Does it Matter How Data are Collected? A Comparison of Testing Conditions and the Implications for Validity

Carol L. Barry and Sara J. Finney
James Madison University

Abstract

The effects of gathering test scores under low-stakes conditions has been a prominent domain of research in the assessment and testing literature. One important area within this larger domain concerns the implications of a test being low-stakes on test evaluation and development. The current study examined one variable, the testing context, that could impact students' responses during low-stakes testing, and subsequently the decisions made when using the data for test refinement. Specifically, the factor-structure of college self-efficacy scores was examined across three low-stakes testing contexts, and results indicated differential model-data fit across conditions (the very controlled context yielded the best model-data fit), implying that testing conditions should be seriously considered when gathering low-stakes data used for instrument development.

Introduction

As the emphasis on accountability in education has increased, so has the need for a clear understanding of the validity of the inferences made from examinee scores. This need is more imperative when one considers that many times there are no consequences of poor performance or low effort for the examinee. In fact, oftentimes the measures given in order to make high-stakes decisions about program effectiveness have relatively little personal meaning or importance to the students completing them. Situations in which there are little to no consequences to the test-taker are termed "low-stakes." This paper focuses on the implications of low-stakes testing on the validity of inferences made from test scores when those scores are used for instrument development.

Low-Stakes Testing and Examinee Motivation

There is a well-documented link between low-stakes testing environments and examinee motivation. Because there are very few, if any, consequences associated with performance and because students may perceive no personal gain from the experience, low-stakes testing often leads to low effort and motivation on the part of the test-taker (Wise & DeMars, 2005). Students may feel that there is nothing in it for them and may not be motivated to perform their best. Thus, their scores may not serve as valid indicators of their true level of the construct of interest (Sundre, 1999; Sundre & Kitsantas, 2004; Wise & DeMars, 2005). Essentially, this decrease in student motivation results in an increase in construct-irrelevant variance, with further implications on the psychometric functioning of the test items.

Uses of Low-Stakes Data and Threats to Validity

One of the main uses of data gathered in low-stakes environments is in evaluating program effectiveness for accountability purposes. Assessment practitioners may gather data to gauge whether or not a certain program delivered its intended effects. However, if low motivation results in test scores that are not truly representative of the construct of interest, the scores are then ambiguous at best and misleading at worst (Wise & DeMars, 2005). Thus, much of the research focused on low-stakes testing and motivation has emphasized either filtering out examinees with low motivation (e.g., Sundre & Moore, 2002; Wise & Kong, 2005; Wise, Wise, & Bhola, 2006) or attempting to increase examinee motivation (e.g., Wise, Bhola, & Yang, 2006) as ways to handle this construct-irrelevant variance.

Although it is often noted that one must exercise caution when making decisions about *program effectiveness* based on data from tests that are low- or no-stakes to the student, few studies note that caution must also be exercised when making decisions about the *test itself.* That is, the more fundamental and pervasive use of data collected in low-stakes environments is for instrument development purposes. Often, tests created to be used in high stakes conditions are evaluated and modified using data from low-stakes conditions (e.g., pilot testing; DeMars, 2000). Specifically, assessment and measurement professionals seem comfortable collecting data in low-stakes environments (e.g., through large-scale testing programs

or university participant pools) and using these data to examine the psychometric functioning of the items in order to inform instrument development decisions. If students do not provide valid responses, test developers may make unnecessary changes (or *not* make necessary changes) to an instrument. The need for sound instrument development practices is made more imperative when one realizes that sound assessment practice begins with appropriate and well-functioning measures.

