Validation of the Teacher's High Stakes Testing Survey

Lantry L. Brockmeier, PhD
Robert B. Green, PhD
James G. Archibald, PhD, LPC
James L. Pate, PhD
Donald W. Leech, EdD

Valdosta State University
1500 N. Patterson St.
Valdosta, GA 31698

Abstract

The purpose of this study was to examine the soundness of the psychometric characteristics of the Teacher's High Stakes Testing Survey. The 49-item instrument is comprised of six hypothesized subscales (i.e., curriculum, teaching, work satisfaction, stress, accountability, and students) measured with a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). An expert panel reviewed the instrument plus an exploratory factor analysis and confirmatory factor analyses were conducted. Expert panel members suggested only a few minor wording modifications to improve the instrument. The confirmatory factor analyses yielded data to support the fit of the model and measurement invariance of the model by gender and race or ethnicity.

Keywords: high stakes testing, accountability, validation, measurement invariance

1. Introduction

Standardized testing began in Massachusetts under the direction of Superintendent Horace Mann during the 1840s to assess student knowledge in several content areas (Resnick, 1982). Comparisons between schools and classrooms were made and the results of these examinations were published (Hamilton, 2003). Within the next 30 years, Tyack (1974) indicated that other states began administering tests and reporting their results in newspapers. Even student promotions that had been based on teacher recommendations became tied to performance on these tests (Engelhart, 1950).

According to Linn, Miller, and Gronlund (2005), the use of standardized testing in the United States did not expand significantly until after World War II. Congress, in an attempt to equalize educational opportunities, passed the Elementary and Secondary Education Act (ESEA) of 1965. Within the ESEA there was a requirement for monitoring of student progress that resulted in additional student testing. Airasian (1988) indicated that during the 1970s concern continued to grow about the quality of education. The minimum competency movement developed from these concerns and transferred some important decisions from individual teachers to increase standardization of content taught to students (Burton, 1978; Camilli, Cizek, & Lugg, 2001). In addition, minimum competency testing attempted to ensure that all students mastered the basic skills (Hamilton, 2003). According to Popham (1978), the minimum competency testing movement halted the devaluation of the high school diploma.

The concept of measurement-driven instruction evolved from the minimum competency testing movement (Hamilton, 2003). The prevailing thought was that testing could influence what was taught. With the release of *A Nation at Risk* (National Commission on Excellence in Education, 1983), there was a heightened concern over student and school performance. This led to increased testing and school-level incentives (Hamilton, 2003). The 1990s standards movement increased the awareness of the links between standards, curriculum, and testing. The links and formal stakes enhanced motivation to increase performance (Smith, O'Day, & Cohen, 1990). Lewis (2000) and Holland (2001) indicated that high stakes testing encourages students and educators to approach the teaching and learning process seriously.

Holland (2001) went on to state, "Without testing, standards are mere suggestions" (p. 2). Congress with the passage of The No Child left Behind (NCLB) Act of 2001 increased the pressure for educational reform. The NCLB Act consists of goals in the form of standards, tests or measures of performance, targets for performance, and consequences for a school's success or failure (Hamilton & Koretz, 2002).

Over the last 50 years of educational reform, the common thread was the increased use of high stakes testing for accountability purposes due to the concern for student, program, and school performance. Similar to the business leaders concerns about students' ability to read and write during the 1970s (Cizek, 2001), researchers today are indicating that students are exiting school without the knowledge and skills to survive in an increasingly competitive world (Wakefield, 2003; Haycock, 2005; Darling-Hammond, 2006).

America's obsession with high stakes standardized testing will not become an endangered species anytime soon (Kaback, 2006). Phelps (2005) indicated that poll and survey data has indicated the general publics' positive view of standardized testing. For example, the percentage point differential between positive responses and negative responses to standardized testing was a +90 for students being required to pass a graduation test. Driesler (2001) reported that 90% of parents wanted information that would allow the comparison about their children and schools. Policymakers, parents, and the general public continue to demand better school performance and view the results of high stakes testing as proof of learning (Wahlberg, 2003; Scherer, 2005). High stakes tests results are being used to demonstrate to taxpayers that their investment of dollars is used effectively to produce quality outcomes (Lederman & Burnstein, 2006). Afflerbach (2005) suggested three possible explanations for high stakes testing's popularity; fairness to all students since no students receives preferential treatment, scientific since the tests undergo examination for validity and reliability, and commonplace due to the frequency of administration. Another potential reason for high stakes testing's popularity is the ability to provide a numerical score that can be indexed to an alphabet that represents quality and achievement (Baines & Stanley, 2004).

