The Condition of Assessment of Student Learning in Arizona: 2004

Thomas M. Haladyna
Arizona State University West Campus

Reviewer: Jerome D'Agostino
University of Arizona

Reviewer: Ronald Carriveau
Arizona Department of Education

Reviewer: Thomas Fetsco
Northern Arizona University

Reviewer: Ildiko Laczko-Kerr
Arizona Department of Education

Background

Arizona has many reasons to be concerned about current assessments of student learning. Teachers and parents need valid information to help children learn and to guide them in their careers and life aspirations. State and local policy makers need valid information in order to build strategies to improve student learning. Businesses seek assurance that they will acquire a work force that is well-educated and capable of helping the economy grow. Finally, in accordance with the well-accepted notion of education as a primary means of advancing and benefiting our nation, recent federal legislation, the No Child Left Behind Act of 2001 (NCLB), requires accurate monitoring of student learning in order to measure yearly progress of students, schools, and school districts.

Validity of Interpretations of Student Learning

Any tool used to assess student learning must be validated. In educational testing, "validity" refers to the degree to which an interpretation of a measure of student learning

is trustworthy. The *Standards for Educational and Psychological Testing*[1] provides guidelines for establishing and improving the validity of test scores and of their interpretation and use in order to draw conclusions about student learning. The American Education Research Association (AERA) has published guidelines to be used when a test has high stakes—that is, when it becomes the primary or sole determinant of whether a student advances a grade or earns a diploma, or whether a teacher, school, or school district is to be rewarded or punished in some way.[2]

Which Tests Offer the Most Valid Information About Student Learning?

The above-mentioned test guidelines consistently state that to assess student learning in a valid manner requires the use of several indicators. Such indicators may include course grades; district, state, or national test scores; and teacher evaluations. An assessment tool is considered more likely to be valid when its conclusions are confirmed by other sources of information about student learning. Additionally, to be valid, an assessment also needs the following:

1. A set of content standards that identifies the content that guides instruction.
2. Evidence that all students have had one or more opportunities to learn the content standards.
3. Tests aligned to the content standards.
4. Information about in-school and out-of-school factors believed or known to influence learning, so as to create a context for interpreting student learning.
5. Evidence that the test information is accurate and not corrupted by threats to validity, such as cheating, scoring errors, inappropriate test preparation, lack of motivation, inappropriate test administration, inability of students to read the items on the test, and other conditions that impede valid testing.[3]

Arizona has three achievement measures of student learning: Arizona's Instrument to Measure Standards (AIMS), the Stanford Achievement Test (Stanford 9), and the National Assessment of Educational Progress (NAEP).

Of these three tests, only the AIMS is directly aligned to the state's content standards. For that reason, if the AIMS is administered without threats to validity, it is more likely than the Stanford 9 or the NAEP to be an accurate measure of student learning in Arizona. Unlike the other two tests, however, AIMS offers no normative information by which to make comparisons with students from other states. Also, AIMS is a relatively new test and lacks longitudinal data by which to assess student learning over time. As AIMS is further developed and administered, and as school districts and their teachers adopt and use the content standards, the data from these tests offer the most promise of accurately measuring student achievement in Arizona.

The Stanford 9 is based on a generic, national curriculum that relies on content standards commonly espoused by national societies such as the National Council of Teachers of Mathematics. It is not known to what extent the Stanford 9 is aligned to Arizona's content standards; therefore, claims that it measures what is being taught in Arizona schools are not valid. In order to make such claims, research would be needed on the alignment of the Stanford 9 to Arizona's content standards for all grade levels and subjects. The Stanford 9 provides nationally representative norms by which Arizona students can be compared to other students across the country. There is some disagreement about the validity of these norms, however, as discussed later in this brief.

The NAEP is the United States' only long-term, scientifically-based system for tracking student achievement. The NAEP's interpretations of state and national trends are supported by a wealth of validity studies, and wield an increasing influence on federal and state policies.[4]

Like that of the Stanford 9, the curriculum framework of the NAEP is national in scope and not necessarily aligned to Arizona's content standards; and, as is true for the Stanford 9, an alignment study is needed to see how well the NAEP aligns to Arizona's content standards.

The NAEP is not a conventional test. It consists of blocks of items randomly assigned to blocks of students. These blocks of item responses are used to make estimates of achievement for different groups of students. This strategy economically provides valid interpretations of student achievement based on national standards.