Although sparse, there is some research that has studied the impact of low-stakes conditions and, consequently, low motivation on the psychometric properties of test items. Most of this research has examined the psychometric functioning of dichotomously scored achievement test items. One study approached this by examining item-by-item differences in performance by two groups of students that differed in the stakes associated with the test (Wolf, Smith, & Birnbaum, 1995). These researchers found that mentally taxing items exhibited differential item functioning; when matched on ability, the group of students for whom the test was low-stakes performed worse than those for whom the test was high-stakes. Similarly, student performance has been shown to be lower in the low-stakes condition of pilot testing than in a high-stakes testing condition, which may lead to poor instrument refinement decisions if the item difficulties estimated under pilot conditions are thought to represent item difficulties under operational conditions (DeMars, 2000). An additional study focusing on the problem of low motivation found that the inclusion of examinees who demonstrated rapid-guessing (i.e., examinees with extremely low motivation) affected the estimation of item parameters (Wise & DeMars, 2006). Specifically, items that were known to have low item difficulties appeared more difficult and more discriminating when rapid-guessers were included in the sample.

Despite the research conducted on the effects of low-stakes and low motivation on the psychometric properties of achievement tests, one area of research that is lacking involves the effects on the psychometric properties of non-cognitive or developmental tests. The items on these instruments are typically polytomous or continuous in nature, and their psychometric properties are generally studied through the use of factor analysis. Interestingly, there appears to be very little, if any, research conducted on whether and how low-stakes environments impact the factor-structure of developmental measures. This is somewhat surprising given that student attitudes/affect are often of interest to student affairs personnel and that assessment specialists are often concerned with both learning and developmental outcomes. It seems reasonable to believe that, similar to achievement tests, the psychometric properties of developmental instruments would also be impacted by the decreased student motivation that accompanies low-stakes testing.

Purpose of the Current Research

Because low-stakes testing environments are unavoidable for many who study the properties of tests, and thus there may be inconsistency in the stakes associated with data gathered for test development versus data gathered for decision making (DeMars, 2000), it is important to determine the best way to collect useful and valid data that are of no- or low-stakes to the participants. Specifically, we were interested in examining if changes in the testing context would impact student responses to low-stakes tests. Would a more controlled testing context improve the quality of low-stakes data? To answer this question, we examined the factor structure of college self-efficacy scores from multiple samples gathered in several different testing contexts. Our main focus was the effect of testing context on model-data fit. That is, did the same factor structure emerge under the different contexts, or were the relationships between items different across context, resulting in different psychometric properties associated with the measure. The theoretical underpinnings of the models tested are discussed at length in another paper (Barry & Finney, 2007) and will not be the elaborated upon here; rather, the focus of this paper will simply be on comparing model-data fit across testing contexts.

We believe this study helps answer the call of Birenbaum (2007) to evaluate the validity of the *full* testing program. Specifically, Birenbaum emphasized the need for entrenching the comprehensive assessment process within an overarching validity framework. That is, one should not focus solely on the validity of inferences made based on scores, but rather should consider these inferences within the wider frame of how the assessment instruments map to the domain of study, the psychometric functioning and internal structure of the instruments (e.g., factor structure), and the *contexts* in which the data were collected (Birenbaum, 2007). In other words, the entire assessment process needs to be evaluated with respect to validity. Again, this study focuses on the impact of context on examinee test-taking behavior when tests are of no-stakes to students.

Methods

Participants and Procedures

Five samples of data were collected across a variety of testing conditions at a mid-sized, mid-Atlantic university. In all conditions, student responses were of no stakes to the individual student, and students were not provided with any information regarding their scores. Each sample and the method by which data were gathered are described below (see also Table 1 for a description of all samples).

Table 1
Description of Samples

Sample	Acronym	Age	Testing Condition	Item Order
Uncontrolled Freshman-1	UnFr-1	Freshmen	Students completed an online version of the instrument on their own time, unsupervised, prior to arriving on campus for the start of the Fall 2006 semester.	Non-Randomized
Uncontrolled Freshman-2	UnFr-2	Freshmen	same as above	Non-Randomized
Very Controlled Upperclassman	VCUp	Sophomores, Juniors, Seniors	Students completed the instrument in a small (~20 seats) classroom in the presence of a trained proctor. Care was taken to slow response time.	Non-Randomized
Controlled Upperclassman	CUp	Sophomores and Juniors	Students completed the instrument in large (i.e., number of seats ranged from 63-250), lecture-style classrooms, in the presence of trained proctors.	Non-Randomized
Controlled Upperclassman-Randomized	CUp-R	Sophomores and Juniors	same as above	Randomized

