

POLICY *brief*

Goldwater Institute

No. 07-04 | June 11, 2007

A Test of Credibility: NAEP versus TerraNova Test Score Results in Arizona

by Matthew Ladner, Vice President of Research, Goldwater Institute; Gregory E. Stone, Assistant Professor, Research and Measurement, The University of Toledo

EXECUTIVE SUMMARY

Does Arizona face a crisis in academic achievement of students in public schools? The answer depends on which test scores you believe to be credible.

The National Center for Education Statistics administers the nation's most respected study of student achievement, known as the National Assessment of Educational Progress (NAEP). NAEP tests representative samples of students in all 50 states and consistently finds that Arizona students score below the national average across subjects and grade levels.

Recently, the state of Arizona created an alternative source of data. The Arizona Department of Education embedded a subset of questions from the TerraNova exam into the AIMS to create a Dual Purpose Assessment (DPA).

The DPA provides a startlingly different picture of the performance of Arizona students than does NAEP. For example, on May 4, 2007, Arizona Superintendent of Public Instruction Tom Horne published a letter in the *Arizona Daily Star* lauding the performance of Arizona's public schools:

The TerraNova test is designed for comparisons among states. Arizonans can be motivated to strive for the top tier of states, knowing their students' test scores are eight percent above the national average, despite being last in resources per student, because of the efforts of Arizona teachers and administrators, and our emphasis on academic rigor in the classroom.

In short, either NAEP or the DPA has a serious credibility issue. The public, parents, and policymakers have a tremendous need for reliable and accurate data concerning public school performance.

The following pages assess this discrepancy through a review of available academic literature on Arizona's DPA and the exam's technical reports. The combined evidence strongly supports the notion that Arizona's DPA contains deep flaws, providing an inaccurate view of school performance in Arizona.

A Test of Credibility: NAEP versus TerraNova Test Score Results in Arizona

by Matthew Ladner, Vice President of Research, Goldwater Institute; Gregory E. Stone, Assistant Professor, Research and Measurement, The University of Toledo

How Do Arizona Students Compare? NAEP versus TerraNova

How do Arizona’s public school students compare with students nationwide? The National Assessment of Educational Progress (NAEP) and the TerraNova exams tell different stories about the academic achievement of Arizona students. Specifically, the TerraNova exam finds that Arizona students are above the national average in every subject (math, reading, and language).¹

Meanwhile, NAEP finds that Arizona public school students are *below* the national average in every subject and at every grade level. Arizona students have been administered 29 different NAEP examinations since 1992—in reading, mathematics, science, and writing—and Arizona’s students have scored below the national average on all 29 of them.

The TerraNova exam finds that Arizona students are above the national average in every subject (math, reading, and language). Meanwhile, NAEP finds that Arizona public school students are below the national average in every subject and at every grade level.

A substantial gap exists in results from these two exams. For example, on the most recent 4th-grade mathematics NAEP, 30 percent more Arizona 4th-graders score “below basic” than the national average. Compared with the national average, 40 percent fewer Arizona students score at the highest level of achievement. Meanwhile, the TerraNova exam results find that Arizona students score 10 percent *above* the national average on mathematics.

These contrary results present a mystery. NAEP and TerraNova tell very different stories about the academic achievement of Arizona public school students. The purpose of this brief will be to resolve this discrepancy.

Background on NAEP

The U.S. Department of Education and its forerunners have administered NAEP exams, also known as the Nation’s Report Card, to samples of students in all states since 1969. The NAEP website describes the project as “the only nationally representative and continuing assessment of what America’s students know and can do in various subject areas.” NAEP provides the common measure to judge the academic achievement of American students across various subject areas. NAEP tracks changes in performance over time and allows for cross-state comparisons of academic achievement.

NAEP periodically administers tests in reading, mathematics, science, writing, U.S. history, civics, geography, and the arts. The Department tests 4th- and 8th-graders more than those at other grade levels, although it does irregularly test 12th grade students. The No Child Left Behind Act required Title I schools selected to participate or risk loss of federal funding.

The National Center for Education Statistics (NCES) in the U.S. Department of Education executes the NAEP project. In 1988, Congress created the National Assessment Governing Board, appointed by the Secretary of Education but independent of the Department, to set policy for NAEP. The Board is responsible for developing the framework and test specifications that serve as the blueprint for the assessments. The Board is a bipartisan group including governors, state legislators, local and state school officials, educators, business representatives, and other members of the public.

NCES does not give NAEP exams to every student in a state but instead to a sample of students. NCES goes to great lengths to ensure a *representative* sample from each state, using a stratified random sampling technique. NAEP selects a representative sample of students by first randomly selecting schools and then selecting the students within those schools who will participate in a given NAEP assessment. Every school has some known chance of being selected for the sample. Within a selected school, all students within a participating grade have an equal chance of being selected for testing.² The credibility of the entire NAEP project rests upon drawing a representative sample of students. Not surprisingly, NCES spends a great deal of effort to sample students in a scientific and representative fashion.

NCES spends a great deal of effort to sample students in a scientific and representative fashion. NAEP is a long-standing national examination with an extremely high reputation.

In short, NAEP is a long-standing national examination with an extremely high reputation. The publishers of *Education Week* surveyed a host of education insiders, analyzed citations in academic journals, and tallied media hits to rank the most influential educational studies and information sources of the last decade. NAEP came in first place in both categories by a wide margin, scoring 100 on a 100-point scale as the most influential education study.³

Background on Arizona's TerraNova Exam

Arizona statutes require that public schools administer both the Arizona Instrument to Measure Standards (AIMS) and a national norm-referenced test. AIMS seeks to measure academic progress against a set of education standards developed for Arizona public schools. The idea behind AIMS and similar tests around the country is for the state to develop a set of academic standards that

describe what students should know in different subjects and grade levels, and then to assess student's proficiency against those standards. Officials established a minimum passing threshold for different subjects and grade levels of AIMS, and the state ranks schools based on AIMS performance.

