Technical Characteristics of State Assessments of Skills and Knowledge

Christopher H. Tienken
Monroe Township School District
Rutgers University, Graduate School of Education (Part-time)

Michael J. Wilson
Western Connecticut State University

Abstract

Most large scale state tests (e.g. New York, North Carolina, California, Florida, South Carolina, Michigan, Georgia, New Jersey, Washington, Massachusetts, and 23 other states) have documented flaws that limit their usefulness to educators as diagnostic or decision-making tools. In this article the researchers examine the technical characteristics of New Jersey's (NJ) state-mandated assessment for grades 3 and 4 to provide examples of problems associated with large-scale state testing programs. Additionally, we report the outcomes of a statewide stratified random sample survey that investigated the ways district leaders use the test results provided by the state. The authors present an analysis of the technical characteristics in light of the high stakes attached to the test results and discuss the ways educators use those results. The analysis suggests that the sub-scale levels (content clusters) of the assessments demonstrate undesirably low levels of reliability, lack content validity, and the results are strongly correlated to socioeconomic status. District leaders and policy-makers should not use the results of the tests as the sole evaluation tool for curriculum, instruction, or student achievement at the local level. Educators, policy-makers, and parents should question the technical characteristics of their states' large-scale tests and require that the tests meet at least minimum levels of reliability and content validity.

Technical Characteristics of State Assessments of Skills and Knowledge

Introduction

The Problem

Most large scale state tests (e.g. New York, North Carolina, California, Florida, South Carolina, Michigan, Georgia, New Jersey, Washington, Massachusetts, and 23 other states) have documented limitations and serious flaws that prevent them from providing diagnostic student achievement data that teachers and school leaders can use to make important decisions (Standard and Poors 2005; Darling-Hammond, Rustique-Forrester, Pecheone & Andree, 2005; American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Neill, 1997). The most common of these problems is the myopic focus of measurement (Klein & Hamilton, 1999).

In this article we examine the technical characteristics of the state-mandated New Jersey Assessment of Skills and Knowledge (NJASK) for grades 3 and 4 to provide specific examples of problems associated with large-scale testing programs. The issues presented are indicative of technical problems faced in other states. We analyze the usability of the results for school leaders and teachers and present commentary about the technical characteristics in light of the high-stakes attached to the test results.

Background

Most states, 49/50 (except Iowa), instituted core curriculum standards, and tests to monitor the implementation of those standards, by 2002. The New Jersey Legislature created the current standards-based state testing program in 1997 to monitor implementation of the New Jersey Core Curriculum Content Standards (NJCCCS). The New Jersey Department of Education (NJDOE), and the other 49 states, administered language arts and mathematics tests in grades 3-8 and 11 during the 2005-2006 school year to comply with the No Child Left Behind Act of 2002 (NCLB, PL 107-10).

The NJDOE categorizes students as *Partially Proficient, Proficient* or *Advanced Proficient* based upon their performance on the annual tests. The NJDOE defines *Proficient* as attaining a score of 200 to 249 and *Advanced Proficient* as attaining a score of 250 or above. Students are categorized *Partially Proficient* if they score below 200. The theoretical maximum test score is between 285-300. The state administers the tests one time per year in March or April.

High Stakes

According to Popham (2001), two conditions must be present for a test or testing program to be considered high-stakes: (a) There must be a significant consequence related to individual student performance; and (b) The test scores (of the students) must be the basis for the evaluation of quality and success of individual teachers and school districts. The NJASK 3 & NJASK 4 meet the conditions of “high-stakes” tests as defined by Popham.

Consequences for Students

The NJDOE uses a combination of social and economic elements to categorize its approximately 600 school districts into District Factor Groups (DFG). The NJDOE uses DFG as a proxy for socioeconomic status (SES) of the district’s residents. Districts are rated A though J. The “J” factor group represents school districts located in highest SES communities as measured by the NJDOE and “A” school districts are located in the state’s poorest communities.

We conducted a proportional stratified random sample survey, via phone, of 74 districts in New Jersey, representing all eight DFG’s, to investigate two aspects of the ways district leaders use the NJASK 3 & 4 test results. We made contact with central office staff, after up to three attempts, from 47 school districts (64%) representing the continuum of DFG's. We asked the administrators five questions[1] about two aspects of how they use state test results. Aspect 1: The role of state testing results in local decision-making about student placement in elementary school Title I basic skills instruction (BSI) programs; Aspect 2: The influence of the NJASK 3 & 4 results on curricula and program evaluation. Districts in our sample used students' NJASK scores to make student placement decisions and to evaluate the effectiveness of curricula and programs.

