: a theoretical analysis. *Developmental Review, 12*, 265-310. doi:10.1016/0273-2297(92)90011-P
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81. doi:10.1006/ceps.1999.1015
- 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. doi: 10.1207/s15324818ame1902\_2
- Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: problems and potential solutions. doi:10.1207/s15326977ea1001\_1
- Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation

filtering in improving test score validity. *Educational Assessment, 11*, 65-83.

doi:10.1207/s15326977ea1101\_3

Wolf & Smith, (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=7363996&site=ehost-live&scope=site>

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. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=7366037&site=ehost-live&scope=site>

Yeo, G., & Neal, A. (2008). Subjective cognitive effort: A model of states, traits, and time. *Journal of Applied Psychology, 93*, 617-631. doi: 10.1037/0021-9010.93.3.617

Zilberberg, A., Brown, A. R., Harmes, J. C., & Anderson, R. D. (in press). How can we increase student motivation during low-stakes testing? Understanding the student perspective. In McInerney, D. M., Brown, G. T. L., & Liem, G. A. D. (Eds.), *Research on sociocultural influences on motivation and learning: Vol. 9. Student perspectives on assessment: What students can tell us about improving school outcomes*. Greenwich, CT: Information Age Publishing.

### Footnotes

<sup>1</sup>The first method is to incorporate the external variables into the mixture model as predictors or outcomes of class membership. This method is advantageous in that it takes into account the probabilistic nature of class membership (i.e., it takes into account that each person has a probability of belonging to each class). However, it is important to note that the solution will change depending on whether external variables are included in the model (Marsh, Ludtke, Robitzsch, & Trautwein, 2009). That is, when external variables are included in the model, class membership is not solely defined by test-taking effort, but instead the relationship between test-taking effort and external variables.

A second method assigns each individual to the class for which they have the highest posterior probability (modal assignment) and then conducts analyses such as ANOVA or regression to examine differences between the classes on each of the external variables. Unlike including the external variables in the mixture model, class membership is based only on effort scores, which often makes interpretation much easier. However, the method treats all members of a class as interchangeable, thereby introducing error variance, which biases the class membership-external variable relationships.

A third method derives class membership using only the effort scores, but unlike modal assignment, uses the posterior probabilities associated with each case to compute class means for each of the external variables (Asparouhouv & Muthén, 2007; Clark & Muthén, 2009). This method has the advantage of fixing class membership without introducing error due to modal assignment. Because of this advantage, and because this method has recently been implemented in software used for mixture modeling (posterior probability-based multiple imputation to test

equality of means across latent classes in Mplus 5.2; Muthén & Muthén, 1998-2007), we employed this approach for gathering validity evidence for the latent classes.

Table 1.  
*Demographic Characteristics of Sample (N = 887)*

---

Age	
Mean (SD)	17.49 (0.38)
Gender	
%Male	37.43
%Female	62.57
Ethnicity	
%White	80.04
%Black	2.93
%Asian	4.62
%Hispanic	2.48
%Other	9.92

---

Table 2.  
*Descriptive Statistics for Overall Sample*

Measure	Mean	SD	$\alpha$
<i>Effort</i>			
Effort1	3.95	3.95	0.78
Effort2	3.91	3.91	0.81
Effort3	3.51	3.51	0.82
Effort4	3.96	3.96	0.81
Effort5	3.85	3.85	0.79
<i>Need for Cognition</i>	3.31	0.56	0.84
<i>Achievement Goals</i>			
Mastery approach	5.81	0.97	0.80
Performance approach	5.48	1.30	0.89
Mastery avoidance	4.57	1.19	0.51
Performance avoidance	4.94	1.30	0.66
Work avoidance	2.78	1.11	0.81
<i>Quantitative "Ability"</i>			
SAT math	579.72	579.72	N/A
NW-9 Score	44.03	7.79	0.80
<i>Big Five</i>			
Extraversion	3.53	0.77	0.87
Agreeableness	3.98	0.56	0.79
Conscientiousness	3.60	0.57	0.78
Neuroticism	2.75	0.70	0.80
Openness	3.50	0.58	0.77

Table 3.  
*Bivariate Correlations for Five Effort Scores N= 887*

	Eff1	Eff2	Eff3	Eff4	Eff5
Eff1	1				
Eff2	0.627	1			
Eff3	0.419	0.483	1		
Eff4	0.507	0.543	0.345	1	
Eff5	0.474	0.552	0.410	0.629	1

Table 4.