Fremer (2005) and Linn, Miller, and Gronlund (2005) indicated that arguing against the use of high stakes tests results dismisses relevant information that might lead to better decision making. Standardized tests are essential to confirm grading systems that vary from teacher to teacher and from school to school (Holland, 2001). Grade point averages and course grades are too unreliable for use as outcome measures (Phelps, 2003). Evers (2001) stated that "a divergence between grades from classroom teachers and scores on standardized tests can be a wake-up call for parents, taxpayers, and school boards – telling us that students don't really know the subject matter and that teachers are too soft in their grading practices. Getting rid of standardized tests is like getting rid of thermometers, X-ray machines, and blood pressure gauges in a doctor's office" (p. 2).

Stone (2003) indicated that if one were to read the educational literature on high stakes testing, one would get the impression that high stakes testing has few advantages since so much of the recent literature is negative. Stone pointed out that during the 20th century that teachers and schools at the local level routinely used standardized tests for documentation of student, teacher, and school performance. Stone elaborated that it was not until policymakers began to hold schools accountable for test results that the limitations became fatal flaws.

Teachers were very supportive of high stakes standardized testing in the 1970s and 1980s when the stakes were only for students (Phelps, 2005). While still supportive of standards, testing, and accountability, teachers support has declined with the implementation of the NCLB Act of 2001. Teachers are under ever increasing pressure to increase performance of their students. Is their an incongruence between what teachers believe is their instructional role in the teaching and learning environment and what high stakes testing requires of teachers? Are high stakes tests a valid measure of teaching ability considering the impact of prior student achievement and a host of other student, family, and community factors that impact student performance on a high stakes test (Meyer, 2000; Linn, 2006)? Furthermore, potential consequences for teachers include a negative evaluation, removal, reassignment, and a decrease in financial compensation. How are these potential consequences impacting teachers? Research into the impact of high stakes testing on teachers is important to continue (Vogler, 2002).

1.1 Instrument History

Hope, Brockmeier, Lutfi, and Sermon (2006) developed the *Teacher's High Stakes Testing Survey* to obtain information from teachers about their attitudes towards high stakes testing. The authors identified that items selected for the instrument were based upon a review of the literature and that items represented both positive and negative attributes of high stakes testing.

Furthermore, Hope et al. identified that the developmental process included the identification of specific domains for the construct of interest, item building, and content validation of each item. Hope et al. constructed the instrument with 48 items on a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) across six subscales. The six hypothesized subscales were (a) curriculum, (b) teaching, (c) work satisfaction, (d) stress, (e) accountability, and (f) students.

Upon using the instrument, Hope et al. (2006) reported that Cronbach's alpha reliability coefficient was .95 for the 48-item instrument. The subscale Cronbach's alpha coefficients were .70 for curriculum, .89 for teaching, .81 for work satisfaction, .88 for stress, .84 for accountability, and .47 for students. Overall, Cronbach's alpha was excellent for the 48-item instrument and Cronbach's alpha was good to very good for 5 of 6 subscales. However, the students subscale total scores would not be sufficiently reliable for analysis by itself.

2. Purpose of the Study

The purpose of this study was to investigate the psychometric properties of the *Teacher's High Stakes Testing Survey*. While Hope et al. (2006) carefully constructed this instrument, the authors presented little evidence of validity in their study. We wanted to conduct an in-depth analysis into the psychometric properties of the instrument before beginning a new study utilizing the instrument. First, each item was examined to determine whether the item was technically well-written. Second, the instrument was examined to determine whether any items should be added, modified, or deleted in order to improve the instrument. Finally, the instrument was analyzed to determine whether the items fit the hypothesized six-factor model and whether the instrument was measurement invariant by gender and race or ethnicity.

3. Methodology

The methodology section is divided into two subsections. First, we will discuss the population, sample, and sampling procedure. Second, we will present the participants demographic information.