The survey results indicated that districts stream students into Title I BSI programs (in elementary and middle school) based partially or totally (depending on the district) on their results from state tests. Most of the districts responding to the survey (46/47 or 98%), used state test results as one factor in their decision making for placing students into BSI programs. Approximately 55% of the districts (26/47) used state test results as the deciding factor for placing individual students into BSI classes. Historically, many students placed into a BSI program during elementary school do not exit that program by the end of eighth-grade (Borman & D'Agostino, 1996). The act of placing a child into a BSI program represents a significant consequence for students.

Consequences to School Districts

Schools and school districts experienced negative effects of high-stakes testing in NJ before the implementation of the No Child Left Behind (NCLB) Act passed on January 8, 2002. However, NCLB's prescribed penalties increase the pressure. For example, in New Jersey, schools whose students fail to make adequate yearly progress (AYP) in any one of 40 categories two years in a row are labeled by the state as "a school in need of improvement" and labeled in the media as a failing school. This year, 2005-2006, the NJDOE labeled 539 schools out of approximately 2,300 as “in need of improvement” and identified 55 out of approximately 600 districts for “corrective action and restructuring”. Similarly, the California Department of Education (CDOE) labeled 1,626 schools out of 9,087 as being in need of improvement during the 2004-2005 school year based on the results of a state test. Virginia, Ohio, Florida, Michigan, New York, Illinois and Georgia each had more than 400 schools in a stage of improvement status during the 2004-2005 school year.

The “failing” schools and districts must provide for intra-district school choice and use Title I funds to pay supplemental services providers (SSP) to serve their students. Schools in which students fail to make AYP for five consecutive years can be taken over by the state, be outsourced to a private firm, have their staff dismissed, or have the school governance structure reconstituted. State take-over of schools occurred in New Jersey prior to NCLB, but on a limited scale (i.e. Jersey City and Newark). Similar consequences exist in the other 49 states under NCLB.

Technical Characteristics

Reliability

States that attempt to measure large subject domains (i.e. mathematics, language arts) using tests with relatively few questions (i.e. 30-40) risk the testing program to reliability threats, especially at the subscale level. Reliability threats reduce the trust educators can place in the assessment of the domain clusters (e.g. Number Sense, Algebra, Geometry, etc.). New Jersey reports student and district level performance at the domain cluster level (sub-scale), through the use of Cluster Scores. Lower levels of reliability mean an increase in score error and a decrease in instrument precision. The lack of instrument precision can produce results that lead district administrators, teaching staff, policy-makers, and the general public to draw faulty conclusions about the effectiveness of staff, programs, and curricula. The administrators of New Jersey’s testing program acknowledge this limitation, yet the program continues (NJDOE, 2003; 2003a):

> When evaluating these results it is important to recall that reliability is partially a function of test length. Therefore, the reliability of a content area is likely to be greater than the reliability of a cluster simply because the content area has more items. (p. 12).

The NJDOE uses Cronbach’s Alpha to determine reliability estimates. Alpha is an indication of the extent to which all the items in a particular scale, or cluster, such as *Geometry & Measurement,* are working together to create consistent results. The level of desired reliability depends on how one intends to use the scores. Some statisticians and national organizations recommend high test reliability when making judgments about individual students (APA, NCME, AERA, 1999).

Our survey results indicated that district leaders use the Cluster Scores provided by the NJDOE to evaluate the achievement of individual students and to make decisions about the effectiveness of curricula. Using the mathematics section of the NJASK 3 & 4 as an example, consider the reliability concerns for school district leaders and the possible impacts on students (See Table 1).

Educators and policy-makers should use a minimum reliability coefficient of at least .85 (Frisbie, 1988) when making high-stakes decisions about students, although an argument can be made for greater reliability. A minimum reliability coefficient of .69 is desired when making judgments about groups ($n \geq 25$) of students or programs (Frisbie, 1988, p.29). Each Cluster Score on the NJASK represents the students’ performance on those items from the test, (only six to eight items in most clusters) not the complete scope of skills and knowledge contained in that cluster of the state mandated NJCCCS.