*Fit for Various Models of Effort Scores across the Testing Session (N = 887)*

	AIC	BIC	SSABIC	LMR	Entropy	LL	# of free parameters
1-Class	10,243.80	10,291.68	10,259.92	N/A	N/A	-5,111.90	10
1-Class C	8,614.08	8,709.84	8,646.32	N/A	N/A	-4,287.04	20
2-Class A	9,064.26	9,140.86	9,090.05	$p < 0.001$	0.783	-4,516.13	16
2-Class B	8,929.63	9,030.17	8,963.48	$p = .0017$	0.909	-4,443.82	21
2-Class C	7,907.42	8,055.85	7,957.40	$p < 0.001$	0.687	-3,922.71	31
2-Class D	7,859.90	8,056.20	7,925.99	$p < 0.001$	0.699	-3,888.95	41
3-Class A	8,734.79	8,840.12	8,770.25	$p < 0.001$	0.787	-4,345.39	22
3-Class B	8,339.10	8,492.31	8,390.68	$p = 0.011$	0.829	-4,137.55	32
3-Class C	7,740.95	7,942.04	7,808.66	$p = 0.330$	0.753	-3,828.48	42
3-Class D	7,595.59	7,892.43	7,695.53	$p = 0.138$	0.890	-3,735.79	62
4-Class A	8,648.16	8,782.22	8,693.30	$p = 0.164$	0.829	-4,296.08	28
4-Class B	8,121.61	8,327.48	8,190.92	$p = 0.257$	0.786	-4,017.80	43
4-Class C	7,669.75	7,923.51	7,755.19	$p = 0.639$	0.648	-3,781.88	53
4-Class D <sup>1</sup>							

<sup>1</sup> The best log-likelihood did not replicate.

Table 5.

*Classification of most likely class membership based on posterior probabilities (row) and modal assignment (column) for three-class Model D*

	1	2	3
1	0.981	0.000	0.019
2	0.075	0.909	0.016
3	0.093	0.028	0.879

Table 6.  
*Means, SD, and Correlations for Effort Scores for the 3-Class Solution*

Class 1 ( $N = 73.83$ )					
	Eff1	Eff2	Eff3	Eff4	Eff5
Eff1					
Eff2	0.16				
Eff3	0.23	0.06			
Eff4	0.00	0.38	-0.01		
Eff5	0.03	0.10	-0.14	0.32	
Mean	4.78	4.95	4.02	4.96	4.45
SD	0.28	0.09	0.78	0.09	0.80
Class 2 ( $N = 624.58$ )					
	Eff1	Eff2	Eff3	Eff4	Eff5
Eff1					
Eff2	0.49				
Eff3	0.28	0.35			
Eff4	0.32	0.32	0.13		
Eff5	0.35	0.43	0.27	0.52	
Mean	3.84	3.79	3.32	3.86	3.76
SD	0.68	0.73	0.80	0.74	0.75
Class 3 ( $N = 188.59$ )					
	Eff1	Eff2	Eff3	Eff4	Eff5
Eff1					
Eff2	0.85				
Eff3	0.80	0.92			
Eff4	0.81	0.93	0.99		
Eff5	0.79	0.92	0.98	0.99	
Mean	3.99	3.93	3.92	3.93	3.92
SD	0.65	0.69	0.73	0.75	0.75

Table 7.

*Equality of Class Means based on Posterior Probabilities<sup>1</sup>*

Measure	Class 1 <i>N</i> = 73.83	Class 2 <i>N</i> = 624.58	Class 3 <i>N</i> = 188.59	$\chi^2(2)$ , <i>p</i> -value
<i>Need for Cognition</i>	3.47	3.28	3.35	5.305, <i>p</i> = .070
<i>Achievement Goals</i>				
Mastery approach	6.32 <sup>a</sup>	5.76 <sup>b</sup>	5.78 <sup>b</sup>	37.257, <i>p</i> < .001
Performance approach	5.97 <sup>a</sup>	5.43 <sup>b</sup>	5.44 <sup>b</sup>	11.591, <i>p</i> = .003
Mastery avoidance	4.83	4.59	4.42	5.292, <i>p</i> = .071
Performance avoidance	5.11	4.93	4.93	1.095, <i>p</i> = .578
Work Avoidance	2.19 <sup>a</sup>	2.81 <sup>b</sup>	2.90 <sup>b</sup>	26.502, <i>p</i> < .001
<i>Quantitative "Ability"</i>				
SAT math	569.54 <sup>a</sup>	576.81 <sup>a</sup>	593.22 <sup>b</sup>	8.727, <i>p</i> = .013
NW-9 Score	44.56 <sup>a,b</sup>	43.52 <sup>a</sup>	45.49 <sup>b</sup>	6.885, <i>p</i> = .032
<i>Big Five</i>				
Extraversion	3.64	3.54	3.44	3.010, <i>p</i> = .222
Agreeableness	4.22 <sup>a</sup>	3.96 <sup>b</sup>	3.96 <sup>b</sup>	17.340, <i>p</i> < .001
Conscientiousness	3.87 <sup>a</sup>	3.56 <sup>b</sup>	3.60 <sup>b</sup>	21.519, <i>p</i> < .001
Neuroticism	2.73	2.79	2.64	4.303, <i>p</i> = .116
Openness	3.72 <sup>a</sup>	3.47 <sup>b</sup>	3.52 <sup>b</sup>	10.277, <i>p</i> = .006