Uncontrolled freshman samples. Data were collected from 3,562 freshman students who completed the college self-efficacy measure as part of an on-line survey designed by the university to gather information about the incoming class. These surveys were approximately 60 items in length, with the college self-efficacy measure administered last. These students completed the instrument on their own time, unsupervised, prior to arriving on campus for the start of the Fall 2006 semester. Given this, we considered this a very uncontrolled testing context. The total sample was randomly split for replication purposes, and after screening the data and removing any outlying cases, sample sizes were 1,586 and 1,585 for Samples 1 (Uncontrolled Freshman 1: UnFr-1) and 2 (Uncontrolled Freshman 2: UnFr-2), respectively.

Very controlled upperclassman sample. Sample 3 consisted of 237 university upperclassmen (i.e., 66% sophomores, 22% juniors, and 11% seniors) recruited from the psychology participant pool during the Fall 2006 and Spring 2007 semesters. These students completed the instrument along with several other motivation-related measures in a small (~20 seats) classroom setting. The instruments were administered

to the students by handing out a manila envelope containing all measures. It took approximately 40 minutes to complete the battery of instruments, and the college self-efficacy measure was administered first in all sessions. Participants completed the measures one at a time and were not allowed to begin responding to the next measure until everyone had completed the current measure. Each measure's instructions were read aloud by a trained proctor prior to student beginning the measure. This process was employed as an attempt to slow response rates, in the hopes that it would produce more thoughtful responses. Thus, Sample 3 (Very Controlled Upperclassman: VCUp) completed the college self-efficacy instrument in a highly controlled context.

Controlled upperclassman and controlled upperclassman-randomized samples. Data for samples 4 and 5 were collected from a total of 854 upperclassman students. These participants completed the college self-efficacy measure during a mandatory university-wide assessment day during the Spring 2007 semester. The data collected during the assessment day were used for program effectiveness initiatives on campus. That is, the data were high-stakes for the administrators of programs on campus but of no stakes to the students completing the measures. Students completed a three-hour battery of tests in large (i.e., number of seats ranged from 63-250), lecture-style classrooms with proctors. The order of the tests differed across rooms, but the college self-efficacy measure tended to be administered during the last third of the testing session. We deemed this a slightly controlled testing context because, although there were proctors present, students were allowed to attend to the test as much or as little as they wanted. The combination of the larger room and the decreased proctor attention resulted in a higher degree of anonymity for the students and a potential for decreased motivation. After removing outliers and cases with missing data, Sample 4 (Controlled Upperclassman: CUp) consisted of 397 students and Sample 5 (Controlled Upperclassman-Randomized: CUp-R) consisted of 449 students. Sample 5 completed a version of the college self-efficacy instrument in which the order of the items was completely randomized.

Measures

College Self-Efficacy Inventory. The College Self-Efficacy Inventory (CSEI: Solberg, O'Brien, Villarreal, Kennel, & Davis, 1993) was used to assess college self-efficacy and consists of 20 items written to represent participants' beliefs in their capabilities to successfully complete college-related tasks. Participants were asked to respond by indicating how confident they are in their ability to complete the task [1 (not at all confident) to 10 (extremely confident)]. The instrument, with its original and randomized item order, is presented in the Appendix.

Although the CSEI was administered for program evaluation purposes, a second, and equally important, purpose for its administration was to examine its psychometric properties. There had been little previous research on the properties of the instrument, and what existing research there was led us to believe that additional work on the measure may be necessary before trusting the inferences we made from its scores. It was important to collect data to evaluate its properties in the same context it would be gathered when used for program assessment: no-stakes. Moreover, it is difficult to imagine a situation in which students would complete this type of measure in a high-stakes environment. Therefore, we believe the contexts used in this study have high external validity.

