RUNNING HEAD: Skipping the Test

Skipping the Test:
Using Evidence to Inform Policy Related to
Those Students Who Avoid Taking Low-Stakes Assessments in College

Peter Swerdzewski

Sara J. Finney

J. Christine Harmes

James Madison University

Paper presented at the annual meeting of the Northeastern Education Research Association,

Rocky Hill, CT, October 2007

Abstract

Students who avoided low-stakes testing were studied to determine if their scores on cognitive and developmental tests differ from those of students who do not avoid such testing. When later forced to complete the testing, the students who initially avoided it were further subdivided based on the effort they exerted on the tests. It was found that students who avoided low-stakes testing but did exert effort on these tests once they were required to take them, scored similarly or higher on cognitive tests compared to students who did not initially avoid the low-stakes testing. The "avoiders" also had different affective profiles compared to the "attenders" and these profiles were a function of the amount of effort exerted on the tests. The results from this study suggest that assessment practitioners need to include test scores from students who avoid taking low-stakes tests to ensure that the results are reflective of the overall population of students under study. The results also underscore the need for practitioners who are making inferences from test scores to consider the effort that students exert on low-stakes tests.

KEYWORDS: low-stakes testing, motivation, large-scale testing, accountability, higher education

Skipping the Test:
Using Evidence to Inform Policy Related to
Those Students Who Avoid Taking Low-Stakes Assessments in College

Despite the best efforts of psychometricians and assessment practitioners to make valid inferences from the instruments they develop and administer, these inferences are predicated on two important student behaviors: the students must *show up* to take the test in order for the student population to be accurately represented, and the students must *give effort* when responding in order for responses to be valid. These two behaviors can be major issues in low-stakes testing contexts. Testing companies (e.g., ETS, ACT, Pearson), researchers who pull from voluntary subject pools, and the assessment programs located on nearly ever college campus in the country often rely heavily low-stakes tests. These low-stakes tests have few—if any—repercussions for students who choose not to attend the test administrations, despite the importance of the results to the programs being assessed (Sundre & Kitsantas, 2004). Who are these students who don't attend low-stakes testing sessions, why do they choose not to attend, and most importantly, what are the implications of not including their data in both the test development process and in the assessment of instruction? These questions have implications for student learning at both a localized level (such as assessing a university general education program) and at a national level (such as assessing NCLB).

Purpose of the Study

By conducting this study, we hope to learn more about those students who avoid low-stakes tests so the testing and educational communities, who consistently employ low-stakes tests for the purposes of accountability, piloting items, or exploring innovative measurement techniques, will have evidence to make informed decisions about the policies related to this

unique population of students. Although our data are collected in the context of a low-stakes university testing endeavor, the results are informative to any low-stakes testing context (e.g., the National Assessment of Educational Progress, or NAEP).

Motivation Issues in Low-Stakes Testing

Despite the recent increase of low-stakes testing for accountability purposes (e.g., NCLB, accreditation, state mandates), there is little scholarship on the subject of test avoidance. Researchers in the domain of low-stakes testing are generally concerned with the low motivation of those who *do* attend testing sessions; so, given the lack of extant literature that examines whether or not a given student decides to engage in a low-stakes test, we must instead look to the literature on how much effort students exert once they decide to attend a low-stakes testing session. In an extensive study on student motivation as it relates to low-stakes assessments, the authors found that test-taking motivation was a predictor in test performance (Sundre & Kitsantas, 2004). A similar study by Wise and DeMars (2005) proposed that a student's effort on a low-stakes test is a function of a four-pronged framework: (a) the student's perceived success on the test, (b) how much effort the student believes the test will consume, (c) the student's perceived importance of the test, and (d) the student's affective and emotional reaction to the various test items.

Although the above studies focused on the role of students' effort when completing the test, researchers are beginning to question the role that test avoidance plays in the validity of the inferences one can make from low-stakes test scores. Among the most prominent testing programs that are experiencing participation issues (in addition to effort, issues) is the NAEP. In 2002, the 12th grade participation rate dropped to 55% (National Commission on 12th Grade Assessment and Reporting, 2004). The National Assessment Governing Board highlighted the

need for information related to the reasons for non-participation, and the effect of non-participation on the quality of the results from the NAEP (Chromy, 2005).

Given the important validity implications at many levels for investigating test avoidance, we sought to understand more about those students who choose not to participate in low-stakes testing and how their test results (or lack thereof) affect the validity of an assessment program predicated on measures that are low-stakes for examinees. Specifically, we investigated three research questions:

1. What percentage of students did not show up for a set of low stakes tests, who are these students, and do their cognitive and developmental profiles differ from the attenders?
2. Of the students who did not show up, what percentage of them exhibited effort on all low-stakes tests during a required make-up session, and who are these students?
3. What are the characteristics of those students who did not exhibit effort during the required make-up sessions?

Methods

Data collected during the 2005-2006 academic year for the assessment program at a mid-sized mid-Atlantic university were used to profile students who avoided taking a battery of low-stakes tests. All students who had between 45 and 70 earned credit hours (i.e., most students were in their second year) were required to participate in a series of assessments of their skills and knowledge. Notification of this requirement came to each examinee via a personalized letter and an additional personalized e-mail addressed to each student who fit the testing criteria. Additionally, fliers were placed in campus busses and on campus dining room tables and the campus newspaper ran an advertisement about the required assessments. The results from these assessments were used to inform the university's general education program. Students were also

asked to complete a number of affective and developmental scales, and the data from these scales were used for the assessment of the university's student affairs programs. The "Assessment Day" took place on the second Tuesday of February 2006. All classes at the university were cancelled to allow for the assessments. Each student who was required to complete assessments was randomly assigned to either a morning or afternoon session. All sessions were three hours in length and were managed by trained proctors. Due to the assessment design at the university, many students in the current study took some of the tests when they entered the institution as freshmen (i.e., a year-and-a-half earlier), thus the February 2006 administration served as a posttest for a subset of students.