<sup>1</sup>Means and tests of equality were computed using posterior probabilities estimated via the mixture model analyses. If the omnibus test was significant, pairwise comparisons are reported with non-common superscripts indicating means that are statistically different at *p* < 0.05

Figure 1. Example effort scores across a testing session.

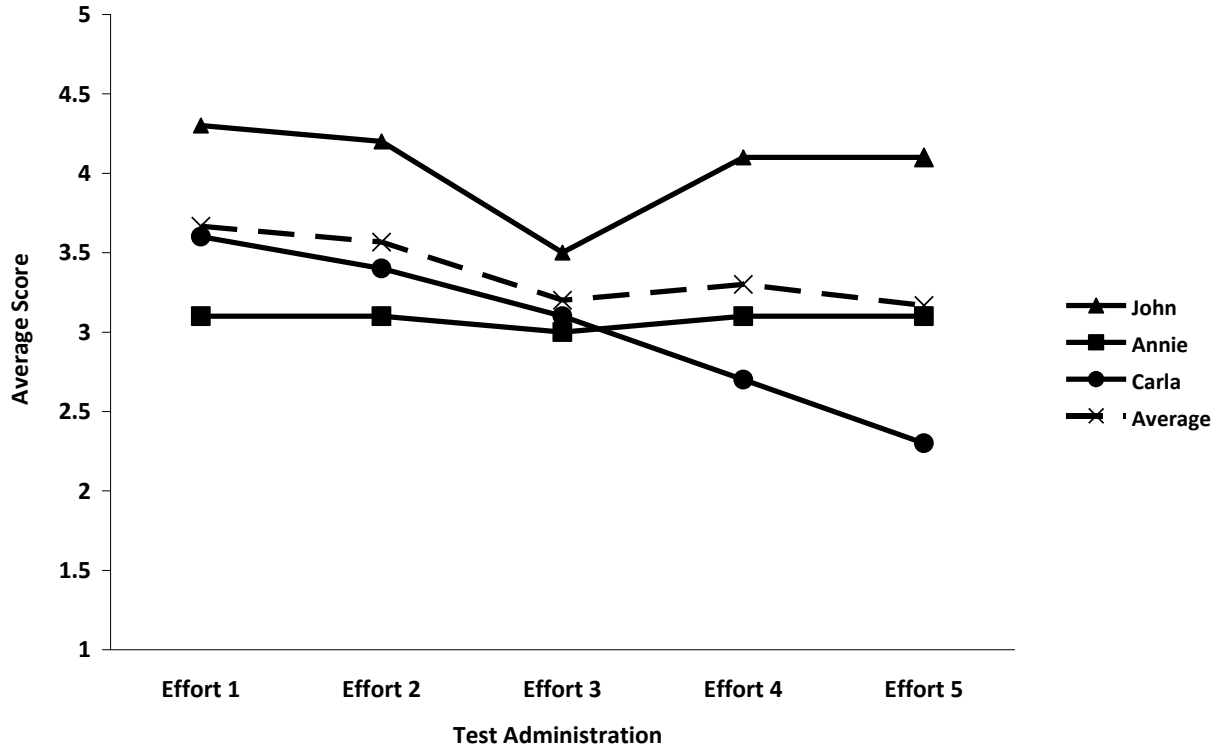


Figure 2. One class model: Average effort scores across the five time points ( $N = 887$ ).