Results

Confirmatory Factory Analyses

Confirmatory factor analysis (CFA) was used to test four models. All CFAs were conducted using LISREL 8.72 (Jöreskog & Sörbom, 2005). Because data screening indicated that the data for all samples were multivariate nonnormal, the Satorra-Bentler (S-B) correction was used in conjunction with maximum likelihood estimation to produce a corrected χ^2 and corrected standard errors. Global model-data fit was evaluated using the χ^2, along with the standardized root mean square residual (SRMR, with values of .08 or less indicating good model-data fit), the S-B adjusted root mean square error of approximation (RMSEA, with values of .07 or less indicating good model data fit), and the S-B adjusted comparative fit index (CFI, with values of .95 or above indicating good model-data fit). Areas of local misfit were identified by examining the standardized covariance residuals, which describe how well a model is able to reproduce each pair-wise relationship among items. These values can be positive or negative, indicating under- or over-representation of relationships, and absolute values of three or greater have been suggested as values to indicate a poorly reproduced relationship (Raykov & Marcoulides, 2000).

We were interested in examining whether model-data fit differed across the testing conditions. Although we expected there to be model-data misfit for all testing contexts given previous study of the measure, we expected greater overall misfit in the less controlled contexts compared to the controlled context. Given model-data misfit, we were then interested in examining how localized areas of misfit replicated across the testing conditions. Specifically, we questioned whether the more controlled condition would have fewer but *similar* areas of misfit than the other conditions or whether the more controlled condition would have fewer and *different* areas of misfit than the other conditions. Because specific areas of misfit often guide scale modification, ultimately we were interested in whether we would make different recommendations regarding scale modifications and refinement across the testing conditions.

Uncontrolled Freshman Samples 1 and 2

The theoretical model (Model 1) was fit to the data for the UnFr-1 sample and did not fit the data well (Table 2). Specific areas of misfit associated with this model were diagnosed by examining the standardized covariance residuals (Table 3). For Model 1, there were 41 standardized covariance residuals greater than three in absolute value, providing further evidence of model misfit. Theoretical and empirical considerations were used to derive and test a series of modified models through an iterative process until finding a model that fit the data adequately. Specifically, modifications were made to address areas of localized misfit, given that there was a theoretical or practical reason for doing so (e.g., redundancy in items, misalignment between item and subscale content). Three modified models were tested, with a 15-item three factor model providing adequate global model-data fit (Table 2).

Table 2
Fit Statistics for Hypothesized and Modified Models

	Model	$\chi^2_{S\text{-}B}$	df	SRMR	$RMSEA_{S\text{-}B}$	$CFI_{S\text{-}B}$
UnFr-1						
	1) 17-item, three-factor	2135.30	116	0.075	0.10	0.92
	2) 16-item three-factor a	1444.85	101	0.075	0.09	0.94
	3) 16-item three-factor b	1286.40	101	0.054	0.09	0.95
	4) 15-item three-factor	976.92	87	0.051	0.08	0.95
UnFr-2						
	1	1886.78	116	0.069	0.10	0.94
	2	1282.44	101	0.069	0.09	0.96
	3	1224.22	101	0.055	0.08	0.96
	4	820.45	87	0.048	0.07	0.97
VCUp						
	1	349.39	116	0.083	0.090	0.90
	2	223.05	101	0.079	0.070	0.94
	3	225.05	101	0.076	0.070	0.94
	4	182.78	87	0.070	0.070	0.95
CUp						
	1	575.68	116	0.079	0.09	0.95
	2	396.97	101	0.079	0.09	0.95
	3	368.79	101	0.068	0.08	0.96
	4	305.09	87	0.062	0.08	0.96
CUp-R						
	1	674.41	116	0.076	0.10	0.92
	2	522.85	101	0.078	0.10	0.93
	3	487.85	101	0.073	0.09	0.94
	4	451.72	87	0.069	0.10	0.93

Note. The 17-item, three-factor model is a model in which three of the original items were removed prior to analyses due to poor functioning found in prior studies and item content issues; the 16-item three-factor a model is the model in which item 5 was removed; the 16-item three-factor model b is the model in which item 5 was removed and item 1 was moved to the Roommate subscale; the 15-item three-factor model is the model in which item 1 was removed from the scale.