Table 1

Reliability estimates for all items from the NJASK 3 & 4 mathematics clusters (NJDOE, 2004).

Cluster	*n* Dichotomous Items/Grade		Reliability Estimates		*n* Open-ended Items/Grade	
	3rd	4th	3rd	4th	3rd	4th
Total Test Items	27	32	-	-	3	5
Number sense, concepts, Applications	12	11	.71	.73	0	2
Geometry/Measurement	5	7	.44	.59	1	1
Data analysis, probability, Statistics, discrete math	5	7	.60	.61	1	1
Patterns, Functions, Algebra	5	7	.57	.60	1	1
Problem solving [a]	17	18	.83	.86	3**	5**

Note: Reliability estimates reported for entire cluster.
[a] Problem solving is embedded in the first four clusters. The NJDOE claims that the questions on the test address up to three progress indicators at one time.
** Sum of open ended items on the test. Not discreet items for problem solving. (e.g. One problem solving item from geometry, data analysis and algebra clusters on the NJASK 3).

For example, there are 38 cumulative progress indicators and sub-indicators for the Number Sense/Operations/Estimation cluster of the NJASK4. Only two mathematics clusters (i.e. Problem Solving NJASK3 and NJASK4) out of ten, 20%, meet the minimum reliability threshold for making group decisions and none meet the minimum desired reliability for making decisions about individual students. Similar reliability conditions exist in the language arts sections of the NJASK 3 & 4 (See Table 2). Validity issues are discussed in a subsequent section.

Table 2

Reliability estimates for all items from the NJASK 3 & 4 language arts clusters (NJDOE, 2004).

Cluster	n Dichotomous Items/Grade		Reliability Estimates		n Open-ended Items/Grade	
	3rd	4th	3rd	4th	3rd	4th
Total Test Items	16	16	-	-	4	5
Reading [a]	12	11	.80	.82	-	-
Working with Text [b]	9	6	.70	.55	0	0
Interpreting Text [b]	3	5	.65	.78	2	3
Writing	0	0	.72	.75	2	2

Note: [a] Reliability reported for entire cluster. [b] Working with Text and Interpreting Text are sub-categories of Reading as measured by the NJASK.

Because testing populations in school districts are considerably smaller than the state testing population used to compile the technical characteristics of the NJASK 3 & 4, it is quite possible that test reliability will be much lower at the local level than for the state as a whole. The state does not provide district leaders with school-level reliability results, thus leaders cannot gauge the actual amount of error they must factor into interpretation of test scores. In a district with very unreliable results the leaders might unknowingly change a successful program or not recognize deficiencies in a program, but due to the size of the error in scores, the success or need for improvement might not be reflected in the results provided by state. This may occur more frequently in districts with smaller testing populations because scores can fluctuate more from year to year because of changes in student characteristics.

The NJDOE does not freely distribute the technical manuals for New Jersey's tests to district leaders nor does the NJDOE discuss publicly the error or measurement. It took over 18 months and multiple emails and phone calls to acquire the technical manuals for the NJASK 3 & 4. Thus district personnel seldom, if ever, understand the extent to which their scores are influenced by testing error.

We acknowledge that the Cluster Scores may represent an acute example and some statisticians may argue that it is inappropriate to discuss the reliability of individual clusters. However, the results of our survey demonstrated that leaders in NJ use them to make programmatic and curricular changes and to make specific recommendations for students' Title I Basic Skills program placement. Monitors from the state DOE also use Cluster Scores to make recommendations for improvement to district leaders. While it may be inappropriate for leaders to make decisions and programmatic changes based on performance indicators from individual clusters, the NJDOE (2004) insists that the results can be used to make curricular decisions:

> The NJASK is designed to give an early indication of the progress students are making in mastering the knowledge and skills described in the Core Curriculum Content Standards. The results are to be used by schools and districts to identify strengths and weaknesses in their educational programs." (p.1)

Since 1997, officials at the NJDOE insisted the tests were diagnostic. In 2004 and 2005 they admitted in limited public forums that the tests were for monitoring purposes only and should not be used to make in-depth diagnoses of students' strength and weaknesses (Doolan, J, personal communication May 9, 2004; Robinson, B. personal communication May 19, 2005). Yet the myth of diagnosis persists and, in actuality, educators use the test results for making decisions about individual students. Indeed, if test results cannot be used for diagnosis and decisions for improvement, of what practical value are they?