A national norm-referenced test serves a different purpose. A student cannot pass or fail a national norm-referenced exam, since it compares the performance of an individual student to a representative national sample of students. The norm-referenced test allows comparisons of performance against those of the sample. Many such exams express the performance of students as a national percentile ranking. For example, a student scoring in the 90th percentile means that 90 percent of the students in the sample scored lower than that student.⁴

Arizona schools had previously administered the Stanford 9 exam to fulfill the legislative mandate to provide a national norm-referenced test. In 2005, however, Arizona began using a Dual Purpose Assessment (DPA). The Arizona Department of Education, working with CTB/McGraw-Hill, developed the DPA by embedding a subset of TerraNova exam questions into state AIMS exams.

The Arizona Department of Education held that because 18 states had employed a DPA model and because the increase in testing error fell within what it felt was an acceptable limit, it argued against field-testing the new exam.

The Department designed the DPA to reduce the amount of time students spend taking tests. Figure 1 illustrates the change. Rather than two separate tests, the DPA has three types of test items: purely AIMS questions, dual-use questions (counting for both AIMS and TerraNova), and a subset of questions counting only on the TerraNova exam.

Note that the time-savings results not from giving one test rather than two, but from actually asking fewer total questions. Figure 1 shows the total number of questions on the TerraNova exam falls from 78 questions to 20-30 questions. The Arizona Department of Education, however, held that because 18 states had employed a DPA model and because the increase in testing error fell within what it felt was an acceptable limit, it argued against field-testing the new exam.⁵

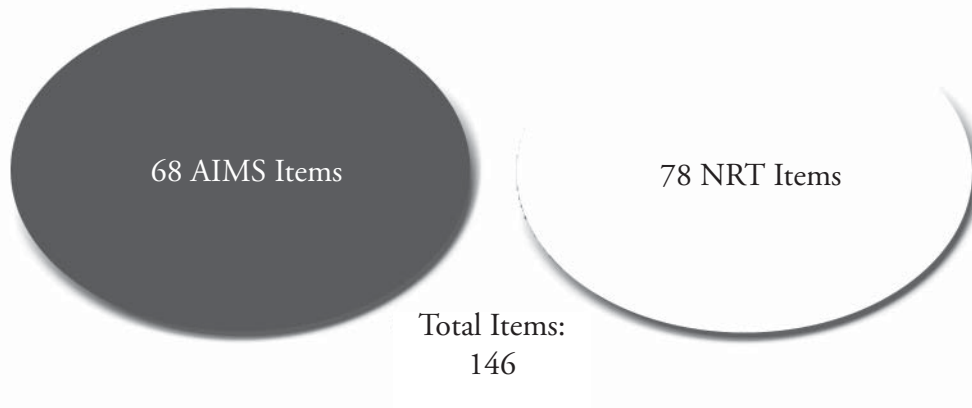
NAEP versus TerraNova in Arizona

Perhaps the most basic method for reconciling the results of TerraNova (which find Arizona above the national average) with those of NAEP (which find Arizona public school students below the national average) is to assess the relative advantages and challenges of teaching Arizona students compared with those other states.

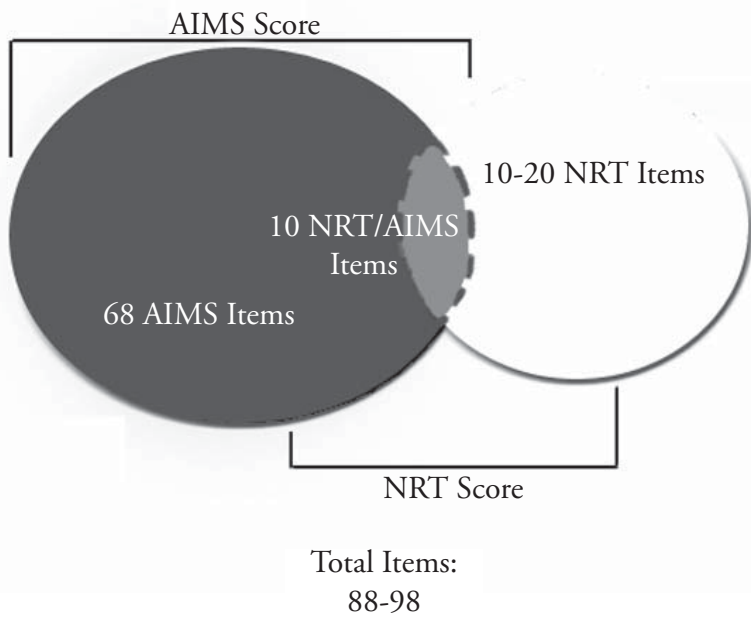
Greene and Forster measured the teachability of students by examining 16

Figure 1: The Arizona Department of Education Norm-Referenced Embedded AIMS, Fifth Grade Mathematics

Separate Assessments



**Format Beginning
2004-2005**



Source: Arizona Department of Education, "The Right Test at the Right Time." Norm-Referenced Test (NRT).

social factors that researchers generally agree influence student test scores. Items included measures of student poverty rates, family income, drug use, family structure, disabilities, low-weight births, parental education levels, the percentage of nonnative English speakers, and so on.