Individual student scores are not obtained, however. Valid state-to-state comparisons can be made only if the national curriculum frameworks are common to all states being compared and if the sampling plan for the NAEP thus provides truly representative estimates of student achievement.

Each of these tests offers some benefits but also has some limitations for assessing student learning. The use of multiple indicators of student achievement provides a stronger, more valid assessment of student learning. Obtaining more measures of student learning is costly, however, making this strategy unlikely when resources are limited.

Recent Developments

Most states have used standardized achievement tests to measure student learning. In recent years, educational reform has motivated a more systematic approach to measure and assess student learning: content standards are created, instruction is based on these content standards, and state-level tests are designed to sample these content standards accurately. The precision involved in aligning curriculum, instruction, and testing has long been a principle of effective teaching.

In 1996, Arizona participated in this reform effort by creating a set of highly regarded content standards and creating Arizona's Instrument to Measure Students (AIMS).[5] In the first four years, AIMS testing became a very controversial matter with the public, particularly because it was intended to influence high school graduation, and field testing indicated a high rate of failure. Currently, the AIMS is undergoing considerable revision, aimed at improving its validity for the purposes explained at the beginning of this brief.

The No Child Left Behind Act of 2001 (NCLB) was intended to improve student achievement. One of its primary goals is to improve the learning of chronically low achievers. With NCLB Act, Congress reauthorized the Elementary and Secondary Education Act (ESEA) as it affects education from kindergarten through high school. Four key principles governing NCLB are (1) accountability for results; (2) an emphasis

on using research to drive innovations; (3) expanded options for parents; and (4) expanded local control and flexibility.

Arizona has recently received and evaluated bids for new testing services, and has selected the TerraNova CTBS test created by the California Test Bureau/McGraw-Hill (http://www.ctb.com/). Consequently, the current state-mandated test that provides norm-referenced interpretations, the Stanford 9, will be replaced by the TerraNova CTBS, which also produces nationally normed test scores.

Although the next section presents results that include the Stanford 9, Arizona will be using a new test in the future for its norm-referenced interpretations. However, the discussion of the Stanford 9 and the findings reported is relevant to the adoption of the TerraNova and its use in Arizona in the future.

Available Data

This section presents information about the achievement of Arizona students in comparison with student achievement in other states, and about changes in the achievement of Arizona students in reading, writing, and mathematics over the past decade. An inquiry about the status of student learning can be framed in two questions:

1. How do Arizona students compare with students in other states in reading, writing, and mathematics?
2. How has the performance of Arizona students in reading, writing, and mathematics changed over the past decade?

Both NAEP and Stanford 9 test results have been used to answer these two questions.

Results and Discussion for NAEP

Table 1 shows the NAEP report of reading achievement for grades four and eight compared to achievement throughout the nation, and for Arizona from 1992 to the present. The results suggest that Arizona students are below the nation's average in

achievement. In fact, Arizona is one of the lowest achieving states in the United States, based on mean performance in all years where reading was assessed in grades four and eight. Of 53 states and jurisdictions tested, scores were lower in three states, higher in 42, and about equal to Arizona's in seven states. Furthermore, the NAEP results show a flat trend in reading achievement both in Arizona and in the nation. The fluctuations in scores are very small and may be due to random error and sampling bias. Thus, reading achievement does not seem to have changed very much, either in Arizona or in the nation at large.

Table 1: Reading Achievement in Arizona and in the Nation for Grades 4 and 8

Grade 4	**1992**	**1994**	**1998**	**2002**	**2003**
Arizona	209	206	206	205	209
U.S.	215	212	213	217	216
Grade 8	**1992**	**1994**	**1998**	**2002**	**2003**
Arizona	*	*	250	257	255
U.S.	*	*	261	263	261

Source: Nation's Report Card, http://nces.ed.gov/nationsreportcard/
*Not Assessed

Table 2 shows writing achievement for just two assessment years. Arizona is well below the nation's average in both grades assessed, grades four and eight.

Table 2: Writing Achievement in Arizona and in the Nation for Grades 4 and 8

Grade 4	1998	2002
Arizona	*	140
U.S.	*	153
Grade 8	**1998**	**2002**
Arizona	143	141
U.S.	148	152

Source: Nation's Report Card, http://nces.ed.gov/nationsreportcard/
* Not Assessed

Table 3 shows results for mathematics achievement. As Table 3 indicates, Arizona is four to five points below the nation's average in all years assessed for grade four. There is a two- to three-point discrepancy for the eighth-graders for a comparable assessment period. Unlike achievement in reading, mathematics achievement shows small gains both in Arizona and in the nation.