Individual Interpretation and Measurement Error

The standard error of measurement (SEM) is a measure of variability of the errors of measurement and it is related to the error score variance (Harville, 1991). Measurement error can cause district leaders to make inaccurate interpretations about student achievement.

The SEM for scale scores exceeds 8.96 points for each section of the NJASK 3 & 4 (See Table 3). Using reasonable limits, it is likely that the real score is +/- 8.97 points (a total of 17.4 points) around the reported student score. The lower the reliability estimate, the larger the SEM area around the score.

Table 3

Standard error of measure (SEM) for language arts and mathematics sections from the NJASK3 & 4 (NJDOE, 2004).

	Standard Error of Measure/Test Content	
Test Level	Language Arts	Mathematics
NJASK3	9.12	11.62
NJASK4	8.97	11.21

We present SEM's for the language arts and mathematics sections of the NJASK tests as an example of the sizable error associated with large-scale state tests. A more appropriate statistic is the Conditional SEM associated with proficiency level cut-scores (Harville, 1991).

The NJDOE provides *Conditional* raw score SEM's for the *Proficient* and *Advanced Proficient* levels (See Table 4). A concerned educator would expect a small SEM and a correspondingly high level of reliability if the test were to be used for high-stakes decisions (e.g. determining the effectiveness of curricula or the programmatic placement of students). Yet sizeable levels of SEM exist, 6%-15.4%, at the cut-score levels on the NJASK 3 & 4.

Table 4

Estimated Conditional Standard Error (SE) in raw points at each cut-score level on the NJASK 3 & 4 language arts and mathematics sections (NJDOE, 2004).

Section/ Rating	Raw Cut Score		Estimated SE in Raw Points		% of Error	
	3rd	4th	3rd	4th	3rd	4th
Language Arts						
Partially Proficient	<18.0	<19.0				
Proficient	18.0	19.0	2.5	2.5	14.0	13.0
Advanced Proficient	30.5	33.5	2.0	2.0	7.0	6.0
Mathematics						
Partially Proficient	<17.0	<19.5				
Proficient	17	19.5	2.5	3.0	15.0	15.4
Advanced Proficient	27.5	32.5	2.0	3.0	7.0	9.2

Given the statistics in Tables 1-4, district leaders and teachers should base decisions about student achievement and school effectiveness on information more precise than the results provided through the state's tests. States and school districts should use multiple measures. The NCLB Act requires the use of multiple measures, including measures of higher-order thinking and understanding (USDOE, 2002) as part of state accountability assessment systems. Using more than one test and an assessment system built upon multiple measures (e.g. portfolios, performance assessments, and class grades/or Grade Point Average: GPA) could provide district leaders and teachers with more complete information about students, but that requires sophisticated knowledge of testing to produce and understand the use of multiple measures. No single large-scale standardized test provides the desired levels of reliability and validity necessary to be used as the sole basis for making important decisions about students (AERA, APA, NCME, 1999)

Content Validity

"Validity is, hands down, the most significant concept in assessment" (Popham, 2002, p. 47). Content validity is the extent an assessment adequately represents the subject domain being sampled (Popham, 2002). Babbie (2001) defined content validity as how well a measure covers the range of meanings included within the concept. Error due to content sampling is generally the greatest factor of measurement error (Rudner, 1994). Large-scale state testing programs generally attempt to measure a wide array of knowledge and skills with a relatively limited number of test questions. The sheer number of indicators to be measured and the small quantity of test items guarantee that the results will be unhelpful to a teacher attempting to diagnose the learning processes of individual students. The NJDOE provides limited guidance on how the state's tests relate to individual aspects of the core curriculum standards. The technical characteristics of the NJASK suggest that the tests have little, if any value in the evaluation of individual students, or school districts, and they are not helpful for the type of planning required to customize instruction.

“Content validity goes awry when the assessment instrument does not represent the content accurately” (Tanner, 2001, p. 178). The NJASK 3 & 4 test items sample only a small part of a larger domain of content. It is difficult to determine from the information available publicly how well the items on the state tests represent the domains they attempt to assess. How well can 37 items on the NJASK4 math section, or 16 items on the NJASK3 Language Arts section, measure over 200 curriculum progress indicators from two large curriculum domains?