Table 3

Areas of Localized Misfit for Models

	Model 1		Model 2		Model 3		Model 4	
	# resids > | 3 |	item pairs w/resid > | 5 |	# resids > | 3 |	item pairs w/resid > | 5 |	# resids > | 3 |	item pairs w/resid > | 5 |	# resids > | 3 |	item pairs w/resid > | 5 |
UnFr-1	41	1 with 2, 15, 16, 20; 2 with 3, 4; 3 with 5, 6; 4 with 18, 19; **5 with 6**, 11; 6 with 11, 18; 9 with 17	33	**1 with 2**, 15, 16, 20; 2 with 3, 4; 4 with 18, 19; 6 with 17; 9 with 17; 11 with 18	29	1 with 3; **2 with 3**, 4; 4 with 18, 19; 6 with 17; 9 with 17; 11 with 18	22	**2 with 3**, 4; 4 with 18, 19; 6 with 17; 9 with 17; 11 with 18
UnFr-2	43	1 with 2, 15, 20; 2 with 3, 4; 3 with 4, 5, 6, 20; 4 with 18; **5 with** 6, **11**; 6 with 11, 18; 8 with 18	33	**1 with** 2, 3, 15, **20**; 2 with 3, 4; 3 with 4, 13, 20; 4 with 18; 8 with 18; 11 with 15	32	1 with 3, 6, 16, 20; **2 with 3**, 4; 3 with 4; 4 with 18; 8 with 18; 11 with 15	25	**2 with 3**, 4; 3 with 4, 20; 4 with 18; 8 with 18
VCUp	8	4 with 15; **5 with 6**	9	**4 with 15**	9	**4 with 15**	5	4 with 15
CUp	22	1 with 20; 2 with 3; **5 with 6**; 9 with 14	22	1 with 20; 2 with 3; **6 with 13**; 9 with 14	18	1 with 3; 2 with 3; **6 with 13**; 9 with 14	16	2 with 3; 6 with 13; 9 with 14; **11 with 16**
CUp-R	19	3 with 5; **5 with 6**; 6 with 18; 8 with 19	16	6 with 19; **8 with 19**	15	6 with 17, 19; **8 with 19**; 9 with 15	13	6 with 17, 19; **8 with 19**

Note. Largest residual indicated by bolded item pair.

Although modifications to the tested models resulted in improved fit for the UnFr-1 sample, there are several problems associated with re-specifying and testing modified models on the same sample (MacCallum, Roznowski, & Necowitz, 1992). Because the fit of the modified models may capitalize on chance (i.e., fitting the idiosyncrasies of the sample), the fit of modified models may not generalize to other samples. Given this, all models were tested again using the UnFr-2 sample to (a) determine whether the pattern of misfit associated with the four theoretical models was reproduced in an independent sample, (b) provide the first a priori testing of the modified models. As expected, results for UnFr-2 were extremely similar to UnFr-1, both in regard to global fit and areas of local misfit (Tables 2 and 3). This was not a surprise given that the two samples were derived by randomly splitting the overall sample and both were fairly large in size, which results in more stable estimates.

Despite the adequate global fit for the 15-item, three-factor model, several areas of local misfit remained for both samples, as evidenced by a number of large residuals. Especially puzzling were the large residuals associated with the relationships between items 2, 3, and 4. These three items represent different subscales and appear to represent completely different areas of confidence. One possible explanation lies in the fact that these items were presented in succession, and the strong relationships may have been caused by an item-ordering effect; especially when expressing attitudes, preceding questions can influence the responses given to subsequent ones (e.g., Schwarz, 1999; Tourangeau & Rasinksi, 1988). It is possible that these items were correlated with one another simply because they were located next to one another on the instrument.