3.1 Population, Sample, and Sampling Procedure

The Georgia Department of Education School Directory was used to select a stratified random sample of 100 elementary schools, middle schools, and high schools. Once schools were identified, teachers were randomly sampled from within school levels. The goal was to obtain sufficient data for a ±5 percent margin of error at a 95 percent confidence level (i.e., 386 teachers). After two mailings, 300 teachers completed the survey. Due to incomplete data, 15 surveys eventually were removed from the data analysis. Although almost 78% of the desired responses were obtained, the actual response rate was approximately 38% due to oversampling in order to account for nonrespondents.

An examination was made to determine how closely the sample matched the statewide demographics. The number of teachers in the sample reporting to be at the elementary level and at secondary level were similar to the statewide population, χ^2 (1) = 2.74, p = .10. The number of teachers in the sample reporting to be male and female were similar to the statewide population, χ^2 (1) = 1.54, p = .22. The number of teachers in the sample reporting to be Black, Hispanic, and White were not similar to the statewide population, χ^2 (2) = 16.68, p < .001. The sample had slightly fewer Black teachers and Hispanic teachers along with more White teachers than the statewide population.

3.2 Demographic Information

Demographic information collected on the survey included gender, race or ethnicity, educational level, and school level. Table 1 presents the number and percentage of teachers responding to the survey by demographic variable. One might note that three participants identified the school level as "Other" indicating that these teachers were teaching at a combination school. Either the teachers coded this variable incorrectly or the school directory did not identify the school correctly as a combination school. We chose to include these three participants in the analysis.

Table 1: Number and Percentage of Teachers Responding to the Demographic Variables

Variable	n	Percentage
Gender		
Female	236	83.39
Male	47	16.61
Race or Ethnicity		
African American	37	13.07
Asian or Pacific Islander	1	0.35
Caucasian	239	84.45
Hispanic	2	0.71
Other	4	1.41
Education Level		
Bachelor's Degree	80	28.27
Master's Degree	120	42.40
Ed. Specialist's Degree	67	23.67
Doctorate	16	5.65
School Level		
Elementary	154	54.42
Middle	53	18.73
High	73	25.80
Other	3	1.05

Note. n = 285 with 2 missing values.

4. Results

This results section consists of four subsections. First, we will report the results on the instrument validation by the expert review panel. Second, we will report the reliability analysis results. Third, we will present the results of the exploratory factor analyses. Finally, we will report the results of the confirmatory factor analyses along with the results about measurement invariance across subpopulations.

4.1 Instrument Validation

To begin the process, an Expert Panel Review Form was developed to collect information from our experts. Eventually, two review panels were formed. One review panel consisted of three college faculty members of the Educational Leadership program and another review panel consisted of eight practicing teachers at the elementary school, middle school, and high school levels. The two panels reviewed the *Teacher's High Stakes Testing Survey* for clarity of directions, adequacy of items to meet the intended purpose, item clarity, and grammatical correctness. In addition, panel members were asked to identify items that might be added or deleted to improve the instrument.

Feedback from the expert review panels was extremely positive. All expert panel members agreed that the survey directions were clear and the items matched the stated purpose. The expert panel identified seven items that potentially required modification. One panel member suggested for item 3, "Student achievement on a high stakes test accurately portrays the quality of a school's curriculum," be changed to "Students' scores on a high stakes test accurately portray the quality of a school's curriculum." Another panel member suggested that item 6, "High stakes testing promotes certain subjects' content over other subjects' content," be changed to "High stakes testing promotes certain subject area content over other subject area content." A panel member suggested that 'the' be added in item 12, 'Students' scores on a high stakes test are a valid way to determine the quality of education." Three panel members identified that in item 14, "High stakes testing requires test preparation that diminishes time to teach other subject content," that the word 'test or test administration' be added to the item. Another panel member suggested that item 29, "Punitive components of high stakes testing induce teacher stress," be changed to "Punitive measures associated with high stakes testing increases teacher stress."

One panel member suggested that item 37 "High stakes testing has increased teachers' awareness of accountability," be changed to "High stakes testing has increased teachers' awareness of the accountability issues in education." The final item that received a comment from the expert panel was item 49. One panel member suggested that we add 'the nature of" after changed in the item. Item 49 will now appear as "High stakes testing has changed the nature of student-teacher interactions."