Nationally, Greene and Forster found that the difficulty of educating students has declined, not increased, since the 1970s. In 2001, Arizona ranked as the state with the second-most-difficult-to-educate student population, ahead of only New Mexico (see Figure 2). Greene and Forster found a positive statistical relationship between the amount of school choice in states and their ability to perform better than their rankings on the index would suggest. Consistent with that finding, Arizona scores ranked 30th, in comparison to their teachability ranking of 49th, and generally shows above-average “bang for the buck” on results versus spending.⁶ These latter two findings stand in contrast to the simplistic labeling of Arizona as the “dumbest state” by Quinto Press.⁷

The teachability index makes NAEP’s consistent finding of below-average performance seem much more credible than the TerraNova finding of higher-than-average test scores. The TerraNova exam results would require one to believe that Arizona public schools had not only overachieved but had done so in spectacular fashion.

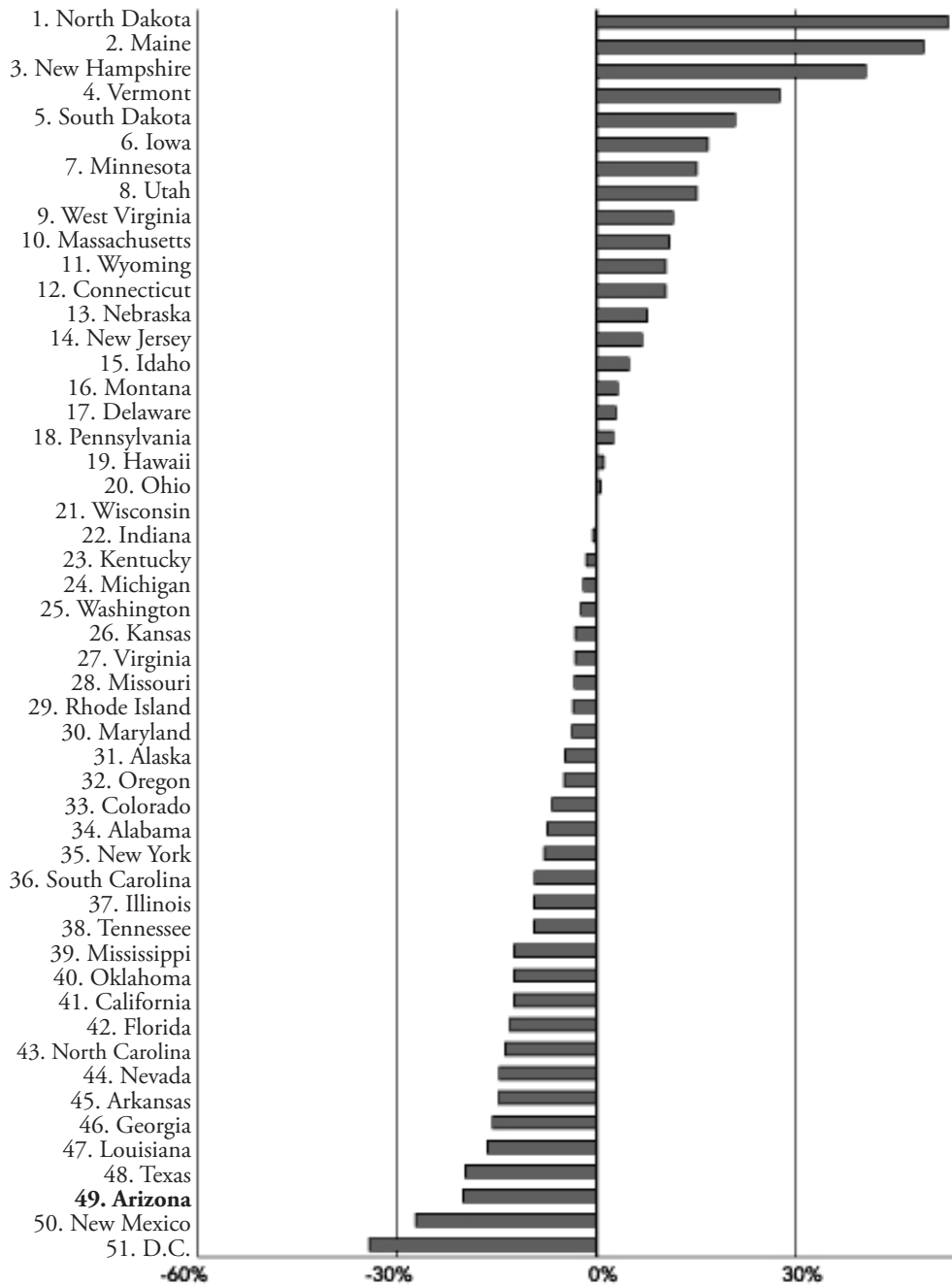
Overall, however, the Manhattan Institute study shows that Arizona has a challenging public school population to educate. But Arizona schools must meet this as a challenge rather than an excuse.

The teachability index makes NAEP’s consistent finding of below-average performance seem much more credible than the TerraNova finding of higher-than-average test scores. The TerraNova exam results would require one to believe that Arizona public schools had not only overachieved but had done so in spectacular fashion. Overall, a degree of overachievement seems plausible, while a huge amount appears suspicious.⁸

Discrepancies between Stanford-9, TerraNova, and NAEP

Scholars have noted discrepancies between trends in the previous national norm-referenced exam (Stanford-9) and trends in Arizona’s NAEP scores. In the years before the test’s replacement, Arizona’s Stanford-9 scores showed a trend toward improvement notably different from that of NAEP. A research team from the University of Arizona led by Professor Darrell Sabers attributed the positive trend in Stanford-9 to the “Lake Woebegon” effect, named after Garrison Keillor’s mythical Minnesota town where “all the women are strong, all the men are good-looking, and all the children are above average.” The “Lake Woebegon” report documented that all 50 states were testing above the national average in elementary achievement, concluding that America’s public schools testing programs were corrupt.⁹

Figure 2: Teachability Index



Source: Manhattan Institute, The Teachability Index: Can Disadvantaged Students Learn?