Very Controlled Upperclassman Sample

The results from the UnFr-1 and -2 samples indicated that the three-factor model (Model 1) did not fit the data well and that, even after removing two items that consistently performed poorly across samples, a great deal of localized misfit remained. Again, the important point is that areas of misfit replicated across the two random samples from the uncontrolled condition, and if these were our only samples, we may claim there was no clear structure to the data and most likely recommend not using the measure for assessment purposes. We were now interested in evaluating if these same results would emerge for data collected in a controlled condition. Thus, data for the VCUp sample were gathered to address these concerns.

Similar to the UnFr-1 and -2 samples, the theoretical model did not fit the data (Table 2). Additionally, the *patterns* of local misfit for Model 1 were similar, although not identical, to those found using the Freshman samples. In order to fully compare the results across samples, the three modified models tested using the UnFr-1 and -2 samples were fit to data from the VCUp sample, and the reduced 15-item, three-factor model provided fairly good model-data fit. Moreover, it is quite interesting to note that the local misfit associated with items 2, 3, and 4 no longer was present, and overall, standardized covariance residuals were fewer in number and smaller in magnitude, with values between 0 and 1 for most items (Table 3).

Obviously, one possible explanation for the substantially better local fit concerns the method of administration. Unlike UnFr-1 and -2, students in the VCUp sample completed the instrument in a much more controlled testing context. It is very likely that the high number of large residuals were not present for this sample because these students provided more thoughtful answers to the questions and were not able to simply rush through the questionnaires. However, it is important to note that the age of the student in the controlled condition was different from that in the uncontrolled condition; students in the controlled condition were older and had more experience in college. Because there were two variables that changed between these samples (i.e., freshman vs. upperclassman and uncontrolled vs. controlled condition), it is not possible to disentangle which was the cause of the better model-data fit.

Controlled Upperclassman Samples

The CUp and CUp-R samples were used to collect data to address questions raised by the results from the previous three samples. Specifically, one question concerned why there were fewer areas of local misfit when using the VCUp sample compared to the UnFr- and -2 samples. As noted above, one possibility could be the method of administration (an uncontrolled condition vs. a very controlled setting with explicit instructions to answer slowly and carefully); however, it is possible that the year in school of the participants was the underlying factor contributing to these differences. The CUp sample (i.e., upperclassmen in a slightly controlled condition) was gathered to help disentangle these variables. We believed the reduction of misfit for the very controlled condition was due to the testing environment and not the age of the student. Therefore, we expected to find more misfit associated with the CUp sample (upperclassmen in slightly controlled condition) compared to VCUp (upperclassmen in a very controlled setting).

A second question that remained was why items 2, 3, and 4 in particular exhibited large residuals. We believed we were seeing an item-order effect (e.g., Schwarz, 1999; Tourangeau & Rasinksi, 1988) due to low motivation. Specifically, if students don't respond in a thoughtful manner, they may choose similar response options for items placed next to each other on the measure. The CUp-R sample was used to test this hypothesis. That is, if items 2, 3, and 4 were no longer positioned next to each other on the scale and the testing condition was slightly controlled, would the items still have large standardized residuals? We hypothesized that they would not; instead, items positioned next to each other in this new randomized order would have large residuals.

As found previously, the theoretical model did not fit the data for the CUp sample (Table 2). Examination of the standardized covariance residuals for the models (Table 3) indicated that, overall, patterns local misfit was similar to that found for the UnFr-1 and -2 samples. As expected, the number of standardized covariance residuals was higher than that found using the VCUp sample, and the specific misfit associated with items 2, 3, and 4 was again found, suggesting that its presence was a function of the testing condition (i.e., degree of control) rather than age.