Participants

A total of 2,965 second-year students who were required to take the assessments were asked to attend an assigned session on a specific day during the spring semester; however, approximately 753 students, or about 25%, chose not to attend. The 753 students compose this study's test "avoider" group. Those who attended assessments on the assigned day and were assigned to computer-based testing (CBT) sessions comprise the comparison group, and are referred to as the "attenders." One may ask, "Why are only student who tested via CBT included in this study?" The amount of time each student spent on each item was collected by the CBT software employed in this study as an indicator of each student's test-taking motivation. It is obviously impractical to collect similar data from students who complete tests via paper and pencil, so only CBT data was used in this study. Both the avoiders and the attenders completed the same tests and tested under the same conditions. Those students who did not attend the original assessment session (avoiders) were required to attend a make-up session to remove an administrative hold on their student record.

The computer-based testing sessions consisted of six affective/developmental measures (measures of (1) academic motivation and one's sense of belonging, (2) beliefs about learning related to the collegiate environment, (3) social self-efficacy, (4) worry, (5) diversity, and (6) test-taking motivation), and two criterion-referenced measures of academic knowledge (quantitative/scientific reasoning and fine arts/humanities; see Table 1 for a chart of measures). Additionally, the 621 of the 753 students who attended the make-up assessment completed a questionnaire that elicited value judgments about assessment and reasons for not attending the original testing session.

Data Cleaning and Management

Interpretable analyses required that stringent data management procedures were followed, which reduced the useable sample sizes for both the avoider and attender samples. Of the 753 avoiders, 621 completed a required make-up assessment within three months of the original required assessment date. An additional 52 avoiders were additionally removed from the study because their data from the student information system was insufficient for the purposes of this study. Another 81 avoiders were removed from the data set because they had incomplete or unusual test data due to technical issues (the CBT software required responses for every item, so traditional missing data issues were not a concern in this study). This resulted in a final data set of 488 students labeled as "avoiders."

The comparison group (the "attenders") initially included 326 students who were randomly assigned to complete their required assessments via computer and who did attend their originally-scheduled session. Of these 326 attenders, 17 were removed from the study due to incomplete student information system data and an additional 6 were removed due to incomplete

or unusual test data (again, due to technical issues, not due to student-related issues). This resulted in a final data set of 303 students labeled as "attenders."

Unless otherwise noted, a total of 791 students were involved in this research project: 488 who skipped the original required assessment day (avoiders), and a comparison group of 303 who did attend the original assessments day and who took the computer-based assessments (attenders).

Instrumentation

Students completed a total of eight tests of varying length and content. Two of the tests, referred to in this study as the cognitive tests, evaluate the effectiveness of the fine arts and humanities component (Fine Arts test) and the quantitative and scientific reasoning component (UOW test) of the university's general education program. The remaining six tests, referred to in this study as the developmental tests, assessed students in terms of their psychological and educational development. Information about the scales used in this analysis is reported in Appendix A and a short description of each test, as well as the order in which each was given, is in Table 1.

Table 1

Short Description of Each Test

Order*	Test Name	Type**	Description
1.	Attitudes Toward Learning (ATL)	NC	Measures students' knowledge and beliefs related to learning. Scale consists of seven separate scales.
2.	Learning Environment Questionnaire (LEQ)	NC	Measures students' academic motivations and beliefs about learning under various environmental conditions in college.
3.	The Fine Arts	Cog	Measures students' understanding of the fine arts, humanities, and literature.
4.	Scale of Perceived Social Self-Efficacy (PSSE)	NC	Measures students' perceived abilities to engage in the social interactional tasks necessary to function in a college environment.

5.	Student Worry Questionnaire (SWQ)	NC	Measures both the process and content of worry as it is experienced by the traditional college student population.
6.	Interacting with Others (MGUDS)	NC	Measures students' appreciation for, and comfort with diversity.
7.	Understanding our World (UOW).	Cog	Measures students' quantitative and scientific reasoning skills.
8.	Student Opinion Survey (SOS)	NC	Measures the effort a student exerts on a test (or tests) and the importance he or she places on the test(s).

* The order column refers to the order in which all students encountered the eight tests.

** NC = Non-cognitive test; Cog = Cognitive test.

Identifying Motivated and Unmotivated Test-Takers

A direct, yet non-intrusive measure of examinee motivation that capitalizes on CBT technology is Response Time Effort (RTE; Wise & Kong, 2005). Within a CBT, an examinee views items one-at-a-time, and the interface records the amount of time the examinee spends on each item before he or she chooses to view the next item. This response time can then be compared to a predetermined threshold representing the minimum possible time necessary to read and respond to the item (for a discussion of methods for setting item thresholds, see Kong, Bhola, & Wise, 2005). An examinee's overall index of RTE for the whole test is calculated as the proportion of items on the test in which the examinee's response time exceeds the threshold. For this research study a student was said to have tried on a given test if he or she had an RTE of 0.90 or higher for the test. In other words, a student is said to have been motivated on a test if he or she spent enough time on the test to read 90% of the items. Each student is thus said to have tried or not tried on each of the eight tests administered in the study.

Results and Conclusions

Profile of the Test Avoidance Group

The first research question sought to develop a profile of the test avoider group. Of the 2,965 students who were required to complete assessments, 753 students (25.40%) chose not to attend the original testing session. These avoiders were required to come to a make-up session. Examinees from the avoidance group were compared to examinees from the attendance group on test performance along with various demographic variables (see Table 2 for an overview of results; see Appendix B for a full table of results).