In summary, the expert review panel provided very positive feedback about the directions and items comprising the *Teacher's High Stakes Testing Survey*. Panel members made a few substantive suggestions that will improve the instrument. In addition, expert panel members were asked to identify any items needed to improve subscale coverage. After a thorough review of these items, we selected one of these items for inclusion on the final version of the instrument. With the addition of one item, we now had a 49-item instrument (see Appendix A) on a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) across six subscales.

4.2 Statistical Analyses

The statistical analyses revealed a significant amount of information about the structure of the *Teacher's High Stakes Testing Survey*. The process included generating Cronbach's alpha reliability coefficients for the total scale and subscales, conducting an exploratory factor analysis, confirmatory factor analysis, and an examination of the measurement invariance by gender and race or ethnicity. Muthén (2004) suggested these last three analyses for instrument development in his lecture series on *Statistical Analysis with Latent Variables*.

4.2.1 Cronbach's alpha.

Cronbach's alpha reliability coefficient was used to assess the reliability of scores on the *Teacher's High Stakes Testing Survey*. Cronbach's alpha for the 49-item instrument was .94. The subscale Cronbach's alpha coefficients were .64 for curriculum, .90 for teaching, .75 for work satisfaction, .86 for stress, .83 for accountability, and .67 for students. The reliability estimate was excellent for the 49-item instrument and good to very good for most of the subscales. The reliability estimates for the curriculum subscale and the students subscale were adequate on this administration. Note that negatively worded items were reverse-coded for the estimates of reliability and subsequent analyses.

4.2.2 Exploratory factor analyses.

The exploratory factor analyses were run allowing the *Teacher's High Stakes Testing Survey* items to load on an unspecified number of factors. Kaiser's criterion, Cattell's scree test, and residuals were examined for each of the factor models (see Stevens, 2002) in order to select the most appropriate parsimonious factor model. All three of the criteria indicated that at least four factors were present. Kaiser's criterion of 1 indicated that there were up to 10 factors present, while Cattell's scree test indicated that at least four factors fit the model. An examination of the residuals indicated a decrease in the root mean square residual from .05 to .03 as one went from 4 to 10 factors.

The four-factor model had 22 items, 9 items, 6 items, and 4 items load on the factors, whereas the five-factor model had 19 items, 9 items, 3 items, 10 items, and 4 items load on the factors. The six-factor model had 19 items, 8 items, 2 items, 7 items, 4 items, and 2 items load on the factors. On each of the two-item factors in the six-factor model, the loadings were in the range of .33 to .54. These two-item factors were disregarded as was the six-factor model. The seven-factor model had 18 items, 8 items, 3 items, 7 items, 4 items, 2 items, and 0 items load on the factors. The two-item factor in the seven-factor model had item loadings in the range of .57 to .65. Like the six-factor model, the seven-factor model was disregarded. The eight-factor, nine-factor, and 10-factor models had even more factors with zero items to two items loading on a factor and were subsequently disregarded.

This left the four-factor model and the five-factor model to examine. Items in each of the factor models were inspected to determine how well these items fit together. The four-factor model had 22 of 49 items load on a single factor, 23 other items that loaded almost evenly across the three other factors, and four items did not load on a factor. However, it was extremely difficult to determine how items on three of the four factors were related to one another. Similarly, the five-factor model had one dominant factor with 19 of 49 items loading on that factor, two factors with approximately 10 items, two factors with approximately four items, and four items that did not load on a factor. It was determined that two factors had similar items, one factor had fairly similar items, and two factors had dissimilar items. Neither the four-factor model nor the five-factor model as generated by the exploratory factor analysis seemed logical. So, in the end these two models were disregarded too.

Finally, the original factor structure of the instrument as developed by Hope et al. (2006) was examined. While the original factor structure was not reproduced by the exploratory factor analysis, it was determined that this factor structure would be employed due to its simplicity and ease of understanding. Item scores within factors were totaled to use in the subsequent confirmatory factor analyses.