The Sabers team revealed that the most common reason for the Lake Wobegon effect involves teaching to the test:

The most publicized reason for the increase in test scores within a state is the possibility that teachers tailor their curriculum and focus to reflect the test's weighting of objectives. Along with teaching of testing skills, the conformity of curriculum to testing is often referred to as teaching to the test.... At the extreme end, teaching *to* the test can become teaching *the* test, where teachers learn the specific test items that assess various objectives, and teach those test items.¹⁰

Essentially, test scores have an upward bias when teachers become more familiar with the exam. Teachers can shape their curriculum around what they know to be on the test, making it unclear as to whether score gains relate to learning gains or to teaching to the test. In the worst-case scenario, teachers learn the *actual test items* and teach students the individual questions rather than the standards, thus teaching the test rather than teaching to the test.

Essentially, test scores have an upward bias when teachers become more familiar with the exam. Teachers can shape their curriculum around what they know to be on the test, making it unclear as to whether score gains relate to learning gains or to teaching to the test.

Some argue that “teaching to the test” is exactly the point in a standards based exam like AIMS, but in fact, the idea is to teach to the standards, not to the test. The state has developed grade-level standards, and it tests students on their achievement against those standards. Foreknowledge of test items can bias the results of the exam in favor of those students accessing the exam.

The Sabers team, in fact, notes that items are publicly available for both teachers and students.

[Test] frameworks (or blueprints) and released/sample test items for Arizona's Instrument to Measure Standards (AIMS), TN, and the Dual-Purpose Assessment (DPA) are publicly available on the Arizona Department of Education (ADE) website. CTB-McGraw Hill's website also has information about the TN blueprints as well as a plethora of “teaching tools” marketed to help teachers teach to the test.

Obviously, the availability of such “teaching tools” could explain some of the variation between Arizona's NAEP and TerraNova scores. The Sabers team addresses this issue directly:

NAEP has been considered a more valid index of state achievement because of its low-stakes nature. Teachers and schools may have felt less pressure to improve performance on NAEP as compared to high-stakes tests such as AIMS. Although states may feel pressure to show relative improvement on

NAEP, documentation of states encouraging districts or schools to improve does not exist. It is not expected that teachers teach to a test that is not administered every year or in the same school every testing cycle. NAEP frameworks are much more general and descriptive, and less instructional, compared to the other websites and supplemental materials.

Ultimately, the Sabers team recommends dispensing with the TerraNova and using NAEP to compare Arizona's academic achievement with that of other states:

National standardized norm-referenced tests (NRTs) are intended to provide information for a state about its students'/schools'/districts' performance relative to a national norm. However, because the validity of the information NRTs generate is in question, its usefulness is also in question. Because NAEP is considered a more valid source of information about state assessment trends, NRTs that do not show trends similar to NAEP do not serve their intended purpose.

Has Arizona Successfully Embedded TerraNova into AIMS?

Another set of questions about Arizona's DPA involves the embedding process. The available evidence strongly suggests that Arizona's TerraNova scores are inflated, and that "teaching to the test" is a possible source of that inflation.

This would especially seem to be the case for the dual-purpose items, that subset of DPA questions counting for both AIMS and TerraNova scores. In Figure 1, dual-purpose items represent between one-third and one-half of the diminished set of questions making up the TerraNova portion. Irrespective of how much "teaching to the test" Arizona schools have been engaging in regarding TerraNova (instructors making use of TerraNova "teaching tools" such as test blueprints and sample questions), the process of "teaching to the test" on AIMS will inevitably inflate scores on dual-purpose items, and thus on TerraNova scores as well. Arizona's TerraNova scores suffer from three separate possible sources of score inflation: teaching to the test for TerraNova, teaching to the test for AIMS, and teaching the test for both exams.

Even more fundamental questions surround the dual-purpose exam. The Goldwater Institute asked Gregory Stone, a psychometrician (psychometrics is branch of psychology dealing with the design, administration, and interpretation of quantitative tests for the measurement of psychological variables such as intelligence and aptitude), to review 2005 and 2006 *Arizona Instrument to Measure*

Arizona's TerraNova scores suffer from three separate possible sources of score inflation: teaching to the test for TerraNova, teaching to the test for AIMS, and teaching the test for both exams.

Standards Reports published by CTB/McGraw-Hill and associated documents related to the creation of the Dual Purpose Exam.¹¹

Stone reviewed the reports for psychometric soundness and legal defensibility. During the review, Stone assessed two substantive questions:

- (1) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) a valid and reliable proxy for the standard Terra Nova examination, such that outcomes are comparable nationally or across states?
- (2) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) comparable to the National Assessment of Educational Process (NAEP) examination in accomplishing its stated objectives?

This report presents Stone's full memo responding to these questions as Appendix A. Stone's analysis raises fundamental questions about the validity of Arizona's dual-purpose exam, finding deficiencies that "are serious enough to render the outcome of the examination untenable."

A large number of questions on the DPA were not measuring what they were intended to measure. This is not uncommon in testing, and it is usually corrected by field-testing exams. The Arizona Department of Education, however, did not field-test the DPA.

Broadly, Stone raises two fundamental objections. First, he asserts that the designers of the Arizona DPA display a misunderstanding of the difference between norm- and criterion-based items and norm- and criterion-based exams. Stone says that test items do not intrinsically possess properties of being either norm or criterion referenced, but rather it is the scoring and use of such items within the context of an exam that make a test either norm or criterion based. He adds that while both normed and criterion scores may be generated from a single examination, such a practice is not generally recommended and that "the comparability of scores from such an examination is questionable."