Table 2

Overview of Key Results for Research Question One

	Attenders (All Students)			**Avoiders (All Students)**			
	Mean	*SD*	*N*	Mean	*SD*	*N*	*d*
GPA	3.02	0.53	303	2.80	0.60	488	0.39*
Number of Earned Credits	53.10	7.28	303	55.27	8.21	488	-0.28*
Age (in years)	20.06	0.67	303	20.42	1.14	488	-0.36*
ATL Subscale Scores:							
Mastery-Approach	5.80	0.95	303	5.46	1.11	488	0.32*
Performance-Approach	5.36	1.38	303	4.97	1.53	488	0.27*
Work-Avoidance	2.92	1.24	303	3.37	1.27	488	-0.35*
LEQ Subscale Scores:							
Autonomy	5.23	0.87	303	5.03	0.99	488	0.21*
Competence	5.87	0.87	303	5.57	1.03	488	0.32*
Intrinsic Motivation to Experience Stimulation (IMES)	3.79	1.52	303	4.14	1.57	488	-0.23*
Amotivation (AM)	1.44	0.92	303	2.05	1.48	488	-0.48*
FineArts % Correct Score	57%	9%	303	50%	10%	488	0.74**
SWQ Subscale Scores:							
Financial-Related Concern	2.63	1.05	303	2.90	1.08	488	-0.25*
UOW % Correct Score	58%	15%	137	45%	18%	484	0.77**
SOS Subscale Scores							
Effort	3.16	0.74	303	2.84	0.75	488	0.42*
Importance	2.70	0.76	303	2.50	0.82	488	0.25*

* "Small" Cohen's *d* (between .2 and .5)

** “Medium” Cohen’s *d* (between .5 and .8)

Given that the primary purpose of the required university assessment day was to collect data on various cognitive tests to inform the institution’s general education curriculum, the scores of greatest interest are those on the Fine Arts and the Understanding our World tests. Avoiders scored substantially lower on average than attenders on both the Fine Arts test (d = 0.74) and the Understanding our World test (d = 0.77). The staggering differences between the groups is clearly a cause for concern; yet the probable explanation for the disparity is likely found in the results of the last instrument that students took: a self-report survey in which students indicated the degree to which they exerted effort on the tests, as well as the degree to which they felt the tests were important. Avoiders indicated that they exerted much less effort on the tests (d = 0.42) and felt the tests were less important (d = 0.25) compared to attenders.

Effort Levels of Test Avoiders

The second research question focused on the effort exerted by the avoiders and the attenders. Specifically, when trying to understand the difference between the attenders and the avoiders, simple attendance does not paint the entire picture: one must also look to students’ motivations *in concert* with whether or not they chose to skip the tests. More specifically, not all attenders “tried” on all of the tests; therefore, some of the data analyzed above is not a valid representation of what students understand about the Fine Arts and Humanities. This same logic follows for avoiders: when they attended the required make-up session, they may not have exerted effort when responding, thus yielding data that is not useful for evaluation purposes. Therefore, an important question remains: do the attenders and avoiders actually put forth effort when they complete the low-stakes tests?

If an examinee did not try on at least 90% of the items on a given test (based on the item response time), the examinee was classified as not having exerted effort (or "tried") on the test. Students who exerted effort on all eight tests were termed "total triers." A comparison of the total triers who attended the regular assessment day to the total triers from the test avoidance group illustrates some interesting differences (see Table 3 for an overview of results; see Appendix C for a full table of results). Of the attenders, 47.19% were total triers versus 12.70% of avoiders who were similarly total triers. Students classified as total triers from the test avoidance group had lower GPAs (d = 0.40) and were older (d = 0.42) than the total triers from the attenders group. These differences are nearly identical to the differences between the larger groups of avoiders and attenders.

Table 3

Overview of Key Results for Research Question Two

	Attenders (Total Triers Only)			**Avoiders (Total Triers Only)**			
	Mean	*SD*	*N*	Mean	*SD*	*N*	*d*
GPA	3.06	0.59	122	2.80	0.77	53	0.40*
Number of Earned Credits	53.07	6.99	122	56.04	8.13	53	0.40*
Age (in years)	20.10	0.90	122	20.56	1.44	53	-0.42*
ATL Subscale Scores:							
Mastery-Avoidance	4.44	0.94	122	4.68	0.80	53	-0.26*
Performance-Approach	5.43	1.24	122	4.99	1.63	53	0.33*
Work-Avoidance	2.76	1.18	122	3.05	1.15	53	-0.25*
LEQ Subscale Scores:							
Extrinsic Motivation Introjected (EMIN)	4.99	1.36	122	4.61	1.46	53	0.27*
Generalized Anxiety Disorder Symptoms	3.13	0.92	122	2.95	0.80	53	0.20*
UOW % Correct Score	65%	11%	45	69%	11%	53	-0.32*

* "Small" Cohen's *d* (between .2 and .5)

** "Medium" Cohen's *d* (between .5 and .8)

Developmental profiles of total triers across groups. Interestingly, the developmental differences between the attender and avoider groups look very different when only the total triers are considered from both groups. For example, when simply comparing the total group of test avoiders to the total group of attenders (e.g., not considering motivation), the attenders scored higher in mastery-approach than the avoiders (d = 0.32). However, when evaluating only the total triers (e.g., those students who were motivated on all eight tests), the goal orientation profile changes: the difference between the two groups on mastery-approach virtually disappears, indicating that, when compared to the total triers in the attenders group, those students who initially skipped the tests (avoiders) but exhibited motivation on the tests as they were taking them (triers), were just as likely to have the goal of learning as much as they can possibly learn during the semester (mastery-approach). Furthermore, the total triers in the avoider group were higher on mastery-avoidance than those in the attender group (d = 0.26), a result that was masked when looking at the difference between attenders and avoiders overall. It is interesting to note that triers who initially avoided the test tend to endorse this goal more than attenders who tried on the test. Additionally, the total triers in the two groups varied on the extrinsic motivation introjected scale, indicating that the test avoiders are *less likely* than the attenders to be in college to prove to themselves that they are capable of completing a college degree or to show that they are intelligent people. Interestingly, the total triers in the two groups did *not* differ greatly in terms of autonomy (d = 0.05), competence (d = 0.11), or amotivation (d = 0.01).