Table 3: Mathematics Achievement in Arizona and in the Nation for Grades 4 and 8

Grade 4	1990	1992	1996	2000	2003
Arizona	*	215	218	219	229
U.S.	*	219	222	224	234
Grade 8	**1990**	**1992**	**1996**	**2000**	**2003**
Arizona	260	265	268	269	271
U.S.	262	267	271	272	276

Source: Nation's Report Card, http://nces.ed.gov/nationsreportcard/
*Not Assessed

In the absence of context, any test result reported can lead to faulty causal reasoning or an incorrect conclusion. Context variables should include in-school and out-of-school information that may influence learning. Two examples of out-of-school factors are social capital and intelligence. Students and classes of students with high

social capital and above-average intelligence tend to score very high on achievement tests, and students and classes with low social capital and below-average intelligence tend to score very low on these tests (Nation's Report Card, available at http://nces.ed.gov/nationsreportcard/).

Tables 1, 2, and 3 show that Arizona student achievement scores are substantially below the nation's average. An examination follows regarding conditions that may have accounted for these results, and considers two themes: threats to valid interpretation, and the need for obtaining greater clarity about findings by disaggregating results.

Table 4 shows student reading achievement for grades four and eight in three at-risk categories for Arizona and the nation: students who have an Individual Education Plan (IEP), students classified at Title I and who receive free lunch, and students who are English Language Learners (ELLs). These results are not remarkable, except for the fact that the percentages of at-risk students in Arizona exceed the percentages of at-risk students nationally. Thus, Arizona's national standing appears to be directly related to the differences in proportions to variables associated with chronically low-performing students. Table 4, however, also shows that for the three at-risk categories, Arizona students consistently score a few points lower than their national at-risk counterparts. It is also important to note that the way that states classify students as having an IEP or being an ELL may vary; thus it is difficult to make valid comparisons between one state and the aggregate of all other states if methods for identifying at-risk students differ. Nonetheless, a valid observation can be made that the achievement trend for all students is flat over assessment years considered in this report.

Table 4: Analysis of Student Reading Achievement by At-Risk Factors

		Arizona		U.S.	
Grade 4		**Percentage in Sample**	**Average Score**	**Percentage in Sample**	**Average Score**
IEP	**Yes**	7%	177	9%	185
	No	93%	211	91%	221
Free/Reduced Lunch	**Yes**	52%	194	46%	201
	No	48%	225	54%	229
Title I	**Yes**	38%	187	36%	201
	No	62%	217	64%	226
ELLs	**Yes**	18%	177	8%	186
	No	82%	216	92%	221
Grade 8		**Percentage in Sample**	**Average Score**	**Percentage in Sample**	**Average Score**
IEP	**Yes**	8%	214	5%	222
	No	92%	259	95%	265
Free/Reduced Lunch	**Yes**	43%	241	38%	247
	No	57%	265	62%	271
Title I	**Yes**	26%	240	21%	245
	No	74%	263	79%	267
ELLs	**Yes**	16%	219	5%	222
	No	84%	261	95%	265

Source: Nation's Report Card, http://nces.ed.gov/nationsreportcard/

Results and Discussion of Stanford 9

Table 5 provides Arizona's percentile rank on the Stanford 9 in reading, mathematics, and language for the years 1999 to 2003. These data provide the basis for the summary data and for the discussion that follows. Because of the complexity of the data found in Table 5, an analysis of variance (ANOVA) was performed to examine main effect differences among grade levels, subject matters, and time and among all three first-

order interactions. A technical note on this analysis and the results of that ANOVA are given at the end of this paper.

Table 5: Reading, Mathematics, and Language Percentile Ranks for Grades 2-9 for a Five-year Period

Grade	Subject	1999	2000	2001	2002	2003
2	Reading	50	52	53	57	49
	Mathematics	51	55	57	61	63
	Language	40	43	44	48	57
3	Reading	47	48	50	50	60
	Mathematics	49	52	54	56	59
	Language	51	54	56	57	54
4	Reading	54	54	55	55	52
	Mathematics	54	55	57	58	60
	Language	49	48	50	50	57
5	Reading	51	51	51	53	49
	Mathematics	54	55	57	59	61
	Language	44	45	45	47	54
6	Reading	54	53	54	56	49
	Mathematics	59	60	63	65	66
	Language	44	44	45	47	57
7	Reading	53	52	53	55	59
	Mathematics	55	56	58	60	61
	Language	54	54	55	58	55