Stone further finds that the technical justifications for the DPA provided were very much wanting. He writes, "While the authors report few misfits in their narrative, even a peripheral review of the fit statistics provided in the included tables indicates that there are a great number of strands that do not match the construct. This speaks not only to an inability for comparison, but to the problems associated with use of the 3-PL IRT [three-parameter item response theory] model for constructing 'measures.'"¹²

In layman's terms, Stone found a large number of questions on the DPA that were not measuring what they were intended to measure. This is not uncommon in testing, and it is usually corrected by field-testing exams. The Arizona Department of Education, however, did not field-test the DPA. The smaller number of testing items leaves little margin for error in the form of bad questions.

Stone directly addresses the Arizona Department of Education's assertion that the experience in other states with DPA exams finds a 90 percent correlation between DPA norm-referenced results and actual norm-referenced examinations. Stone asserts:

By mixing methods ... the reported item difficulties do indeed correlate well. The 3-PL IRT model is simply adjusting to the norms of the state and producing inevitably inflated correlations.... This *fait accompli* provides a false notion of comparability that was inevitable from the beginning. Such statistical manipulation is designed simply to produce the results desired from the outset. In short, the process is pseudo-science and meaningless. No practitioner of measurement, including Dr. Frederick Lord cited liberally throughout this section, would have supported such a manipulation of data.

Ultimately, Stone answered the two questions put to him about Arizona's DPA:

(1) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) a valid and reliable proxy for the standard Terra Nova examination, such that outcomes are comparable nationally or across states?

Based upon a lack of critical evidence of predictive validity (including an absence of predictive or comparison studies) and the apparent failure to observe what may be considered to be standard measurement protocol, the answer is unequivocally no. While the DPA may be a reasonable examination for some purpose, that purpose specified as the rationale for its construction is without foundation.

(2) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) comparable to the National Assessment of Educational Process (NAEP) examination in accomplishing its stated objectives?

No reasonable comparisons between the DPA and NAEP or any other nationally normed assessments are possible within the framework presented, given the above noted failures. There is the potential for comparison but only through the use of reasonable equating procedures, none of which were presented in the current reports.

Stone's final conclusion:

CTB/McGraw Hill is an exceptionally well-regarded assessment group and I

"No reasonable comparisons between the DPA and NAEP or any other nationally normed assessments are possible within the framework presented, given the above noted failures."

am surprised at the fundamental errors made throughout the defined process. Indeed, the terminology and procedures employed in the [Arizona Instrument to Measure Standards Reports] violate measurement practices outlined in several of the textbooks published by this same company in the field of measurement. In my reviews, I have seen these errors in only one place—state sponsored reports. It is suggested, therefore, that states are given precisely the information they wish to see, rather than the information that would be considered standard-of-practice in the professional field. This too is troubling.

Conclusions and Recommendations

Arizona's DPA contains deep flaws, providing an unrealistically rosy picture of how Arizona students compare with their peers nationwide. The exam completely lacks the hard-won credibility of the NAEP exam.

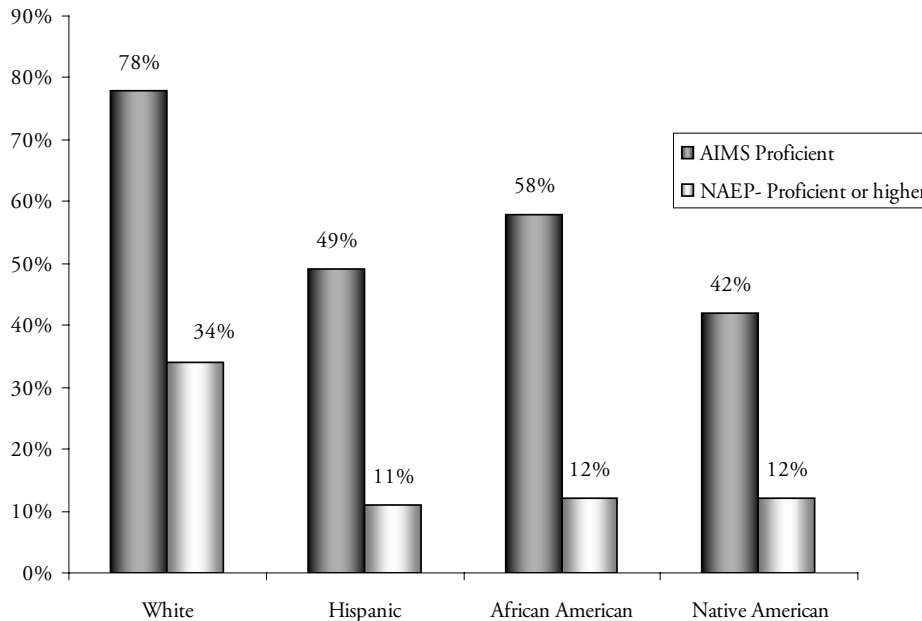
Arizona's DPA contains deep flaws, providing an unrealistically rosy picture of how Arizona students compare with their peers nationwide. The exam completely lacks the hard-won credibility of the NAEP exam.

The Sabers team's recommendation of doing away with TerraNova altogether and solely using NAEP for purposes of comparing Arizona students with students nationwide has some appeal. Under the guidance of Arizona Department of Education, we have witnessed both the "Lake Woebegon" effect of Stanford-9 and the indefensible fielding of a deeply flawed DPA. NCES simply has a much better record of accomplishment in fielding credible national norm-referenced exams than does the Arizona Department of Education thus far.

Unfortunately, the Sabers recommendation contains some serious drawbacks. Only a representative sample of students within each state takes the NAEP exam. NAEP therefore does not provide school- or typically even district-level data. It might be possible to rely on AIMS for purposes of cross-school comparisons, but unfortunately, a lowering of passing standards (specifically the cut scores of the exam) have limited the utility of the AIMS exam for transparency purposes.¹³ The State Board of Education has made AIMS easier to pass, apparently in reaction to federal sanctions under NCLB, and may experience further lowering of standards in the future. On the current course we have set with NCLB, as federally required passing rates approach 100 percent in 2014, minimum passing thresholds are approaching zero.¹⁴

Figure 3 presents the percentage of students scoring "proficient" on AIMS compared with those scoring proficient on NAEP. In short, it would not be wise to put all of the school-level transparency eggs in the AIMS basket, especially since NCLB will place additional pressure on the lowering of AIMS cut scores in a few years.