Misleading Results

The NJDOE reports results for the state assessments in several forms. One form is by DFG (see Table 6). The majority of the “A” districts are located in urban areas and the majority of New Jersey’s poorest and minority students live in urban areas. The districts in the groups closest to DFG “A” generally score lower than those districts closest to group “J” (Table 5).

The DFG results produce Pearson correlation coefficients ranging from .91 to .97 and a Spearman’s Rho of 1.0. The correlations between DFG and NJASK test score are statistically significant ($p \leq .01$) highlighting the inequities that exists in the societal conditions and access to educational opportunities available to students in the state. The correlation of DFG to student achievement on the state assessments is not surprising. A recent study (Michel, 2004) of the 2004 NJASK4 results found SES to be the strongest predictor of student achievement for that test, similar to nationwide analyses (e.g, Coleman, Hobson, McPartland, Mood, Weinfeld, & York, 1966; Cooley, 1993; Hart and Risely, 1995; Rothstein, 2004; Standard and Poors, 2005).

Table 5

Percentage of regular education students in each DFG who scored proficient or advanced proficient on the language arts and mathematics sections of the NJASK3 & 4 (NJDOE, 2004).

DFG	Language Arts		Mathematics	
	3rd	4th	3rd	4th
A	66.2	74.2	57.8	59.6
B	81.1	85.6	75.2	70.6
CD	83.9	89.4	80.4	75.3
DE	90.3	93.4	85.9	80.7
FG	91.8	94.6	86.9	82.6
GH	93.3	95.5	90.1	85.4
I	96.5	97.9	93.6	90.5
J	96.8	98.6	93.8	91.7
Pearson (r)	.91[a]	.91[a]	.91[a]	·97[a]
Spearman Rho (r_s)	1.0[a]	1.0[a]	1.0[a]	1.0[a]

[a] = Statistically significant at $p \leq .01$

The strong correlations between DFG and test-score results on the NJASK 3 and 4 suggest the presence of a de facto *access to opportunity* (social and educational) inequity in the state: The test results tell more about a community's social well-being and barriers to opportunities to learn than the quality of the curriculum and instruction taking place in the school district. This is problematic for educators in districts labeled "failing" by the state who look to the test results for answers. Leaders cannot find answers when the test results tell more about the economic and social standing of the students' community than how well the curricula and instruction function.

Summary

States' reliance on imprecise, blunt testing instruments as a measure on which to base important educational decisions is troubling given the high-stakes nature of NCLB and the ways in which district leaders use the results. The problems associated with high-stakes testing are not limited to New Jersey. However, the consistent defense of these testing programs by state departments of education and state boards of education is striking. One would expect higher quality assessment instruments that produce better information to make education decisions given NCLB-imposed penalties for districts associated with poor performance on the test.

Many states struggle with budget deficits and funding restrictions. They cannot allocate the funds necessary to improve the testing programs. States are forced to rely on large-scale assessments with too few questions and a narrow focus on skills and knowledge that are easily measured. For example, representatives from the NJDOE have admitted publicly that finances, not technical integrity, drive the state's assessment program (Robinson, B. Personal communication, May 19, 2005). The current philosophy is "do the best with what we have."

The testing instruments used by many states as part of the federally-mandated testing program do not support the high-stakes attached to their results. The tests do not represent a complete measure of the mandated content standards (content validity). The instruments lack the desired level of precision and consistency (reliability) that must accompany a high-stakes accountability system. District leaders and teachers are unable to use the results in their current form to individualize instruction effectively, or to evaluate the effectiveness of curriculum, yet the performance of individual students on the state's tests determine if schools meet Annual Yearly Progress (AYP) indicators.

Financial pressures are a reality in many states, but what are the long-term costs to the futures of students by providing limited, or misleading assessment information to district leaders and teachers? "Demanding accountability without providing adequate resources can be an evasion of accountability by setting up public schools for failure" (Sloan-McCombs & Carroll, 2005). Accountability is not a solitary endeavor. It must be shared and understood by teachers, administrators, parents, communities, students, and state policy makers (Linn, 2003).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing,* Washington, D.C.: American Educational Research Association, 142.

Babbie, E. (2001). *Practice of social research.* 9th Ed. Belmont, CA. Wadsworth Thomson Learning.

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

Borman, G.D. & D'Agostino, J.V. (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation & Policy Analysis, 18,* 309-326.

