

RESPONSE TO GOLDWATER INSTITUTE
POLICY BRIEF OF JUNE 11, 2007
By Tom Horne

As state superintendent of schools, I know Arizona education statistics in excruciating detail. Last June, the Goldwater Institute put out a policy brief on Arizona education statistics which is false and misleading, and which was followed by a number of blast emails and postcards from the Goldwater Institute that have also been false and misleading. The following response puts the facts before the public.

1. Goldwater Contention No. 1:

The *Policy Brief* of June 11, 2007, starts as follows:

A Test of Credibility: NAEP versus Terra Nova Test Score Results in Arizona by Matthew Ladner...

The National Assessment of Educational Progress (NAEP) and the Terra Nova 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).

Meanwhile, NAEP finds that Arizona public school students are *below* the national average in every subject and at every grade level.

. . .

These contrary results present a mystery. [The *Policy Brief* goes on to argue that NAEP, which shows Arizona students doing poorly is a more reliable test than *TerraNova*, which shows Arizona students doing well.]

Goldwater Institute Policy Brief, p. 2.

Rebuttal.

There are two kinds of tests. The first is a standards-based test (also known as a “criterion-referenced test”), which measures students not against each other, but how well each student has mastered the state’s standards. In Arizona this is the AIMS (Arizona Instrument to Measure Standards) test. Students are not compared. All students could pass; all students could fail. It measures how well each student has mastered the Arizona Standards, which were developed by Arizona teachers. Each question on the AIMS test is a measurement of a performance objective on the standards, which can be reviewed at www.ade.az.gov/standards/contentstandards.asp.

The second type of test is a norm-referenced test. Here students are compared to each other, or, in education language, are measured against a “norm.” The *TerraNova* test is one of the three tests in the United States that is used for this purpose. It has been normed in all 50 states. It is produced by CTB/McGraw-Hill, a major and highly respected publisher.

There was a time when almost all testing was norm-referenced testing. Then, under No Child Left Behind, the federal government required that all testing for federal purposes be standards-based (also known as “criterion-referenced”) testing. A number of states dropped their norm-referenced tests. The Arizona legislature decided to keep ours, in part, because the business community wanted to know how Arizona compared with other states. We spend about \$2 million a year on this test, which tells Arizona parents how our state compares with other states; it also tells how the individual school and the individual student compares. The *TerraNova* test is given to 600,000 students, all students between grades 2 and 9.

The NAEP test, by contrast, is given to 6,000 students, not 600,000. Most experts agree that the number of students affects the credibility of the results. The NAEP does not tell parents how their students, or even their schools, compare. It gives only statewide figures, based on the small sample of students. Unlike the *TerraNova*, it is not given under the same circumstances from place to place, and therefore does not provide a proper comparison. For example, in Texas and in a number of other states, English language learners take the math test in Spanish. In fact, the test booklet has the English questions on one side of the page and the same questions in Spanish on the other side of the page. Because of an initiative based by the voters, we cannot do that in Arizona: everyone must take the test in English. If our English language learners could take the test in Spanish, our scores would be higher. You cannot make a valid comparison from tests that are given under different circumstances in different places.

The NAEP is not a norm-referenced test. It is a standards-based test, for which the standards are the NAEP “framework,” just as the AIMS is a standard-based test for which the standards are the Arizona Standards. If Arizona had adopted the NAEP “framework” as the Arizona standards, that would make NAEP a better test of Arizona student achievement. But I inherited a set of math standards that were not aligned to NAEP. We just completed a study that indicates that 42 percent of the grade four NAEP framework is not covered, or “weakly” covered in the Arizona Standards, and that the same is true of 36 percent of the grade eight NAEP framework. It is easily predictable that students will not do well on a test where they have not been taught how to do 42 percent of the problems with which they are presented. This is the year to revise the Arizona math standards in our cycle of revisions, and one of my goals is to bring the Arizona math standards much more into alignment with the NAEP “framework,” in part, because of the public relations aspect of how well Arizona does on the NAEP.