Differences on cognitive tests for total triers across groups. In terms of the cognitive tests, one might assume that motivated students would score similarly, regardless of their choice to attend the original testing session. However, the total triers in the test avoider group actually scored on average *higher* on the Understanding Our World test compared to the total triers from

the attender group (d = 0.32). The difference on the Fine Arts test is negligible. Based on these results, it appears that in program assessments employing low-stakes testing, excluding the total triers who avoid tests from the analysis *could* bias the results, as the total triers in the test-avoidance group scored higher on one of the cognitive tests than the total triers who chose to attend the regular assessment session. This finding is not only interesting as it is the opposite of what many would expect, but is also important as it provides evidence that a segment of the test avoider population (those test avoiders who try on tests once required to take them) includes some highly-skilled and high-knowledgeable individuals.

Profile of Low-Effort Test Avoiders

Given the results of research question two, one may think that including test avoiders in an assessment of a program is necessary to ensure an accurate representation of the overall population. However, as discussed above, this is too general of a statement because all test avoiders do not put forth effort on all tests, and, therefore, their data should not be use to make evaluative decisions. Only data from avoiders who put forth effort should be merged with attenders who put forth effort for the purposes of analyzing data for decision-making.

To further explore the effect of effort in test-avoider group, students were divided into two groups: those who tried on all tests (again termed the "total triers"; N = 53) and those who did not try on the two cognitive tests ("cognitive non-triers"; N = 266). These two groups were compared on demographic variables, cognitive test scores, and developmental test scores (for only those developmental tests on which students in *both groups* demonstrated effort, thus N varies for these comparisons). For example, with the Attitudes Toward Learning instrument, all total triers in the test avoider group were compared to the cognitive non-tiers who *did* try on the Attitudes Toward Learning test (N = 122), although they may not have tried on other

developmental tests (see Table 4 for an overview of results; see Appendix D for a full table of results). This allowed us to profile students who tried on everything versus those who didn't try on the more difficult cognitive test with respect to developmental outcomes.

Table 4

Overview of Key Results for Research Question Three

	Avoiders (Total Triers)			**Avoiders (Cognitive Non-Triers)**			
	Mean	*SD*	*N*	Mean	*SD*	*N*	*d*
Highest Combined SAT Score	1175.53	98.95	53	1152.10	114.91	266	0.21*
ATL Subscale Scores:							
Mastery-Approach	5.84	0.79	53	5.43	1.00	122	0.44*
Mastery-Avoidance	4.68	0.80	53	4.33	1.01	122	0.36*
Work-Avoidance	3.05	1.15	53	3.43	1.28	122	-0.31*
LEQ Subscale Scores:							
Autonomy	5.33	0.80	53	4.78	1.06	103	0.57**
Competence	5.89	0.95	53	5.48	1.08	103	0.40*
Interest	2.63	0.45	53	2.33	0.63	103	0.52**
Enjoyment	3.78	0.60	53	3.35	0.71	103	0.64**
Intrinsic Motivation to Know (IMTK)	5.42	1.02	53	4.94	1.24	103	0.41*
Extrinsic Motivation External Regulation (EMER)	5.33	1.38	53	5.85	0.95	103	-0.47*
Amotivation (AM)	1.28	0.58	53	1.76	1.05	103	-0.52**
Fine Arts % Correct Score	60%	7%	53	44%	8%	266	2.00***
SWQ Subscale Scores:							
Social Adequacy Concern	2.90	0.84	53	2.73	0.85	87	0.20*
Financial-Related Concern	2.53	1.07	53	2.75	1.07	87	-0.20*
UOW % Correct Score	69%	11%	53	35%	12%	266	2.84***
SOS Subscale Scores							
Effort	3.71	0.60	53	2.71	0.70	266	1.46***
Importance	3.02	0.86	53	2.44	0.80	266	0.71**

* "Small" Cohen's *d* (between .2 and .5)

** "Medium" Cohen's *d* (between .5 and .8)

*** "Large" Cohen's *d* (greater than .8)

In terms of demographics, the total triers are relatively similar to the test avoiders who did not try on the cognitive tests. There are proportionally more women than men in the total trier group, whereas there are approximately equal proportions of males and females in the cognitive non-trier group. Additionally, the highest combined SAT score for the total triers is higher than that of the cognitive non-triers (d = 0.21).

The more substantial differences between the total triers and the cognitive non-tiers from the test avoidance group are apparent when evaluating the developmental measures. The total triers are much higher than the cognitive non-triers in terms of mastery-approach (d = 0.44) and mastery-avoidance (d = 0.36) goal orientations. Not surprisingly, the cognitive non-triers are higher on work-avoidance (d = 0.31) than the total triers. The cognitive non-triers had lower scores than the total triers on the feelings of academic autonomy subscale (d = 0.57), feelings of academic competence (d = 0.40), interest in academics (d = 0.52), and enjoyment in academics (d = 0.64). In terms of the students' reasons for attending college, the cognitive non-triers scored lower on the intrinsic motivation to know subscale (d = 0.41), indicating that they are less interested in attending college for the satisfaction of learning new things. The cognitive non-tiers scored higher on the extrinsic motivation for external regulation subscale (d = 0.52), indicating that they are more interested in attending college for reasons related to obtaining a more prestigious and/or higher-paying job. As one would expect, the cognitive non-triers demonstrated higher levels of academic amotivation (d = 0.52) compared to the total triers, indicating that the cognitive non-triers are either unsure of why they are in school, or simply do not care about attending college.

Comparing the scores on cognitive tests for a group of students who demonstrated effort on the tests (total triers) versus a group of students who did not read at least 10% of the items

(cognitive non-triers) is a comparison that echos the “apples and oranges” analogy and some may say the comparison should not be made because the scores from the cognitive non-triers are not valid due to no expended effort. However, we thought it interesting to present these results because it is common practice to accept test results from students in low-stakes assessment environments without filtering out the test responses for which students did not exhibit effort. Appendix D compares results on the Fine Arts test and the Understanding Our World test for the students who did exhibit effort on all tests included in these tests (total triers) versus students who did not exhibit effort (e.g., did not read at least 10% of the items) on the two cognitive tests (cognitive non-triers). As expected, the cognitive non-triers scored, on average, 16 percentage points lower than the total triers (d = 2.00) on the Fine Arts test and, on average, 35 percentage points lower than the total triers (d = 2.84) on the Understanding Our World test.