Coleman, J., Campbell, E.Q., Hobson, C.F., McPartland, J.M.; Mood, A.M.; Weinfeld, F.D., & York, R.L.(1966). *Equality of Educational Opportunity*. Washington DC: U.S. Government Printing Office.

Cooley, W.W. (1993, Summer). The difficulty of the educational task. *ERS Spectrum*. 27-31.

Darling-Hammond, L., Rustique-Forrester, E., Pecheone, R., & Andree, A. (2005). *Multiple Measures Approaches to High School Graduation.* School Redesign Network.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7*(1), 25-35.

Hart, B. & Risley, T. (1995). *Meaningful differences*. Baltimore, Md.: Brookes Publishing.

Harville, L.M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices, 10*(4) 181-189.

Klein, S.P. & Hamilton, L. (1999). *Large scale testing: Current practices and new directions.* Rand Corporation.

Koretz, D., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991, April). The effects of high-stakes testing: Preliminary evidence about generalization across tests. In R.L. Linn (Ed.), *The effects of high stakes testing.* Symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Linn, R.L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3-13.

Michel, A.P. (2004). What is the relative influence of teacher educational attainment on student NJASK4 scores? Unpublished Doctoral Dissertation. Seton Hall University.

National Research Council, Committee on Appropriate Test Use (1998). *High Stakes: Tests for Tracking, Promotion, and Graduation.* Washington, DC: National Academy Press.

Neill, M. (1997). *Testing our children: A report card on state assessment systems*. FairTest.org Retrieved on March 18, 2006 from http://www.fairtest.org/states/survey.htm.

New Jersey Department of Education. (2003). Grade four elementary school proficiency assessment: 2001 Technical report. Report. #1504.38. NJDOE.

New Jersey Department of Education. (2003a). Grade eight proficiency assessment: March 2002 Technical report. #1503.65. NJDOE.

New Jersey Department of Education.. (2004). Grades 3 and 4 New Jersey assessment of skills and knowledge. #1505.11. NJDOE.

Popham, W.J. (2001). *The truth about testing: An educator's call to action*. Arlington, VA: ASCD.

Popham, W. J. (2002). *Classroom assessment: What teachers need to know*. Boston: Allyn and Bacon.

Rothstein, R. (2004). *Class and schools*. Teachers College: Columbia University.

Rudner, L.M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation*, 4(2). Retrieved November 16, 2005 from http://PAREonline.net/getvn.asp?v=4&n=2 .

Sheldon K. & Biddle, B. (1998). Standards, accountability and school reform: Perils and pitfalls. *Teachers College Record 100 (1)*: 164-180

Sloan McCombs, J. & Carroll, S.J. (2005) Who is accountable for education if everybody fails? *Rand Review*, *29*(1), 10-15.

Standard and Poors. (2005). Leveling the playing field: Examining comparative state NAEP performance in Demographic Context. McGraw Hill. Retrieved December 5, 2005: http://www.schoolmatters.com/pdf/naep_comparative_state_performance_schoolmatters.pdf

Stecher, B.M., & Hamilton, L.S. (2002). Putting theory to the test: Systems of "educational accountability" should be held accountable. *Rand Review*, *26*(1), 16-23.

Tanner, D. E. (2001). *Assessing academic achievement.* Boston: Allyn and Bacon.

Taylor,G., Shepard, L., Kinner, F., & Rosenthan, J. (2001). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost?* Boulder: University of Colorado, School of Education, Education and the Public Interest Center, Retrieved June 11, 2005 from http://www/education.colorado.edu/epic/coloradostudiesandreports.htm

U.S. Department of Education. (2002). No Child Left Behind: A Desktop Reference. Retrieved November 22, 2005: http://www.ed.gov/admins/lead/account / Nclbreference / reference.doc.

Authors' Notes

Christopher H. Tienken, Ed.D. is the Assistant Superintendent for Curriculum and Instruction for the Monroe Township School District in Middlesex County, New Jersey. He is a part-time professor at the Rutgers University Graduate School of Education. Correspondence to this author should be sent to goteach1@hotmail.com.

Michael J. Wilson, Ph.D. is an Associate Professor at Western Connecticut University. He was a former school administrator in New Jersey and also worked as statistician/psychometrician for the NJDOE. Correspondence to this author should be sent to wilsonm@wcsu.edu.