2. Contention No. 2:

...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 Department designed the DPA to reduce the amount of time students spend taking tests...

. . .

Note that the time-savings results not from giving one test rather than two, but from actually asking fewer total 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 is felt was an acceptable limit, it argued against field-testing the new exam.

Id., p. 4.

Rebuttal:

All questions are field tested.

The Goldwater Institute (“GI”) deliberately uses vague language; “because the increase in testing error fell within what it felt was an acceptable limit.” In fact, what an extensive study showed was that the results of the Dual Purpose Assessment was plus or minus one percent of the results of giving a separate norm-referenced test.

The phrase “embedding a subset of *TerraNova* exam questions into state AIMS exams” is also misleading. Here is how the Dual Purpose Assessment was developed.

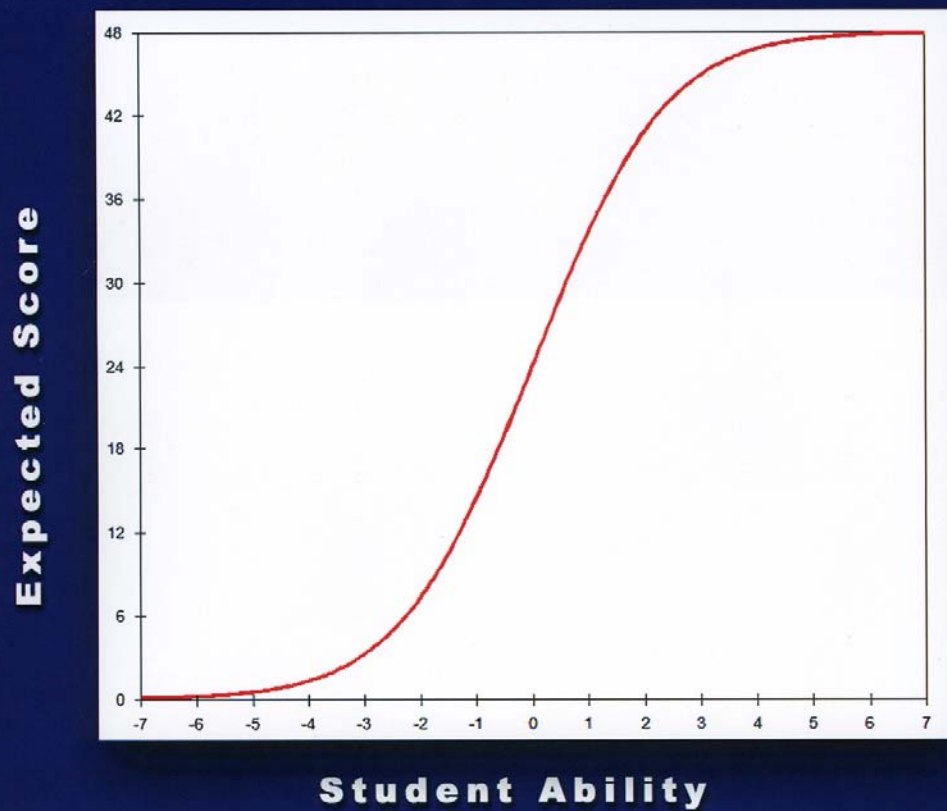
When I took office, there were wide-spread complaints from teachers and parents that students were being over tested. Two weeks each spring were devoted to testing: one week to the standards-based test and one week to the norm-referenced test. Especially with younger students, exhaustion of two weeks of testing was such that it was difficult to get them back on task for the rest of the spring. By reducing testing from two weeks to one week, we could greatly increase the amount of time students spent learning.

Some questions that have been nationally normed also meet Arizona standards. These questions can fulfill both purposes. A student answers the question once, but it counts in the score both of the reported AIMS score and the reported *TerraNova* score. This made us much more efficient in our testing.