In addition to the whopping difference in performance, a nice validity check for the creation of the two groups (total triers vs. cognitive non-triers) are their scores on the test effort and test importance questions., The cognitive non-triers indicated they exhibited much less effort on the tests than the total triers (d = 1.46) and felt the tests were much less important (d = 0.71). Although the results of research question two indicate the low-stakes test results from test avoiders should be included in the assessment of programs, the findings from this research question indicate that *one can only make valid inferences from the results of test avoiders if these results are first filtered based on the effort students exhibited on the various tests.* Therefore, although we believe that ignoring test avoiders presents a biased view of student achievement, we only advocate their inclusion with test attender scores if the avoiders put forth effort on the test(s).

Discussion

Understanding the characterization of the student who chooses to "skip" a low-stakes assessment is important both to understanding the impact of the student's absence on the results of a study, and to the crafting of interventions to encourage this segment of the population to attend low-stakes testing sessions. From this study four key findings emerge that all carry major policy implications. Following, we provide recommendations to address these main findings.

Recommendation 1: Include Test Avoiders. Practitioners should include test avoiders in their assessments and test development processes to ensure their samples are representative of their overall population under study. Given the findings of the current study, excluding test avoiders from a sample seriously limits the generalizability of the test results. For example, when making the simple comparison of test avoiders to attenders (not considering the motivation that students put forth on the low-stakes tests), the two groups in this study differed by a Cohen's *d* of .20 or more on 15 of 32 subscales. It is important to note that this difference is not only quantitative, but also qualitative in nature. In other words, students who tend to avoid tests appear to have different characteristics than students who attend low-stakes tests, suggesting that artificially weighting up or down the scores of students who do attend testing sessions does not fully address the issue. Test avoiders are, essentially, a different kind of student—one that is little understood in the literature because researchers and practitioners have historically not addressed the fact that these test avoiders are not represented in their samples. Including test avoiders can help ensure that test results are representative of the *entire* population under study, not just those students who choose to show up for low-stakes tests.

Recommendation 2: Employ motivation-filtering techniques. Practitioners should implement non-intrusive motivation filtering techniques (e.g., Response Time Effort; Wise & Kong, 2005) to remove construct-irrelevant variance related to motivation on low-stakes tests.

Although the current study provides evidence for the importance of including the scores from test-avoiding students in an analysis that is generalized to an overall population, the low test-taking motivation exhibited by these test-avoiders is similarly a critical factor that one cannot ignore. Blindly including the scores from test avoiders in an analysis has the potential to introduce construct-irrelevant variance in the form of hurried or guessed responses, for which the students may not have even read the question.

Evidence in the current study points to a fascinating finding: test-avoiders who exert effort on a test once they are required to take it may actually score *higher* than non-avoiders who exhibit similar levels of motivation. By not identifying these students using motivation-filtering techniques and subsequently including their scores, a program runs the potential risk of biasing results *downward* (not upward, as one might expect). It is important to note, however, that these "triers" are only one segment of the test-avoiding population. In the current study, cognitive test scores for test-avoiders who did not exert effort (non-triers) were dramatically lower than the scores from test-avoiders who did exert effort (triers). This finding is not surprising, given that the non-triers were identified as such because they could not have spent enough time to simply read the item stems (much less thoughtfully consider the response options). However, in general, it is difficult to say what influence these non-triers' scores would have on the overall results of an assessment because these scores do not represent the students' true level on the construct of interest. Through motivation filtering techniques, practitioners and researchers are able to identify only those students who exert effort on the low-stakes tests of interest, thereby increasing the likelihood that aggregated test results are reflective of students' actual levels on the construct of interest.

Recommendation 3: Create targeted programming. At a minimum, we suggest implementing programming that stresses the importance of the test results, and that focuses on increasing the relevance of these tests for students. As noted previously, this study found that test avoiders may be qualitatively different than non-avoiders, and at least some of this difference appears to stem from the students' motivational profiles. For example, test avoiders are more likely to be work avoidant and less likely to have a performance-approach goal orientation (i.e., they are less likely to engage in academic exercises to demonstrate proficiency). Policy makers, universities, and testing programs may wish to create interventions that focus on the characteristics of the differing profile of the test-avoidant student in an effort to increase participation in low-stakes assessments. Programming, perhaps in the form of a marketing campaign, could advertise the relevance of an assessment to the test-avoider, thereby countering the student's inclination to exercise his or her work avoidant characterization. Similarly, a marketing campaign could appeal to the test-avoider's less performance-approach goal orientation by stressing that the student should not seek to excel on these tests to demonstrate proficiency, but for another reason such as increasing the quality of his or her diploma, or serving the improvement of public education by helping with item pilot testing. Regardless of the specific programming choice, efforts should be designed to directly address the unique non-cognitive profile (e.g., performance approach goal orientation, tendency toward work avoidance) of the test avoider.

Recommendation 4: Attach stakes. Our fourth of four recommendations is made with reservation, but we make it nonetheless given the evidence collected from the current study about the non-cognitive characteristics of many test avoiders. We suggest that practitioners and researchers consider increasing the stakes for tests that have traditionally been administered with

low or no stakes attached to them. This increase in stakes could include placing scores on transcripts or forcing students to retake tests if they exert low effort. This change appeals to the personality profiles of the cognitive non-triers whose effort does not appear to be influenced by external factors such as taking low-stakes tests simply because they are asked to do so. The obvious implication of such an action is to make what was once low-stakes now high-stakes, a move that carriers with it additional implications such as the legal ramifications of the stakes associated with the test as well as the construct-irrelevant variance introduced by test-anxious examinees.