Grade	Subject	1999	2000	2001	2002	2003
8	Reading	54	53	55	56	53
	Mathematics	54	56	58	59	61
	Language	49	49	50	52	56
9	Reading	43	43	43	43	44
	Mathematics	57	59	61	62	63
	Language	39	40	41	42	44

Source: Arizona Department of Education.
http://www.ade.state.az.us/ResearchPolicy/SAT9Results/2003/default.asp

Table 6 provides a summary of data from Table 5 dealing with the differences among grade levels across all subjects and years of data collection. The results were statistically significant, and accounted for 13.4 percent of the variance of these percentile ranks. However, note that grade nine produced the lowest scores and the highest standard deviation. The fluctuations of percentile ranks for grades two through eight were very small indeed. The lower score in grade nine may result due to one or more causes. First, ninth graders may no longer see the Stanford 9 as an important test and therefore make less of an effort because the AIMS takes on high-stakes importance. Second, schools may no longer emphasize the Stanford 9, and therefore may be less likely to use test preparation and other measures designed to increase scores. Excepting the ninth-grade results, scores fluctuate very little across grade, regardless of when the assessments occurred.

Table 6. Overall Achievement Scores (in Mean Percentile Ranks) by Grade

	2	3	4	5	6	7	8	9
Mean	52	53	54	52	54	56	54	48
Standard Deviation	6.7	4.0	3.5	5.1	7.4	2.7	3.5	9.1

Source: Arizona Department of Education.
http://www.ade.state.az.us/ResearchPolicy/SAT9Results/2003/default.asp

Table 7 summarizes data from Table 5 by subject matter. As indicated there, scores are significantly higher in mathematics than in reading and writing. Mathematics scores are well above the national average, and reading and writing scores are very close to the national average. These findings contrast sharply with NAEP findings that show Arizona students at grades four and eight to be substantially below the nation's average. The standardized difference between these mean scores is substantial. Since the standard deviation of all percentile ranks is 5.897, mathematics scores are extraordinarily high.

Table 7: Reading, Mathematics, and Writing Scores Across All Years and Grades

	Reading	Mathematics	Writing
Mean	51.65	58.000	49.200
Standard Deviation	4.234	3.749	5.640

Source: Arizona Department of Education
http://www.ade.state.az.us/ResearchPolicy/SAT9Results/2003/default.asp

Table 8 summarizes data from Table 5 for reading, writing, and mathematics across the years 1999 to 2003. The following results are the most noteworthy. Scores increase substantially over the five assessment years. The differences for mathematics and writing show substantial growth. The magnitude of this growth greatly exceeds more than one standard deviation (5.897) of these percentile ranks. The increase in reading scores is less pronounced, and a slight decline is noted in 2003. These results do not correspond with NAEP trends that show no increases of this magnitude in any of the three subject areas.

Through the years, some speculation has been offered about the nature of increases in published test scores since the time of publication of a particular edition of a test such as the Stanford 9. This test score "creep" or "drift" has been suspected to be caused by teaching to the test or excessive or unethical test preparation.[6] Such practices are known to occur, but their extent is a matter of considerable debate and in need of

more research. Nonetheless, the achievement gains on the Stanford 9 in Arizona do not seem to be correlated to NAEP data.

Table 8: Reading, Mathematics, and Writing Scores for Each Year of the Assessment.

	1999	2000	2001	2002	2003
Reading	50.750	50.750	51.750	53.125	51.875
Mathematics	54.125	56.000	58.125	60.000	61.750
Writing	46.250	47.125	49.286	49.111	54.250
All subjects	50.375	51.292	53.297	53.880	55.958

Source: Arizona Department of Education. http://www.ade.state.az.us/standards/science/standard1.asp

This last section of the study focuses on the reading, mathematics, and writing scores of English Language Learners (ELLs) as compared to the state average. Comparisons may also be made with other at-risk groups, but the results will be similar to those reported here. The omission of other groups from this final analysis should not suggest that their situation is of less concern: the situation with the ELL population seems to be the most severe of all at-risk groups.

Table 9 shows the status of ELLs on a normative percentile rank scale for the Stanford 9. ELL percentile ranks for all grades and subject areas are well below the national norm (50th percentile) and Arizona's above-average percentile ranking, as shown in Table 5. The differences are very large in all instances.