We also found that by reducing the number of norm referenced questions by a predetermined amount, we were still, as noted, plus or minus one percent of the full norm-referenced test. This was determined by using tests that had been given, and looking at the score that would have resulted from a reduced number of questions, compared to the score that resulted from using all of the questions, as actually occurred with actual students. The next two pages are graphs showing how nearly identical the two results were.

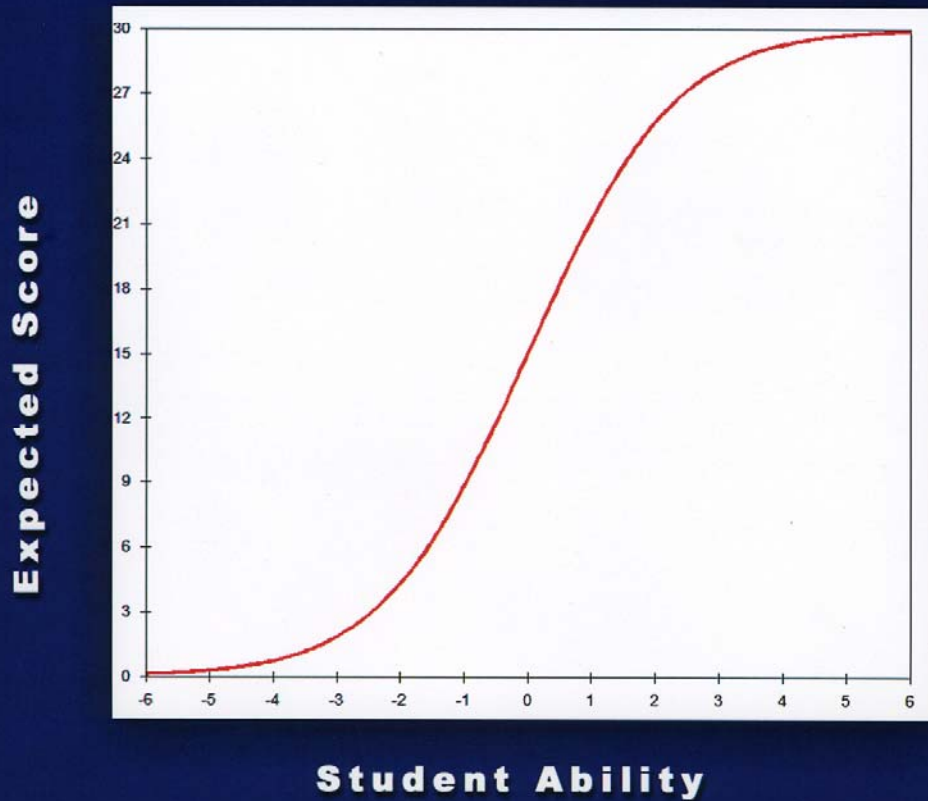
Test Characteristic Curve

Full Test (48 Items)



Test Characteristic Curve

Shortened Test (30 Items)



The Dual Purpose Assessment was then approved by the Arizona National Advisory Committee, which consisted of the following highly respected academics in this field (known as “psychometricians”):

Barbara Dodd, Ph.D., University of Texas, professor
Elizabeth Goertz, Ph.D., University of Pennsylvania, Professor
William Mehrens, Ph.D., Michigan State, retired, Professor Emeritus
Tom Haladyna, Ph.D., ASU West, retired, Professor Emeritus
Edward Wolfe, Ph.D., Virginia Tech, Associate Professor
Jerry D’Agostino, Ph.D., Ohio State, Associate Professor
Joe Ryan, Ph.D., ASU West, retired, Professor Emeritus

The result is that we have reduced testing time by one half, but still get a full AIMS report and a full *TerraNova* report, with the same results plus, or minus one percent on the *TerraNova*.

3. Contention No. 3:

Essentially, test scores have an upward bias when teachers become more familiar with the exam.

. . .

...items are publicly available for both teachers and students.

Id, p. 8.