As the name implies, a "low-stakes test" has limited or no repercussions for the test taker. However, today most low-stakes tests in the K-12 and college environments have serious implications for educational programs, and for the testing companies and institutions that administer the tests and make decisions based on the results. Stakeholders, such as school administrators, parents, legislators, and grant-sponsoring organizations, rely on the validity of the inferences made from low-stakes test scores, despite various student populations who may or may not show up to contribute data to these inferences.

Identifying and describing the students who tend to avoid these important assessments of student learning and development is among the first steps toward understanding the impact these students have on low-stakes tests. Similarly, assessment specialists who work in low-stakes contexts must implement non-intrusive motivation filtering techniques to identify those students who exert effort on tests and those who do not. Assessment specialists must recognize the importance of including low-stakes test results in analyses from those students who are prone to avoiding tests, but who nonetheless exert effort on low-stakes tests once they end up taking them.

Appendix A

Descriptions of Scales Used in Analysis

Attitudes Toward Learning (ATL): Measures students' knowledge and beliefs related to learning. Scale consists of seven separate scales:

Subscale Name	Sample Item	Scale Range
(1) Mastery-approach goal orientation	"I want to learn as much as possible this semester"	1 - 7 ("Not at all true of me" to "Very true of me")
(2) Mastery-avoidance goal orientation	"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")
(3) Performance-approach goal orientation	"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")
(4) Performance-avoidance goal orientation	"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")
(5) Work-avoidance goal orientation	"I want to do as little work as possible this semester"	1 - 7 ("Not at all true of me" to "Very true of me")
(6) Dweck's Theories of Intelligence	"You have a certain amount of intelligence, and you can't really do much to change it"	1 - 6 ("Strongly Agree" to "Strongly Disagree")
(7) Metacognitive Awareness Inventory-Regulation of Cognition Subscale	"I find myself pausing regularly to check my comprehension"	1 - 5 ("Always False" to "Always True")

Learning Environment Questionnaire (LEQ): Measures students' academic motivations and beliefs about learning under various environmental conditions in college.

Subscale Name	Sample Item	Scale Range
(1) Perceived autonomy	"I feel that my instructors provide me choices and options"	1 - 7 ("Almost never" to "Almost always")
(2) Perceived competence	"I feel confident in my ability to learn the material in my courses"	1 - 7 ("Not at all true of me" to "Very true of me")
(3) Interest in academics	"I think what I've learned in my courses is important"	1 - 7 ("Almost never" to "Almost always")
(4) Enjoyment of academics	"I really enjoy the courses I've taken"	1 - 7 ("Almost never" to "Almost always")

The next seven subscales compose an instrument known as the "Academic Motivation Scale", which measures students' reasons for attending college. All begin with the stem "Why are you going to college?"

(5) Intrinsic motivation to know (IMTK)	"Because I experience pleasure and satisfaction while learning new thing."	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(6) Intrinsic motivation toward accomplishment (IMTA)	"For the pleasure I experience while surpassing myself in studies"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(7) Intrinsic motivation to experience stimulation (IMES)	"For the intense feelings I experience when I am communicating my own ideas to others"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(8) Extrinsic motivation identified (EMID)	"Because eventually it will enable me to enter the job market in a field that I like"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(9) Extrinsic motivation introjected (EMIN)	"To prove to myself that I am capable of completing my college degree"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(10) Extrinsic motivation external regulation (EMER)	"In order to obtain a more prestigious job later on"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")
(11) Amotivation (AM)	"I once had good reasons for going to college; however, now I wonder whether I should"	1 - 7 ("Does not correspond at all [to me]" to "Corresponds exactly [to me]")

Fine Arts: Measures students' understanding of the fine arts, humanities, and literature.		
Subscale Name	Sample Item	Scale Range
(The Fine Arts Test is a single scale and does not include subscales)	"In most ancient civilizations, most monumental architecture, ceremonies, and art primarily served to A. provide the focus of a complex economy depending on the skills of many artisans. B. enhance the buildings in which humans resided and worshipped. C. celebrate the power and place of god(s) in shaping human life and society. D. reveal the creative and innovative spirit of the human intellect in facing a frightening world."	All items are dichotomously scores as correct or incorrect. Total scores can range from 0 points to 108 points.

Scale of Perceived Social Self-Efficacy (PSSE): Measures students' perceived abilities to engage in the social interactional tasks necessary to function in a college environment.		
Subscale Name	Sample Item	Scale Range
(The PSSE is a single scale and does not include	Students are asked to indicate using a Likert scale how much confidence they have as a student to engage in various activities, such as:	1 - 5 ("No confidence at all" to "Complete confidence")

subscales)	"Start a conversation with someone you don't know very well."	

Student Worry Questionnaire (SWQ): Measures both the process and content of worry as it is experienced by the traditional college student population.

Subscale Name	Sample Item	Scale Range
(1) Worrisome thinking	"I worry a lot about many daily life events and situations."	1 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(2) Significant others' well being	"I worry that a close family member might become seriously ill or injured."	2 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(3) Social-adequacy concerns	"I worry about saying the right things when expressing my opinion in discussions with other people."	3 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(4) Financial-related concerns	"I worry about not having enough money for the basic necessities of life (for example, clothing, food, rent)."	4 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(5) Academic concerns	"I worry about keeping up with or handling my academic workload."	5 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")
(6) General anxiety symptoms	"I feel physically tired and exhausted when I worry about things."	6 - 5 ("Almost never characteristic of me" to "Almost always characteristic of me")

Interacting with Others (MGUDS): Measures students' appreciation for, and comfort with diversity.

Subscale Name	Sample Item	Scale Range
(1) Relativistic appreciation	"It is very important that a friend agrees with me on most issues."	1 - 6 ("Strongly disagree" to "Strongly agree")
(2) Diversity of contact	"I would like to join an organization that emphasizes getting to know people from different countries."	1 - 6 ("Strongly disagree" to "Strongly agree")
(3) Comfort with differences	"I am only at ease with people of my own race."	1 - 6 ("Strongly disagree" to "Strongly agree")

Understanding our World (UOW): Measures students' quantitative and scientific reasoning skills.