The CUp-R sample was used to determine whether the misfit associated with items 2, 3, and 4 found using the UnFr-1 and -2 samples and the CUp sample was an item ordering effect caused by low motivation. Again, the theoretical model did not fit the data adequately (Table 2). Misfit associated with Model 1 and the three modified models was again examined (Table 3). Consistent with all samples com-

pleting the non-randomized version of the instrument, the largest residual for the CUp-R sample was that for the relationship between items 5 and 6. However, the pattern of residuals overall was *not* very similar to those found in any of the previous samples that administered the non-randomized version. Moreover, Model 4 (the 15-item, three-factor model) did *not* fit the data from this sample, which is understandable since the model was created based on misfit from the previous samples and since the CUp-R sample did not share the same areas of misfit. Of particular importance, there were no longer large residuals associated with items 2, 3, and 4, but there *was* misfit associated with items 6 and 19, which were located next to one another in the randomized version. This suggests that the misfit associated with items 2, 3, and 4 was indeed an item order effect caused by testing context. Essentially, these results highlight the fact that in low-motivation contexts there can be dependencies among items simply because they are located adjacent to one another on an instrument and that randomizing the order of the items will result in *different* sets of items displaying these dependencies. This of course will affect the psychometric properties of the measure (i.e., the factor structure) and subsequent decisions regarding test refinement.

Discussion

Given the risks associated with using low-stakes data and the widespread use of this type of data for instrument development purposes, this research was conducted to examine the dimensionality of college self-efficacy scores from multiple samples gathered in several different testing contexts in order to determine whether the amount of proctor control impacted the fit of the data to the tested models. Although some similarities were found across all samples and testing conditions (e.g., the theoretical model did not fit, there was an extremely large standardized covariance residual for item 5 and 6), there were differences in model-data fit across the three testing conditions. As the testing conditions increased in level of control, the amount of localized misfit decreased. That is, the largest numbers of standardized covariance residuals were found when using data collected in an uncontrolled testing condition (i.e., UnFr-1 and -2 samples), smaller numbers of residuals were found when using data collected in a controlled condition (i.e., CUp and CUp-R samples), and the smallest numbers of residuals were found when using data collected in a very controlled condition (i.e., VCUp). Thus, the measure could have been considered inadequate when employing the two Uncontrolled samples and the two Controlled samples, whereas it may have been considered acceptable when employing the Very Controlled sample. If item deletion was conducted in order to create a "better" measure, more items would be removed from the test using these Uncontrolled or Controlled samples than if conducting the same process using the Very Controlled sample. As items are labor-intensive to construct and, in turn expensive to write, throwing out quality items is something instrument developers and evaluators would like to avoid. Collecting data in a controlled setting appears to minimize the chance of removing quality items.

Thus, one possible way to alleviate these problems is to increase the level of control in the testing condition, as was done with the Very Controlled sample. Specifically, the participants in this sample completed the instrument in a small campus classroom with the experimenter present, were given explicit instructions to carefully answer the questions, and were not allowed to rush through the questionnaires. This was done to slow responding in the hopes that participants would provide more thoughtful responses to the questions. As mentioned previously, the residuals for the tested models were fewer in number and much smaller in magnitude for this sample than they were for the samples who participated in the large-scale testing.[2] Thus, it appears that the testing condition played a very important a role in how much effort participants put into their responses, how thoughtfully they responded, how well the models fit the data, and ultimately the proposed modifications to the measure. Slowing responding eliminated what appeared to be a sort of response style/acquiescence and eliminated some of the dependency of the items on one another.

One related and particularly concerning result of this study involves the dimensionality and pattern of residuals obtained for the CUp-R sample, which received the randomized form of the CSEI. Although randomizing the item order eliminated the residuals between items 2, 3, and 4, the overall patterns of misfit for the models were alarmingly dissimilar when fit to these data than when fit to data from samples who received the non-randomized form. Moreover, the modified models fit the data worse in this sample than any other. It is important to note that this was true when comparing the CUp sample to the CUp-R sample, which involved the same age students (i.e., upperclassmen) in the same testing condition (less controlled testing situation); the only aspect that differed was item order. It is very possible that all modifications made to the instrument in the original item order might not have been made using this randomized order in an uncontrolled setting and other modifications *would* have been made. However, the

results of this study do suggest that students attend to items to a higher degree when in a more controlled testing context, resulting in a clearer understanding of item functioning. It follows that the effects of randomizing the item order on model-data fit may not be so problematic if data are collected in a more controlled testing condition. Presumably, a more controlled testing condition and the subsequent decrease in error variance would allow areas that are truly problematic to be identified. We unfortunately did not have a sixth sample to test this hypothesis, and we call for additional work in this area.