Figure 3: Eighth Grade Reading Proficiency — NAEP versus AIMS, 2005



Source: Arizona Department of Education; U.S. Department of Education.

Policymakers should embrace transparency as their primary goal of state testing regimes. To enjoy the benefits of transparency, Arizonans need a reliable, low-stakes, and unbiased national norm-referenced exam. Lawmakers cannot craft informed policy decisions in the absence of transparency, and citizens require transparency to participate meaningfully in the governance of public schools. Specifically, parents need transparency if they are to choose schools that match the needs of their children. Arizona’s extensive system of public school choice—including open-enrollment transfers between district schools, magnet schools, and charter schools—all require reliable data to allow parents to make informed choices.

The Terra-Nova portion of the DPA simply does not provide reliable norm-referenced data. Whether one can reliably use Arizona’s TerraNova for rough-and-ready comparisons between schools in Arizona remains unclear. The DPA certainly does not represent a valid norm-referenced exam, and there undoubtedly is a need for such an exam.

Arizona should dispense with the DPA and return to the administration of a wholly separate national norm-referenced exam. At the same time, the Arizona Department of Education should take care to avoid the sort of problems that surrounded the administration of both Stanford-9 and the DPA TerraNova.

Arizonans need a reliable, low-stakes, and unbiased national norm-referenced exam. The Terra-Nova portion of the DPA simply does not provide reliable norm-referenced data.

APPENDIX A



Memorandum

To: Dr. Matthew Ladner
Vice-President of Research
Goldwater Institute

From: Dr. Gregory E. Stone
Managing Partner

Date: 23 February 2007

Re: Arizona DPA Program Review

Thank you for the opportunity to review and evaluate the State of Arizona's educational assessment program. Per your request, I have carefully reviewed both the 2005 and 2006 Arizona Instrument to Measure Standards Reports published by CTB/McGraw Hill. I reviewed the reports for psychometric soundness and legal defensibility. During the review, I was asked to address two substantive questions:

(1) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) a valid and reliable proxy for the standard Terra Nova examination, such that outcomes are comparable nationally or across states?

(2) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) comparable to the National Assessment of Educational Process (NAEP) examination in accomplishing its stated objectives?

In addition to reviewing the technical reports, I reviewed the PowerPoint presentation entitled "The Right Test at the Right Time".

What follows are my conclusions.

Overview of the Arizona Dual Purpose Assessment (DPA)

In 2005, the State of Arizona proposed merging the state developed AIMS (Arizona Instrument to Measure Standards) with a selected group of items from

CTB/McGraw Hill's Terra Nova assessment. Scores presented from the AIMS examination are traditionally criterion-referenced whilst those associated with the Terra Nova examinations are traditionally norm-referenced using a national norm group. According to the 2005 AIMS Technical Report, merging items from both instruments would create an effective "*dual purpose assessment*" (DPA) with the benefit of a reduced number of items and a reduction of administration time.

The authors propose that scores from the new DPA would be comparable to that of the traditional Terra Nova examination. While the goal of a blended examination is not without merit, the 2005 AIMS Technical Report and subsequent 2006 report suggest the use of seriously flawed methodological principles. The deficiencies are serious enough to render the outcome of the examination untenable.

Norm and Criterion-Confusion

In the 2005 report (see page 6) the authors discuss *items* as being norm and/or criterion-referenced. The proposed notion is that if norm and criterion-referenced items are blended, the scores on the blended examination would be comparable to the normed Terra-Nova examination. There is a substantive and fundamental problem with this notion. Items are *neither* norm or criterion-referenced. Norm and criterion referencing refers not to items but to the scores on the examinations, or more precisely, the use of the resultant measures.

Norm-referenced examination results place individuals within a distribution allowing for person-to-person performance comparisons either within a norm group or, as is more usual for national assessments, to a smaller, sample norm group designed to represent the desired population. Within achievement testing, this group typically mirrors the population of the country (e.g. the United States).

Criterion-referenced examination results attempt to assess the level of content mastery against a defined construct. Criterion-referenced measures are generally used in high-stakes licensure and certification testing, and have garnered considerable support for use within states under NCLB minimal standard requirements.

The author's approach is fundamentally flawed in two ways:

- (1) Whether used on a normative or criterion-referenced assessment, the items themselves always reflect content in a way that will best lead to the desired outcome (achievement or aptitude, normative or criterion). The items themselves do *not* retain this norm or criterion *quality* once

removed, as the quality does not reflect the characteristic of the item but rather the use and scoring associated with the examination. Therefore, this mixture of items from the AIMS examination with those from the Terra-Nova does not in any way define a so-called “*dual-purpose*” examination. While both normed and criterion scores *may* be generated from a single examination, such a practice is not generally recommended. Furthermore, the comparability of scores from such an examination is questionable.

(2) Student performance on the combined examination may not be used as a proxy for performance on either the full AIMS or full Terra-Nova examination unless predictive validity tests have been performed. No evidence of predictive validity was presented in either report and because of their extensive level of explanation, it must be concluded that none exists. While the DPA may be a reasonable examination, there is no support to conclude it is more or less reliable, and hence more or less valid, than the individual examinations. Indeed, since validity and reliability estimates were better explored within each of the separate examinations, it is unclear what benefit apart from time savings was gained by the merger.