Table 9: Reading, Mathematics, and Language Scores for the 2003 ELL Students

	2	3	4	5	6	7	8	9
Reading	30	23	36	23	26	22	25	27
Mathematics	40	35	37	36	41	38	37	41
Language	22	34	28	24	21	24	20	19

Source: Arizona Department of Education.
http://www.ade.state.az.us/researchpolicy/SAT9Results/2003/default.asp

These results for ELLs are more consistent with NAEP results in two ways. First, the differences between at-risk groups and the national and state averages are very large. Second, the problem seems to be persistent at all grade levels reported. It would also be useful to know how this gap has changed over years, but these data are not available at this time. NCLB has as one of its goals reducing the gap between at-risk and other students. To the extent that this goal is met, future results of state assessments should show a reduction in this gap. However, the current data sources do not seem adequate to the task.

Lake Wobegon Effect

The humorist Garrison Keillor coined this phrase to describe a mythical Minnesota town where all the children are above average. John Cannell[7] adopted this phrase to exemplify a peculiar finding that all states were above the national average in their statewide assessments as based on their publisher's standardized achievement test scores. This state of affairs may arise from practices in the schools that inflate test scores without appreciably improving learning. These practices include selecting content that matches the test (teaching to the test), cheating, unethical test preparation, unethical test administration, and sanitizing answer sheets[8]. Gains reported in Arizona based on Stanford 9 scores may not be validated by trend data reported by NAEP for the same time in Arizona, calling into question the validity of the Stanford 9 results. The trend data in the Stanford 9 also show exceptional gains in student learning, but the validity of an interpretation that says Arizona students are learning more does not seem borne out by other data reported here. Moreover, the Stanford 9 is not necessarily the best achievement measure, as it is not aligned to the state's content standards.

Quality of Available Data

Neither of the original questions about the achievement of Arizona students, either over time or in comparison with the achievement of students elsewhere, can be answered with assurance, because no data source provides the quality of information needed. Although AIMS has the greatest potential to track growth in student

achievement in an accurate manner, this testing program is very young and still under development, and it will take several years before valid trend data can be reported. The Stanford 9 has the serious limitation of not being aligned to content standards, yet the Stanford 9 does have the advantage of providing normative information about student achievement. Like the Stanford 9, the NAEP provides normative data, but, also like the Stanford 9, is not linked to Arizona's content standards.

The NAEP data offers a picture of no achievement gains in reading, mathematics, and writing (language) over recent years both nationally and in Arizona. NAEP findings suggest that Arizona students are doing poorly in comparison to other students across the nation.

The Stanford 9 data, however, suggest that Arizona is performing well above the national average for most grades and subject matters. The trend data clearly show that Arizona students are learning more each year in all subjects reported in this study. This anomaly calls into the question the validity of norms for the NAEP or the Stanford 9 tests and the validity of score interpretations for the Stanford 9, where it has been suggested that the Lake Wobegon effect may in fact be operating.

Key Unanswered Policy Questions

As a result of the foregoing discussion and the data reported, several key unanswered policy questions arise:

1. Which data source, Stanford 9 or NAEP, is most valid in indicating the status of Arizona learners in relation to students from other states?
2. Do Arizona students lag behind their national counterparts in reading, writing and mathematics?
3. To what extent are Arizona students, particularly those who are at risk, making adequate annual progress?

Recommendations

Arizona is not yet able to use student achievement data effectively to inform policy makers about student learning. A comprehensive database of student achievement would serve a number of purposes for the state. Those purposes include:

- Implementing accountability measures called for in NCLB and in state mandates to monitor student achievement and to improve conditions for learning in poorly performing schools.
- Monitoring high-school graduation testing and the graduation and drop-outs in the state.
- Evaluating achievement of at-risk and other students, both in comparison with the nation and across years in the state in order to guide reforms to improve student learning.
- Conducting occasional policy studies for various constituencies, agencies, and organizations.
- Evaluating specific programs or reform efforts in the state.
- Informing teachers and other school leaders about student learning to guide and improve instruction.