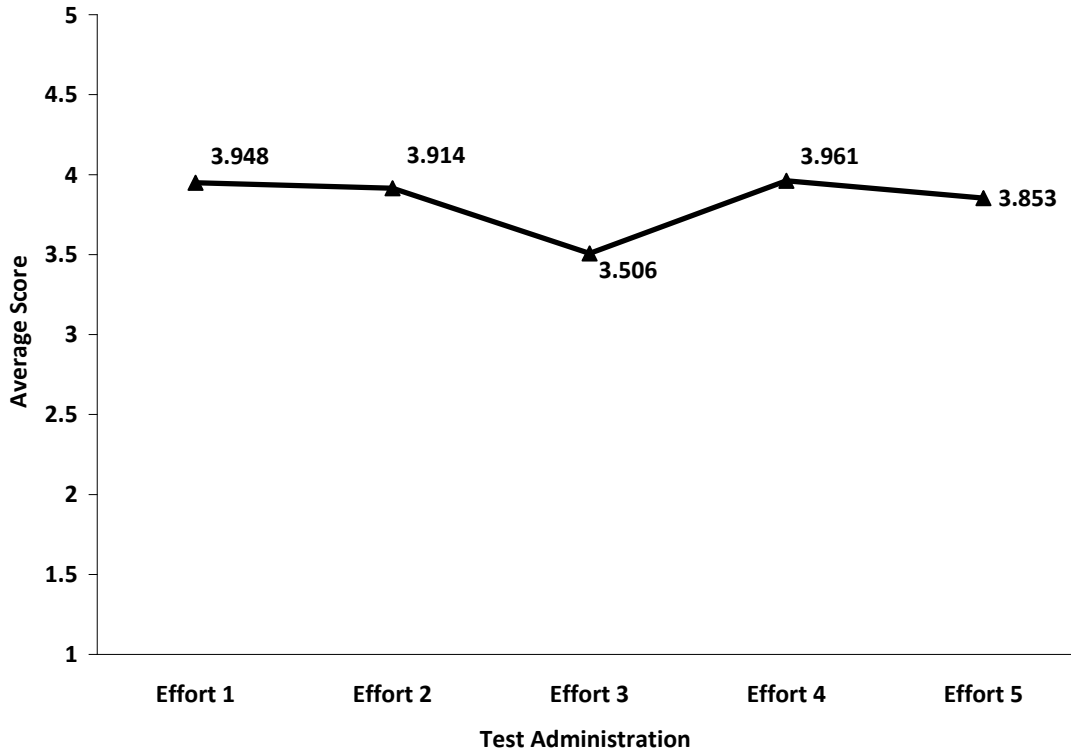


Figure 3. Graphic representation of two-class model D ( $N = 887$ ).