Subscale Name	Sample Item	Scale Range
(Although one can score the UOW using subscales, this analysis only uses the total	"Which of the following is not an essential feature of experimental design? a. manipulation of the independent variable b. at least one control group or comparison c. use of modern technology	All items are dichotomously scores as correct or incorrect. Total scores can range from 0 points to 42 points.

score) d. measurement of the dependent variable"

Student Opinion Survey (SOS): Measures the effort a student exerts on a test (or tests) and the importance he or she places on the test(s).		
Subscale Name	Sample Item	Scale Range
(1) Effort	"I gave my best effort on this test."	1 - 5 ("Strongly disagree" to "Strongly agree")
(2) Importance	"This was an important test to me."	1 - 5 ("Strongly disagree" to "Strongly agree")

Appendix B

Comparison of All Test-Avoidance Students to the Attenders Group

	Attenders (All Students)			Avoiders (All Students)			
	Mean	*SD*	*N*	Mean	*SD*	*N*	*d*
% Male	35.64%			46.31%			
% Female	64.36%			53.69%			
GPA	3.02	0.53	303	2.80	0.60	488	0.39*
Highest Combined SAT Score	1165.05	112.82	291	1162.81	110.26	474	0.02
Number of Earned Credits	53.10	7.28	303	55.27	8.21	488	-0.28*
Age (in years)	20.06	0.67	303	20.42	1.14	488	-0.36*
ATL Subscale Scores:							
Mastery-Approach	5.80	0.95	303	5.46	1.11	488	0.32*
Mastery-Avoidance	4.57	1.01	303	4.48	1.03	488	0.09
Performance-Approach	5.36	1.38	303	4.97	1.53	488	0.27*
Performance-Avoidance	4.74	1.42	303	4.62	1.35	488	0.09
Work-Avoidance	2.92	1.24	303	3.37	1.27	488	-0.35*
Theories of Intelligence	3.46	0.85	303	3.49	0.85	488	-0.03
Metacognitive Awareness - Regulation	3.66	0.46	303	3.59	0.51	488	0.15
LEQ Subscale Scores:							
Autonomy	5.23	0.87	303	5.03	0.99	488	0.21*
Competence	5.87	0.87	303	5.57	1.03	488	0.32*
Interest	2.61	0.60	303	2.59	0.71	488	0.03
Enjoyment	3.74	0.71	303	3.64	0.82	488	0.13
Intrinsic Motivation to Know (IMTK)	5.33	1.13	303	5.32	1.17	488	0.02
Intrinsic Motivation Toward Accomplishment (IMTA)	4.74	1.41	303	4.67	1.41	488	0.05
Intrinsic Motivation to Experience Stimulation (IMES)	3.79	1.52	303	4.14	1.57	488	-0.23*
Extrinsic Motivation Identified (EMID)	6.02	0.89	303	5.82	1.08	488	0.20
Extrinsic Motivation Introjected (EMIN)	5.10	1.40	303	5.02	1.42	488	0.05
Extrinsic Motivation External Regulation (EMER)	5.50	1.17	303	5.60	1.19	488	-0.08
Amotivation (AM)	1.44	0.92	303	2.05	1.48	488	-0.48*
FineArts % Correct Score	57%	9%	303	50%	10%	488	0.74**
SWQ Subscale Scores:							

Worrisome Thinking	3.01	1.02	303	3.05	1.05	488	-0.04
Significant Others' Well-Being Concern	2.94	1.09	303	3.04	1.10	488	-0.09
Social Adequacy Concern	2.96	0.94	303	2.97	0.96	488	-0.01
Financial-Related Concern	2.63	1.05	303	2.90	1.08	488	-0.25*
Academic Concern	3.60	0.87	303	3.47	0.88	488	0.15
Generalized Anxiety Disorder Symptoms	3.11	0.94	303	3.12	0.97	488	-0.01
UOW % Correct Score	58%	15%	137	45%	18%	484	0.77**
SOS Subscale Scores							
Effort	3.16	0.74	303	2.84	0.75	488	0.42*
Importance	2.70	0.76	303	2.50	0.82	488	0.25*

* “Small” Cohen’s *d* (between .2 and .5)

** “Medium” Cohen’s *d* (between .5 and .8)

Appendix C

Students Who Exhibited Effort on All Tests in the Test-Avoidance Group Compared to Students Who Exhibited Effort on All Tests in the Attenders Group

	Attenders (Total Triers Only)			**Avoiders (Total Triers Only)**			
	Mean	*SD*	*N*	Mean	*SD*	*N*	*d*
% Male	35.25%			43.40%			
% Female	64.75%			56.60%			
GPA	3.06	0.59	122	2.80	0.77	53	0.40*
Highest Combined SAT Score	1170.00	119.88	115	1175.53	98.95	47	-0.05
Number of Earned Credits	53.07	6.99	122	56.04	8.13	53	0.40*
Age (in years)	20.10	0.90	122	20.56	1.44	53	-0.42*
ATL Subscale Scores:							
Mastery-Approach	5.81	0.87	122	5.84	0.79	53	-0.04
Mastery-Avoidance	4.44	0.94	122	4.68	0.80	53	-0.26*
Performance-Approach	5.43	1.24	122	4.99	1.63	53	0.33*
Performance-Avoidance	4.64	1.39	122	4.42	1.31	53	0.16
Work-Avoidance	2.76	1.18	122	3.05	1.15	53	-0.25*
Theories of Intelligence	3.48	0.84	122	3.58	0.70	53	-0.13
Metacognitive Awareness - Regulation	3.66	0.48	122	3.62	0.40	53	0.08
LEQ Subscale Scores:							
Autonomy	5.29	0.78	122	5.33	0.80	53	-0.05
Competence	5.98	0.74	122	5.89	0.95	53	0.11
Interest	2.54	0.51	122	2.63	0.45	53	-0.18
Enjoyment	3.77	0.65	122	3.78	0.60	53	-0.01
Intrinsic Motivation to Know (IMTK)	5.29	1.12	122	5.42	1.02	53	-0.12
Intrinsic Motivation Toward Accomplishment (IMTA)	4.66	1.35	122	4.40	1.30	53	0.19
Intrinsic Motivation to Experience Stimulation (IMES)	3.67	1.43	122	3.61	1.50	53	0.05
Extrinsic Motivation Identified (EMID)	6.00	0.80	122	5.88	1.07	53	0.14
Extrinsic Motivation Introjected (EMIN)	4.99	1.36	122	4.61	1.46	53	0.27*
Extrinsic Motivation External Regulation (EMER)	5.35	1.24	122	5.33	1.38	53	0.02
Amotivation (AM)	1.29	0.51	122	1.28	0.58	53	0.01
FineArts % Correct Score	60%	8%	122	60%	7%	53	-0.08