4.2.3 Confirmatory factor analyses.

In this phase a number of confirmatory factor analyses were conducted and fit indices were examined to assess the quality of each model. There is no single statistic that one employs when assessing model fit, rather one examines a number of fit indices to assess model fit. Hu and Bentler (1999) suggested cutoff values for a number of common fit indices. The suggested cutoff values for the comparative fit index (CFI), Tucker and Lewis fit index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) were 0.95, 0.95, 0.06, and 0.08, respectively. Previously, the suggested cutoff values for good model fit were approximately were CFI > 0.90, TLI > 0.90, and RMSEA < 0.08 (Muthén & Muthén, 2001).

The initial baseline model used the original factor structure of the instrument as developed by Hope et al. (2006). This initial baseline model did not allow correlation among the factor scores. The chi-square statistic, CFI, TLI, RMSEA, and SRMR values did not meet the criteria for good model fit (see Table 2). However, a final baseline model was generated that allowed the correlation among factors. For this final baseline model, the chi-square statistic, CFI, TLI, RMSEA, and SRMR either surpassed the minimal criterion or were close to the fit indices suggested for good model fit.

Once the final baseline model was established, separate multiple group analyses were conducted. One multiple group analysis was conducted by gender and another multiple group analysis was conducted by race or ethnicity. Mplus (Muthén & Muthén, 2006) by default constrains intercepts and factor loadings to be equal across groups, allows residual variances to be free, and factor means are held at zero in one group and free in the other groups. Muthén and Muthén contended that these default values are sufficient to establish measurement invariance. In these analyses, male and White were the reference groups and female and Black were the focal groups.

Table 2: Fit Indices by Confirmatory Factor Analysis for the Teacher's High Stakes Testing Survey

	Chi-Square	*df*	*p*	CFI	TLI	RMSEA	SRMR
Initial Baseline Model – no correlation among factors	176.62	9	.000	.82	.68	.259	.103
Final Baseline Model – correlation among factors	5.46	4	.243	.99	.99	.036	.019
Factorial Invariance for Gender – indirect effect	16.56	9	.056	.99	.98	.055	.034
Factorial Invariance for Gender – direct effect	8.46	8	.389	1.00	1.00	.014	.026
Factorial Invariance for Race or Ethnicity (White & Black) – indirect effect	5.74	9	.761	1.00	1.01	.000	.017
Factorial Invariance for Race or Ethnicity (White & Black) - direct effect	1.93	8	.983	1.00	1.02	.000	.008

Note. Comparative fit index (CFI), Tucker and Lewis fit index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR).

When gender was added to the model, the chi-square statistic, CFI, TLI, RMSEA, and SRMR either surpassed the minimal criterion or were close to the fit indices suggested for good model fit (see Table 2). Moreover, the addition of gender in the model was not significant ($p > .05$) indicating that there was not a difference by gender in responding to the *Teacher's High Stakes Testing Survey*. Then, the direct effect from gender to each of the six factors was added to the model.

Of the six factors, only the fifth factor (accountability) was significant ($p < .05$) indicating that female teachers responded more positively with higher scores than did male teachers on this one factor. While not invariant on this one factor, overall one might conclude that the instrument was measurement invariant for gender.

In the second multiple group analysis, only Black teachers and White teachers were considered due to insufficient numbers of teachers from other races or ethnicities. The chi-square statistic, CFI, TLI, RMSEA, and SRMR either surpassed the minimal criterion or were close to the fit indices suggested for good model fit. When race was added to the model, race was significant ($p < .05$) indicating that overall Black teachers responded more positively with higher scores than White teachers on the instrument. Then, the direct effect from race or ethnicity to each of the six factors was added to the model. The direct effect from race or ethnicity to each of the six factors was not significant ($p > .05$) indicating that the instrument was measurement invariant by race or ethnicity (White and Black).

5. Conclusion

In order to utilize the *Teacher's High Stakes Testing Survey* in future studies, the psychometric properties were assessed. First, an expert panel of three university professors and eight educators were selected to participate in the review process. Second, given specific directions as a guide, the expert review panel reviewed the technical quality of the items. Overall, the expert panel found that the items were well constructed. The panel suggested the modification of a couple of items with minor word changes. Third, the expert panel was asked deliberately to address whether any items should be deleted or added for instrument improvement. The panel did not recommend any items for deletion, but an item suggested by the panel review committee was added to the instrument.