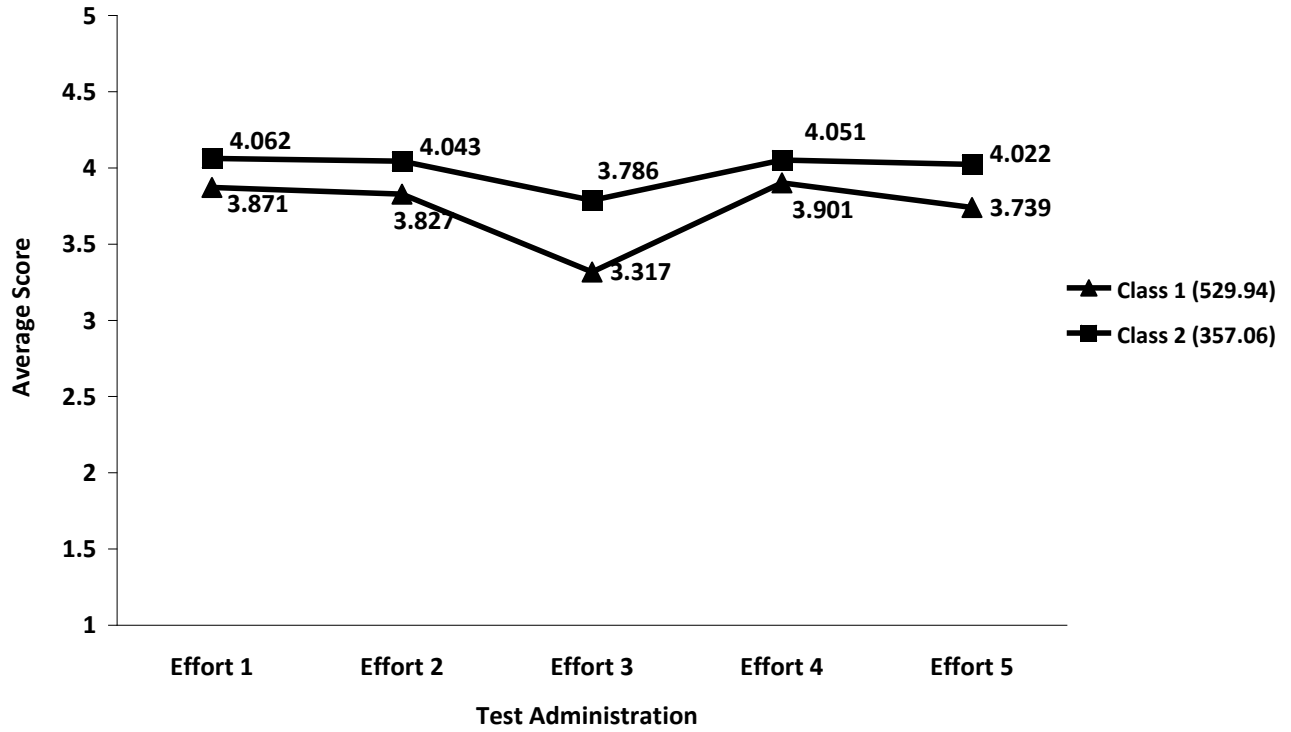


Figure 4. Graphic representation of three-class model D ( $N = 887$ ).

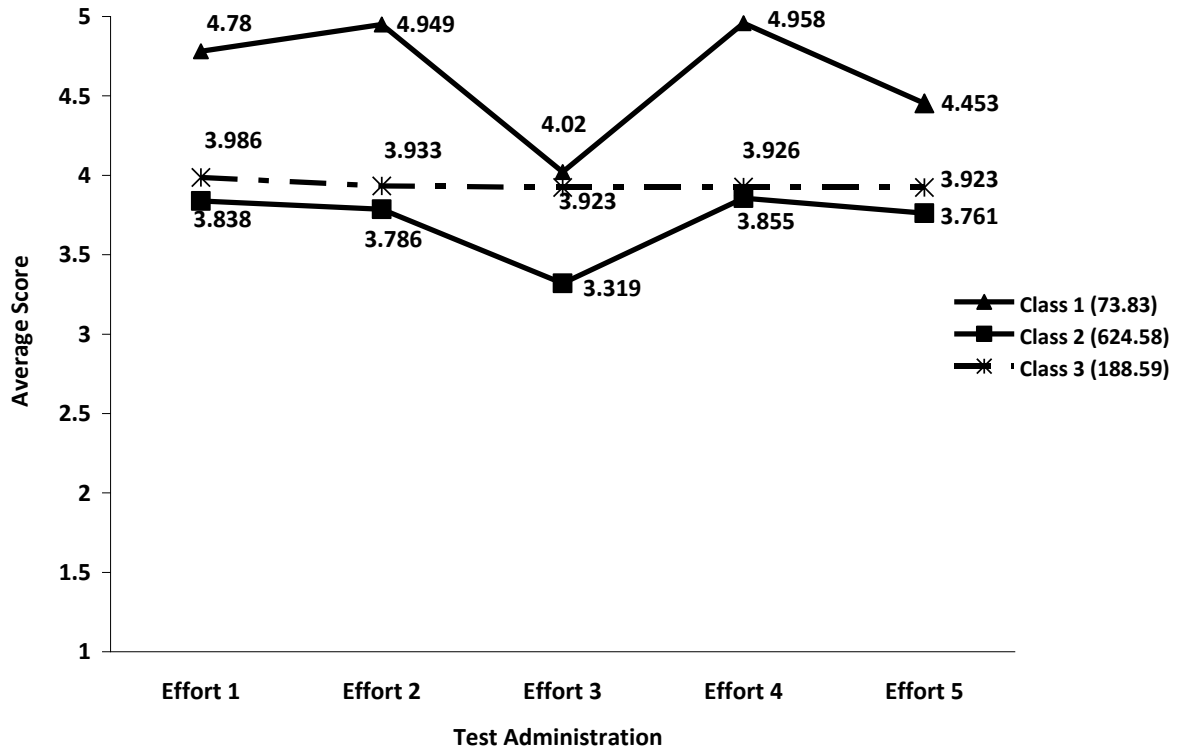


Figure 5. Graphic representation of class averages (z-scores) on external variables ( $N = 887$ ).