Terra Nova Calibrations Are not Accurately Used

As indicated, the Terra-Nova examination is norm-referenced, a practice often used within educational assessment. To that end, item calibrations (i.e. estimates of the difficulty of items) were calculated using a three-parameter item response theory model (3-PL IRT). This differs from the calibration of items on the AIMS examination that makes use of the Rasch model.

3-PL IRT models add parameters called “discrimination” and “pseudo-guessing” to what, superficially, looks like the Rasch model. (Hence, the Rasch model is often inappropriately referred to as the 1-PL IRT model). The two additional parameters act in a way that most accurately reflects the item difficulty associated with the *specific group of students* within the norm sample used (approximately 1,900 students nationwide). 3-PL IRT models are well suited to norm-referenced testing as they provide excellent descriptions of item difficulty *within a specified group*. It is this **sample dependency** that makes their use outside of that realm ill advised. The item difficulty approximations in 3-PL IRT are not measures, but descriptive difficulties associated *only* with the norm group. While these item difficulties may work reasonably well when the characteristics of the norm sample group are *identical* to that of the group being compared, they fail to define a construct. It is most unlikely that the student population within the State of Arizona mirrors the national norm group. There is no evidence that any comparisons have been made. Lack of comparison is a troubling matter.

In the instance of the DPA, the authors propose to use the 3-PL IRT item difficulty estimates obtained from the Terra-Nova nationally normed assessment to anchor the full DPA, therefore claiming to establish a singular ruler through which performance may be compared. This is not possible. While the Rasch model, used for the AIMS examination, produces sample-free measures and *could* have been used to anchor the Terra-Nova to the AIMS, the reverse is not possible. The mirroring of content from the Terra-Nova to the DPA while important does not enable direct comparison.

To observe the failure of this approach, we need only review the significant number of strands that misfit. While the authors report few misfits in their narrative, even a peripheral review of the fit statistics provided in the included tables indicates that there are a great number of strands that do not match the construct. This speaks not only to an inability for comparison, but to the problems associated with use of the 3-PL IRT model for constructing “measures”. This issue will also be discussed in the next section.

The more appropriate models for comparison would be through equating and through predictive validity assessments. If the AIMS and Terra-Nova assessments predict similar mastery and/or performance, such should be established through equating studies. Modern equating models allow for comparisons when no items are in common. Only in this way can the tests be useful in direct comparison.

Invalid Equating and Unjustified Support

The PowerPoint presentation states that “*Research by these states [those employing a DPA examination process] and their test contractor has shown a .90 correlation between core items and full batteries.*” There is no clear evidence to support this contention from the two extensive research reports.

Section 9 of the AIMS Technical Report extensively investigates the reliability and validity of the examination. The investigations are impressive in many ways, and detailed to a level almost excessive. Beginning in sub-section 9.2.2 the authors describe the important process by which the embedded Terra Nova item difficulties and performance parameters from the national norm group were compared to the performance indicators associated with the students in the State of Arizona. The findings were remarkable. Using raw score values (page 229) the authors demonstrate the comparability of scores. The correlations are remarkably high for an educational assessment using such a small number of anchored items. In fact, however, the correlations are spurious and unremarkable because of flawed methodology.

By mixing methods (the Rasch model and the 3-PL IRT model discussed in the second section) the reported item difficulties do indeed correlate well. The 3-PL IRT model is simply adjusting to the norms of the state and producing inevitably inflated correlations. Had the comparisons been undertaken using Rasch measures alone, without the flawed, normative “adjustments” for discrimination and pseudo-guessing, it is likely the result would have been quite different. This *fait accompli* provides a false notion of comparability that was inevitable from the beginning. Such statistical manipulation is designed simply to produce the results desired from the outset. In short, the process is pseudo-science and meaningless. No practitioner of measurement, including Dr. Frederick Lord cited liberally throughout this section, would have supported such a manipulation of data.

Conclusions

Returning to the questions stated at the outset of this memorandum, my conclusions are therefore as follows:

(1) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) a valid and reliable proxy for the standard Terra Nova examination, such that outcomes are comparable nationally or across states?

Based upon a lack of critical evidence of predictive validity (including an absence of predictive or comparison studies) and the apparent failure to observe what may be considered to be standard measurement protocol, the answer is unequivocally no. While the DPA may be a reasonable examination for some purpose, that purpose specified as the rationale for its construction is without foundation.

(2) Is the reformulated State of Arizona Dual Purpose Assessment (DPA) comparable to the National Assessment of Educational Process (NAEP) examination in accomplishing its stated objectives?

No reasonable comparisons between the DPA and NAEP or any other nationally normed assessments are possible within the framework presented, given the above noted failures. There is the potential for comparison but only through the use of reasonable equating procedures, none of which were presented in the current reports.

CTB/McGraw Hill is an exceptionally well-regarded assessment group and I am surprised at the fundamental errors made throughout the defined process. Indeed, the terminology and procedures employed in this paper violate measurement practices outlined in several of the textbooks published by this same company in the field of measurement. In my reviews, I have seen these errors in only one place – state sponsored reports. It is suggested, therefore, that states

June 11, 2007

are given precisely the information they wish to see, rather than the information that would be considered standard-of-practice in the professional field. This too is troubling.