In addition, item responses were analyzed to assess whether items fit the hypothesized six-factor model and to assess whether the instrument was measurement invariant across subpopulations. Exploratory factor analyses and confirmatory factor analyses were conducted. The baseline model (i.e., hypothesized six-factor model) with correlated factors fit the model well. Confirmatory factor analyses supported measurement invariance by gender and for race or ethnicity (i.e., White and Black).

6. References

Afflerbach, P. (2005). National reading conference policy brief: High stakes testing and reading assessment. *Journal of Literacy Research, 37*, 151-163.

Airasian, P. W. (1988). Symbolic validation: The case of state-mandated, high stakes testing. *Education Evaluation and Policy Analysis, 10*(4), 310-313.

Baines, L. A., & Stanley, G. K. (2004). High stakes hustle: Public schools and the new billion dollar accountability. *The Educational Forum, 69*(1), 8-16.

Burton, N. W. (1978). Societal standards. *Journal of Educational Measurement*, *15*, 263–271.

Camilli G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History, and future perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 445–475). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Darling-Hammond, L. (2006). Constructing 21st-century teacher education. *Journal of Teacher Education, 57*, 300-314.

Driesler, S. D. (2001). Whiplash about backlash. The truth about public support for testing. *NCME Newsletter*, *9*(3), 2–5.

Engelhart, M. D. (1950). Examinations. In W. S. Monroe (Ed.), *Encyclopedia of educational research* (pp. 407-414). New York: Macmillan.

Evers, W. M. (2001, August 20). What do tests tell us? Hoover Daily Report.

Fremer, J. (2005). Foreword. In R. P. Phelps (Ed.), *Defending standardized testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hamilton, L. S. (2003). Assessment as a Policy Tool. *Review of Research in Education, 27*, 25-68.

Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), Making sense of test-based accountability in education (pp. 13-49). Santa Monica, CA: RAND.

Haycock, K. (2005). Choosing to matter. *Journal of Teacher Education, 56*, 256-265.

Holland, R. (2001). *Indispensible tests: How a value-added approach to school testing could identify and bolster exceptional teaching*. Arlington, VA: Lexington Institute.

Hope, W. C., Brockmeier, L. L., Lutfi, G. A., & Sermon, J. M. (2006, November). *High stakes test's influence on teachers' beliefs*. Paper presented at the annual meeting of the Florida Educational Research Association, Jacksonville, FL.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Kaback, S. (2006). High stakes are for tomatoes: Supporting teachers and students when the growing conditions are poor. *Kappa Delta Pi, 42*, 101-103.

Lederman, L. M., & Burnstein, R. A. (2006). Alternative approaches to high stakes testing. *Phi Delta Kappan, 87*(6), 429-432.

Lewis, A. (2000). High-stakes testing: Trends and issues [Policy brief], Aurora, CO: Mid-Continent Research for Education and Learning.

Linn, R. L. (2006). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation Education, 19*, 5–15.

Linn, R. L., Miller, M. D., & Gronlund, N. E. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Meyer, R.H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief, 3*(3). Madison, WI; National Center for Improving Science Education, University of Wisconsin-Madison.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus user's guide: Statistical analysis with latent variables.* Los Angeles, CA: Author.

Muthén, B. (2004). *Statistical analysis with latent variables: Multiple-group confirmatory factor analysis*. Retrieved from http://www.ats.ucla.edu/stat/seminars/ed231e/ on September 20, 2007.October.

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus user's guide* (4th ed.). Los Angles, CA: Muthén & Muthén.

National Commission on Excellence in Education. (1983). *A nation at risk.* Washington, DC: U.S. Department of Education.

No Child Left Behind Act of 2001, Pub. Law No. 107–110.

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

Phelps, R. P. (2005). Forty years of public opinion. In R. P. Phelps (Ed.), *Defending standardized testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15*, 297–300.

Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.) *Ability testing: Uses, consequences, and controversies* (pp.173-194). Washington, DC: National Academy Press.

Scherer, M. (2005). Reclaiming testing. *Educational Leadership, 63*, 9.

Smith, M. S., O'Day, J., & Cohen, D. K. (1990). National curriculum American style: Can it be done? What might it look like? *American Educator, 14*(4), 10-17, 40-47.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences (4th ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, J. E. (2003). Preface. In R. P. Phelps (Ed.), *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

Tyack, D. (1974*). The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.