SWQ Subscale Scores:							
Worrisome Thinking	2.98	1.03	122	2.94	1.02	53	0.04
Significant Others' Well-Being Concern	2.92	1.09	122	2.80	1.07	53	0.10
Social Adequacy Concern	2.99	0.94	122	2.90	0.84	53	0.09
Financial-Related Concern	2.40	0.92	122	2.53	1.07	53	-0.13
Academic Concern	3.58	0.87	122	3.46	0.76	53	0.15
Generalized Anxiety Disorder Symptoms	3.13	0.92	122	2.95	0.80	53	0.20*
UOW % Correct Score	65%	11%	45	69%	11%	53	-0.32*
SOS Subscale Scores							
Effort	3.65	0.44	122	3.71	0.60	53	-0.12
Importance	2.99	0.69	122	3.02	0.86	53	-0.04

* “Small” Cohen’s *d* (between .2 and .5)

Appendix D

Students Who Exhibited Effort on All Tests in the Test-Avoidance Group Compared to Students in the Test-Avoidance Group Who Did Not Exhibit Effort on the Cognitive Tests

	Avoiders (Total Triers)			**Avoiders (Cognitive Non-Triers)**			
	Mean	*SD*	*N*	Mean	*SD*	*N*	*d*
% Male	43.40%			50.75%			
% Female	56.60%			49.25%			
GPA	2.80	0.77	53	2.80	0.55	266	0.01
Highest Combined SAT Score	1175.53	98.95	53	1152.10	114.91	266	0.21*
Number of Earned Credits	56.04	8.13	53	55.12	8.42	266	0.11
Age (in years)	20.56	1.44	53	20.40	1.23	266	0.12
ATL Subscale Scores:							
Mastery-Approach	5.84	0.79	53	5.43	1.00	122	0.44*
Mastery-Avoidance	4.68	0.80	53	4.33	1.01	122	0.36*
Performance-Approach	4.99	1.63	53	4.98	1.45	122	0.01
Performance-Avoidance	4.42	1.31	53	4.63	1.43	122	-0.15
Work-Avoidance	3.05	1.15	53	3.43	1.28	122	-0.31*
Theories of Intelligence	3.58	0.70	53	3.48	0.88	122	0.12
Metacognitive Awareness - Regulation	3.62	0.40	53	3.62	0.45	122	0.01
LEQ Subscale Scores:							
Autonomy	5.33	0.80	53	4.78	1.06	103	0.57**
Competence	5.89	0.95	53	5.48	1.08	103	0.40*
Interest	2.63	0.45	53	2.33	0.63	103	0.52**
Enjoyment	3.78	0.60	53	3.35	0.71	103	0.64**
Intrinsic Motivation to Know (IMTK)	5.42	1.02	53	4.94	1.24	103	0.41*
Intrinsic Motivation Toward Accomplishment (IMTA)	4.40	1.30	53	4.16	1.51	103	0.17
Intrinsic Motivation to Experience Stimulation (IMES)	3.61	1.50	53	3.66	1.57	103	-0.03
Extrinsic Motivation Identified (EMID)	5.88	1.07	53	5.86	0.89	103	0.02
Extrinsic Motivation Introjected (EMIN)	4.61	1.46	53	4.89	1.42	103	-0.19
Extrinsic Motivation External Regulation (EMER)	5.33	1.38	53	5.85	0.95	103	-0.47*
Amotivation (AM)	1.28	0.58	53	1.76	1.05	103	-0.52**
Fine Arts % Correct	60%	7%	53	44%	8%	266	2.00***

Score							
SWQ Subscale Scores:							
Worrisome Thinking	2.94	1.02	53	2.85	1.00	87	0.09
Significant Others' Well-Being Concern	2.80	1.07	53	2.86	1.04	87	-0.05
Social Adequacy Concern	2.90	0.84	53	2.73	0.85	87	0.20*
Financial-Related Concern	2.53	1.07	53	2.75	1.07	87	-0.20*
Academic Concern	3.46	0.76	53	3.36	0.92	87	0.11
Generalized Anxiety Disorder Symptoms	2.95	0.80	53	3.00	0.89	87	-0.06
UOW % Correct Score	69%	11%	53	35%	12%	266	2.84***
SOS Subscale Scores							
Effort	3.71	0.60	53	2.71	0.70	266	1.46***
Importance	3.02	0.86	53	2.44	0.80	266	0.71**

* “Small” Cohen’s *d* (between .2 and .5)

** “Medium” Cohen’s *d* (between .5 and .8)

*** “Large” Cohen’s *d* (greater than .8)

References

Chromy, J. R. (2005). *Participation standards for 12th grade NAEP*. Washington, D.C.: National Assessment Governing Board. Retrieved March 6, 2007, from http://www.nagb.org/release/chromy_paper_revised.doc

Kong, X. J., Bhola, D. S., & Wise, S. L. (2005). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education,* Montreal, Quebec, Canada.

National Commission on 12th Grade Assessment and Reporting. (2004). *12th grade student achievement in America: A new vision for NAEP*. Washington, D.C.: National Assessment Governing Board. Retrieved March 6, 2007, from http://www.nagb.org/release/12_gr_commission_rpt.pdf

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.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.