The Nation's Report Card (http://nces.ed.gov/nationsreportcard/), which represents the compiled results obtained from the NAEP, offers a suitable model for such a comprehensive database. Such a database is publicly accessible and affords users a variety of options to answer a wide range of policy-related questions. The Nation's Report Card includes information on student factors, factors existing outside of schools, instructional content and practices, teacher and school factors, community factors, and state government factors, all of which would increase the usefulness of a state database. The Nation's Report Card also offers expertise and background useful for developing a comprehensive and meaningful state database. The NAEP, on which the Report Card is based, also offers public access and training to researchers and policy analysts who desire to ask more complex questions. Although it yields state-specific data, however, the

NAEP does not provide scores for each student, and its test items are not keyed to the state's content standards; therefore, it is unable to fill Arizona's need for a comprehensive database.

Large gains that have not been validated in Stanford 9 scores over the years suggest anomalies between NAEP and Stanford 9 national norms and results. Very different conclusions follow depending on which data source is used, creating a large discrepancy that make suspect the validity of either conclusion.

Of the tests most commonly used to measure achievement, the AIMS appears to have the greatest potential to offer valid interpretations for a variety of purposes. The state's nationally recognized content standards and the AIMS have received national recognition, and have been widely supported by educators and the public. Improving the assessment system can contribute to the better understanding of what it takes to increase student learning, particularly for those students who are at-risk and have the poorest record of achievement.

It is recommended that:

1. The Arizona Department of Education (ADE) develop a comprehensive database that includes valid indicators of student learning as well as in-school and out-of-school indicators reasonably believed to influence student achievement.
2. The ADE concentrate state assessment resources on the AIMS, using it to provide the achievement information sought by a wide range of constituencies and agencies.
3. The ADE resolve the contradiction between the NAEP and Stanford 9 results if Arizona continues to use both these tests.
4. The ADE validate the large gains over time reported by the Stanford 9, and if the gains are shown to be valid, investigate the factors associated with them: whether teaching and student learning are improving, and whether schools or districts have successful programs that are producing unusually high gains in student learning.

The data in Table 5 provided the basis for answering this three-part research question: is there is a difference in student learning as a function of (1) grade level, (2) subject matter, and (3) time? First order interactions were considered in this analysis but a second order (three-way interaction) was not done for two reasons. First, there was no plausible hypothesis to test for a three-way interaction, and the main effects analysis accounted for most of the variance of these scores.

Normally, one would not use ordinal data for analysis of variance. ANOVA is very strong, however, to minor violations in assumptions; and the effects to be detected were large enough that this violation of the assumption about the scale used in this analysis did not threat the validity of this analysis. Effect sizes were calculated as the percentage of total variance accounted by the independent variable. The results appear in the table below and provide the justification for the discussion of this report in the text.

Independent	SS	Df	MS	F	P	R^2
Grade level	552.561	7	78.937	13.981	<.001	10.5%
Subject matter	1643.073	2	821.536	145.505	<.001	31.2%
Year	496.481	4	124.120	21.983	<.001	9.5%
Subject matter	155.284	8	19.411	3.438	.003	2.9%
Grade level *	73.518	28	2.626	0.465	0.985	-
Grade level*	878.105	14	62.722	11.109	<.001	16.7%
Residual	1450.035	106	13.686			
Total	5250.057	119				

Notes and References

[1] American Educational Research Association, American Psychological Association. National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

[2] American Educational Research Association (2000). Position Statement of the American Educational Research Association Concerning High-stakes Testing in Pre K-12 Education. *Educational Researcher, 29*, 24-25. Retrieved April 12, 2004, from http://www.aera.net/about/policy/stakes.htm

[3] Haladyna, T. M. & Downing, S. M. (in press). Construct-irrelevant Variance: A Threat to the Validity in High-stakes Testing. *Educational Measurement: Issues and Practice*.

[4] The Nation's Report Card (2004, April 6). National Center for Education Statistics. Retrieved April 12, 2004, from http://nces.ed.gov/nationsreportcard/

[5] Cross, R. W., Rebarger, T., Torres, J., & Finn, C. E. (2004). *Grading the Systems: The Guide to State Standards, Tests, and Accountability Policies.* Washington, DC: The Thomas B. Fordham Foundation.

[6] Haladyna, T. M., & Downing, S. M. (in press). Construct-irrelevant variance: A Threat to the Validity in High-stakes Testing. *Educational Measurement: Issues and Practice.*

[7] Cannell, J. J. (1989). *How Public Educators Cheat on Standardized Achievement Tests*. Albuquerque, NM: Friends for Education.

[8] Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising Standardized Achievement Test Scores and The Origins of Test Score Pollution. *Educational Researcher, 20,* 2–7.