NOTES

1. Tom Horne, "A Message from Superintendent Tom Horne: 2005-2006 State Report Card," Arizona Department of Education, 2007, p. 2.
2. New Hampshire Department of Education, "Frequently Asked Questions About NAEP Sampling," <http://www.ed.state.nh.us/education/doe/organization/curriculum/Assessment/NAEPfaqs.htm>.
3. See Michael J. Petrilli, "The Key to Research Influence: Quality Data and Analysis Matters After All," *Education Next*, Spring 2007, <http://www.hoover.org/publications/ednext/6080891.html>.
4. Pearson Educational Measurement, "What Is a Percentile rank, and What Does It Mean?" http://www.pearsonedmeasurement.com/research/faq_2c.htm.
5. Arizona Department of Education, "The Right Test at the Right Time," briefing.
6. Note, however, that the Manhattan Institute made use of the official spending statistics for Arizona that have been shown to substantially underestimate public school spending in the state.
7. See John Fahreny, "Arizona Ranked Dumbest State," Arizona Republic, October 18, 2006, <http://www.azcentral.com/arizonarepublic/news/articles/1018dumb1018.html>.
8. Note that for every student participating in choice programs since their creation in 1994, more than three students have come into the public school system through enrollment growth statewide. The actual competitive pressure for Arizona public schools to improve under existing choice programs therefore is real but limited.
9. J.J. Cannell, *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average*, (Daniels, W.Va.: Friends for Education, 1987).
10. Darrel Sabers and Sonya Powers, "The Condition of Assessment of Student Learning in Arizona: 2005," in David R. Garcia and Alex Molnar, eds., *The Condition of Pre-K-12 Education in Arizona: 2005*, (Tempe, Ariz.: Arizona Education Policy Initiative, September 2005) http://epsu.asu.edu/aepi/AEPI_2005_annual_report.htm. Emphasis in original.
11. Gregory Stone is based at the University of Toledo and Managing Partner of MetriKs Amérique, a private, psychometric, statistical, and programmatic evaluation partnership working within the fields of education, the health professions, the not-for-profit sector, and private industry.
12. Within Rasch, the assessment of how closely the performance of an item matches our expectations is evaluated using two fit statistics: the mean-square fit and z-standardized fit. Both measures of fit are sample dependent, and as such, across-the-board ranges are often inappropriate. Mean-square fit is often cited and used by researchers, and it is the fit statistic chosen for use in the AIMS reports. While it is true that Wright and Linacre (see Linacre, J.M. & Wright, B.D, Chi-Square Fit Statistics, *Rasch Measurement Transactions*, 8:2, (1994): 350; and Wright, B.D. & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8:30, pg. 370) do suggest that a range of "good fit" falls between mean-square values of 0.7 and 1.3, it would be misleading to suggest that this range is absolute. Bond and Fox (1997) cite a mean-square range of 0.6-1.4, and many high-stakes testing authorities employing the Rasch model use a strict 0.8-1.2 range (see Bond, T. and Fox, C. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd Ed. New Jersey: Lawrence Erlbaum.) There simply is no clear agreement on appropriate, absolute mean-square fit ranges.

In addition to the lack of an absolute range, the mean-square fit statistic suffers from another problem associated with sample size. Just as reliability increases with sample size, even if the instrument remains the same, so mean-square fit tends toward the mean (1.0) when the sample size increases. The arbitrary ranges work well with sample sizes of a few hundred but do not hold with sample sizes in the several hundred or thousands. If the authors of the AIMS report were serious about using mean-square fit to assess item performance, they would be well served to contact Richard Smith, a leading researcher in this field. Smith has devised a simple equation to calculate the appropriate range of mean-square fit values based on the sample size.

The second Rasch fit statistic, the z-standardized measure is a z-score transformation (although recent literature has begun to consider the measure more akin to a t-score) placed on a normal distribution. While the z-transformation fit is also sample dependent, the range of good fit remains unchanged, thus making interpretation somewhat easier. While the mean-square tends to underreport misfit, the z-transformation tends to slightly overreport misfit.

Because the authors of the AIMS report used fit to assess performance of the items prior to anchoring/equating, the items used should have been carefully selected to fit well, or the resulting comparisons are likely to prove problematic. The AIMS assessment of item fit within the presented Rasch analysis is unsatisfactory because it fails to more holistically assess fit and choose instead to make use of arbitrary ranges more reasonable for small samples. A better assessment would have included a review of both statistics or the proper calculation of an appropriate range. In the AIMS report, the authors suggested that several strands misfit. It is quite possible that this misfit occurred because poorly performing items were allowed to remain on the examinations. Although this suggestion is speculative, if it were found to be correct, it would provide a possible explanation.

13. See Matthew Ladner, "Aimless AIMS," Goldwater Institute Today's News, June 1, 2006, <http://www.goldwaterinstitute.org/aboutus/ArticleView.aspx?id=1012>; and Matthew Ladner, "May the Farce Be with You," Goldwater Institute Today's News, November 2, 2006, <http://www.goldwaterinstitute.org/aboutus/ArticleView.aspx?id=1186>.
14. See Matthew Ladner, "Please Save the Baby in the Bathwater President Bush!," Edpresso (weblog), January 25, 2007, http://www.edspresso.com/2007/01/please_save_the_baby_in_the_ba.htm.

The Goldwater Institute

The Goldwater Institute was established in 1988 as an independent, non-partisan public policy research organization. Through policy studies and community outreach, the Goldwater Institute broadens public policy discussions to allow consideration of policies consistent with the founding principles Senator Barry Goldwater championed—limited government, economic freedom, and individual responsibility. The Goldwater Institute does not retain lobbyists, engage in partisan political activity, or support or oppose specific legislation, but adheres to its educational mission to help policymakers and citizens better understand the consequences of government policies. Consistent with a belief in limited government, the Goldwater Institute is supported entirely by the generosity of its members.

Guaranteed Research

The Goldwater Institute is committed to accurate research. The Institute guarantees that all original factual data are true and correct to the best of our knowledge and that information attributed to other sources is accurately represented. If the accuracy of any material fact or reference to an independent source is questioned and brought to the Institute's attention with supporting evidence, the Institute will respond in writing. If an error exists, it will be noted on the Goldwater Institute website and in all subsequent distribution of the publication, which constitutes the complete and final remedy under this guarantee.