Conclusion

The results from the current study have serious implications for the manner in which data for instrument development should be gathered. In an instrument development context, data are typically gathered through a large-scale testing program or a university participant pool (i.e., an environment that is extremely low-stakes to the test takers providing the data), several models are fit to that data, and changes to the instrument are made based on areas of misfit associated with the tested models. However, this study has shown that the amount of misfit present is dependent upon how controlled the testing condition is. Because of this, data collected from students in an uncontrolled testing condition might lead assessment specialists or test developers to make unnecessary changes (or fail to make necessary changes) to an instrument. On the other hand, a testing condition in which there is a high degree of control, although more costly in terms of time and resources, appears to increase student motivation despite the fact that the test is low-stakes to these students. Consequently, the test developer is more able to trust the validity of inferences made regarding these scores and will therefore make more appropriate decisions about changes to an instrument. This is important given that sound assessment practice begins with appropriate and well-functioning instruments, and before one can trust the inferences made regarding student performance or development and, ultimately, program effectiveness, one must be able to trust the instrument with which these are measured.

References

Barry, C. L., & Finney, S. J (2007, October). *A Psychometric Investigation of the College Self-Efficacy Inventory.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*, 29-49.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13,* 55-77.

Gore, P. A., Leuwerke, W. C., & Turley, S. E. (2006). A psychometric study of the College Self-Efficacy Inventory. *Journal of College Student Retention: Research Theory & Practice, 7,* 227-244.

Jöreskog, K. G., & Sörbom, D. (2005). LISREL (Version 8.72) [Computer software]. Lincolnwood, IL: Scientific Software.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490 – 504.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.

Solberg, V. S., O'Brien, K., Villareal, P., Kennel, R., & Davis, B. (1993). Self-efficacy and Hispanic college students: Validation of the College Self-Efficacy Instrument. *Hispanic Journal of Behavioral Sciences, 15,* 80-95.

Sundre, D. L (1999). *Does examinee motivation moderation the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

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

Sundre, D. L., & Moore, D. L., (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14,* 8-9.

Tourangeau, R., & Rasinksi, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Wise, S.L., Bhola, D.S., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25,* 21-30.

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

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

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

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11,* 65-83.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.

Footnotes

[1] Although the instrument consists of 20 items, three items were removed prior to testing these models. These three items had functioned poorly in past studies of the instrument (e.g., Barry & Finney, 2007; Gore, Leuwerke, & Turley, 2006; Solberg et al., 1993) and were written such that they may not be relevant to all students. Thus, all models tested in this paper were based on the remaining 17 items.

[2] One might question whether the differences in the number and magnitude of the standardized residuals were due to differences in sample sizes rather than differences in the level of control. This is because the standardized covariance residuals used to examine misfit are computed by dividing the covariance residuals by the standard error. Given that standard errors can be affected by the sample size (i.e., smaller samples tend to yield larger standard errors and, in turn, may lead to smaller standardized covariance residuals), it was possible that the large residuals in the UnFr-1 and -2 samples were due to their large N, that the moderate residuals in the CUp and CUp-R samples were due to their smaller N, and that the small residuals in the VCUp sample were due its small N. To ensure that this was not a plausible explanation for the pattern of results, all analyses were conducted a second time, using the correlation matrix (i.e., the standardized covariances) as input; when conducted in this manner, correlation residuals are computed, which are not impacted by the standard error and consequently the sample size. The results indicated that the correlation residuals followed a similar pattern and had similar relative magnitudes, providing evidence that the differences in the number and magnitude of standardized covariance residuals across the five