Vogler, K. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education, 123*, 39-51.

Wahlberg, H. J. (2003). Foreword. In R. P. Phelps (Ed.), *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

Wakefield, D. (2003). Screening teacher candidates: Problems with high-stakes testing. *The Educational Forum*, 67, 380-388.

Appendix A

Items on the Teacher's High Stakes Testing Survey

Curriculum	
1	High stakes testing has led teachers to reassess their beliefs about subject matter that is important to teach.
2	High stakes testing is counter to the idea of a balanced curriculum (equal attention to subjects).
3	Students' scores on a high stakes test accurately portray the quality of a school's curriculum.
4	High stakes testing requires teachers to teach to the test.
5	High stakes test items accurately reflect the content students learn through a school's curriculum.
6	High stakes testing promotes certain subject area content over other subject area content.
7	Students' scores on a high stakes test provide feedback for schools to improve the curriculum.
8	High stakes test content is aligned with a school's curriculum.
Teaching	
09	High stakes testing permits teachers to use the full range of their teaching skills.
10	High stakes testing leads to better teaching.
11	Students' scores on a high stakes test are a valid measure of teaching ability.
12	Students' scores on a high stakes test are a valid way to determine the quality of education.
13	The quality of teachers' instruction is directly related to student performance on a high stakes test.
14	High stakes testing requires test preparation that diminishes time to teach other subject content.
15	Students' scores on a high stakes test provide information for teachers to improve their teaching.
16	High stakes testing reduces the teaching and learning process to a student's test score.
17	High stakes testing motivates teachers to improve the teaching and learning process.
18	High stakes testing has increased cooperation among teachers.
19	High stakes testing has increased teacher and principal cooperation.
Work Satisfaction	
20	Teacher morale has increased because of high stakes testing.
21	High stakes testing diminishes the desire to be an educator.
22	Teachers leave low performing schools because of high stakes test results.
23	The use of high stakes testing as a single measure to determine student achievement leads to teachers leaving the profession.
24	Teachers' work satisfaction diminishes when the focus is on high stakes testing outcomes.
25	Teacher satisfaction increases when she or he has input into the development of a high stakes test.

Note. 1 (*Strongly Disagree*), 2 (*Disagree*), 3 (*Neither Agree nor Disagree*), 4 (*Agree*), and 5 (*Strongly Agree*).

Appendix A (continued)

Items on the Teacher's High Stakes Testing Survey

Stress	
26	High stakes testing leads to competition among teachers.
27	eachers' stress increases when the school receives a failing grade.
28	Teachers' stress increases when the school's accountability grade declines.
29	Punitive measures associated with high stakes testing increase teacher stress.
30	Teachers experience stress in the effort to maintain their school's accountability grade.
31	Teachers' stress increases with public advertisement of a school's high stakes test results.
32	The pressure of high stakes testing may result in teachers cheating to improve scores.
33	District supervisors' pressure to improve high stakes test scores increases teacher stress.
34	Principals' pressure to improve high stakes test scores increases teacher stress.
35	Teachers leave the profession because of stress related to high stakes testing.
Accountability	
36	High stakes testing has increased teachers' accountability for students' academic performance.
37	High stakes testing has increased teachers' awareness of the accountability issues in education.
38	High stakes testing is an effective means of determining the quality of public education.
39	Students' scores on a high stakes test are an indicator of whether a school is staffed with high quality teachers.
40	High stakes testing is a reform measure that improves the quality of education.
41	Teachers are more accountable because of high stakes testing.
42	High stakes testing creates a cooperative environment between teachers and the community.
Students	
43	High stakes testing contributes to the number of students that drop out of school.
44	Students' learning styles are accounted for in high stakes testing.
45	High stakes testing induces anxiety in students.
46	High stakes testing motivates students to achieve.
47	The pressure of high stakes testing may result in students cheating to improve scores.
48	Teachers are concerned about the impact of high stakes testing on minority students.
49	High stakes testing has changed the nature of student-teacher interactions.

Note. 1 (*Strongly Disagree*), 2 (*Disagree*), 3 (*Neither Agree nor Disagree*), 4 (*Agree*), and 5 (*Strongly Agree*).