Rebuttal:

Before we gave the *TerraNova* through the Dual Purpose Assessment, we used another one of the other three major norm-referenced tests: The Stanford 9. *TerraNova* came in with a bid that was about \$7 million less over the five-year life of the contract. While there are disadvantages to switching tests, there is a need for more public officials who are frugal, and we chose to go with the *TerraNova* and the lower bid.

One of the disadvantages of going with a new test was the fear that lower test scores would result. Arizona students had been performing above the national average on the Stanford 9 and there was some fear that this showing might be lost with a new test. The first time we gave the *TerraNova*, Arizona students continued to perform above the national average.

Because this was the first time the test was given in this state, there is no chance that teachers become more familiar with the exam, and yet the scores were still above the national average.

Furthermore, the items on the *TerraNova* test have never been “publicly available for either teachers or students.” The GI report refers to blueprints and sample

test items. Blueprints state the percentage of the test that will deal with a given concept, but do not give any test questions. The students cannot memorize answers to test questions; they must learn the concepts, which is exactly what we want. There are no “sample items” for *TerraNova*. There are sample items for AIMS, but these are different than the questions appearing on the actual test. Here again, this helps the student learn the concept, but not memorize the answers to questions. And this is only for AIMS – not for *TerraNova*.

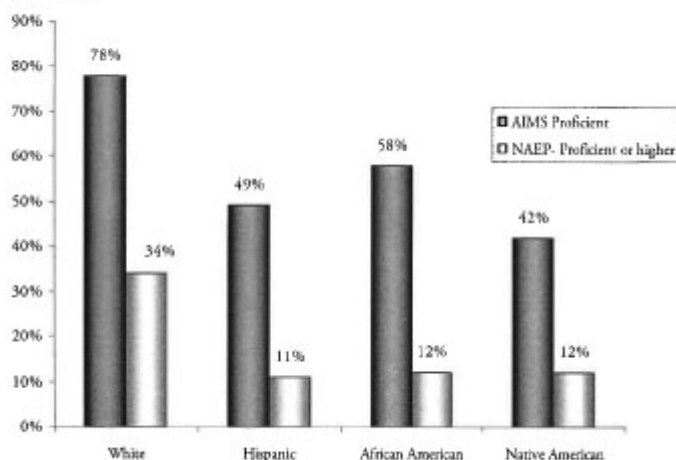
Moreover, the Stanford 9 was given in full, not as part of a Dual Purpose Assessment. Arizona students performed above the national average on the Stanford 9. This is conclusive refutation of the GI’s contention that Arizona students performed above the national average on the *TerraNova* because its combination into the Dual Purpose Assessment eliminated its validity.

There is even more refutation of GI’s position on this issue. Criticisms of the upward bias as teachers become more familiar with an exam had to do, in part, with the fact that the Stanford 9 was kept at the schools, and the same test was given year after year. But this is not true of the *TerraNova* test for grades 3 through 8, because of its combination with AIMS in the Dual Purpose Assessment. The high stakes nature of AIMS has resulted in very strict security rules for the handling of the AIMS test. Nothing was stored at the schools. All excess tests had to be returned. Since *TerraNova* is part of the same test that includes AIMS, the same high security rules now apply to *TerraNova*.

The irony is that criticism that might have had some validity with the Stanford 9 test has no validity with the *TerraNova* test, precisely because it has been combined with AIMS in the Dual Purpose Assessment.

4. Contention No. 4: That GI’s graph of NAEP vs. AIMS presents an honest picture: The GI *Policy Brief*, page 13, presents the following graph:

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



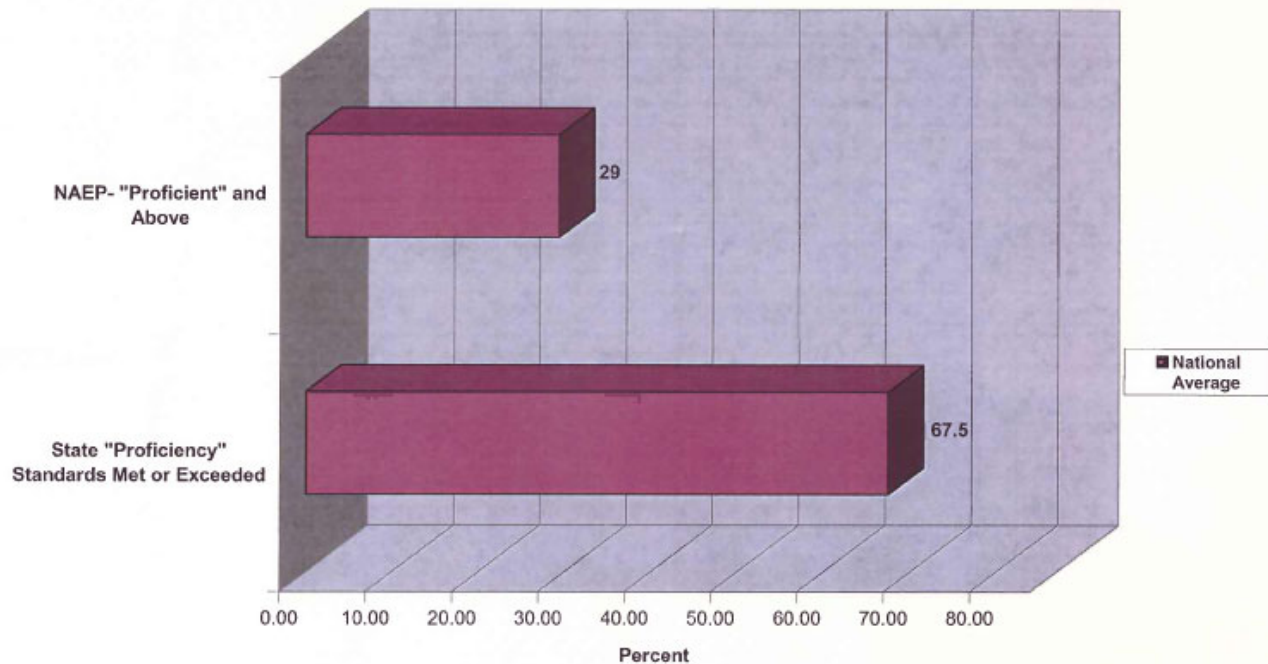
Source: Arizona Department of Education; U.S. Department of Education.

This graph appears to show that Arizona has a deceptive definition of “proficient” which misleads the public about the achievements of Arizona students.

Rebuttal:

The issue around the definition of “proficient” is a national not an Arizona issue. All 50 states use a method for defining “proficient” that follows the state of the art in the testing field, as prescribed by the federal government. The definition is set by a task force of teachers from around each state. NAEP used an entirely different process, in isolation that came up with a much more demanding definition of “proficient.” Every one of the 50 states has a proficiency rate on its state test that is higher than the NAEP proficiency rate. The following graph shows the discrepancy between NAEP “proficient” and the average of the 50 states’ “proficiency.”

National Comparison in Met or Exceeds
State Assessments and NAEP "Proficient" and Above Scores



Students reaching the NAEP "proficient" level demonstrated competency over challenging subject matter, including subject-matter knowledge, application of knowledge to real world situations, and analytical skills appropriate to the subject matter.

I want this rebuttal to be purely factual, rather than opinionated, so I will not comment. But the reader may draw whatever conclusion you feel is appropriate from the fact that the Goldwater Institute chose to create and disseminate Figure 3, without the necessary context shown by this national graph.

5. Contention No. 5: Arguments illogical on their face.

The Goldwater Institute found an assistant professor who is part of a marginal group that attacks a technical method known as the "three parameters model" (p. 16) as a way of attacking the *TerraNova*. We need not get into the technical issue, because the NAEP, touted by the Goldwater Institute, also uses the three parameters method.

The Goldwater Institute report states: "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." (P. 15.)]

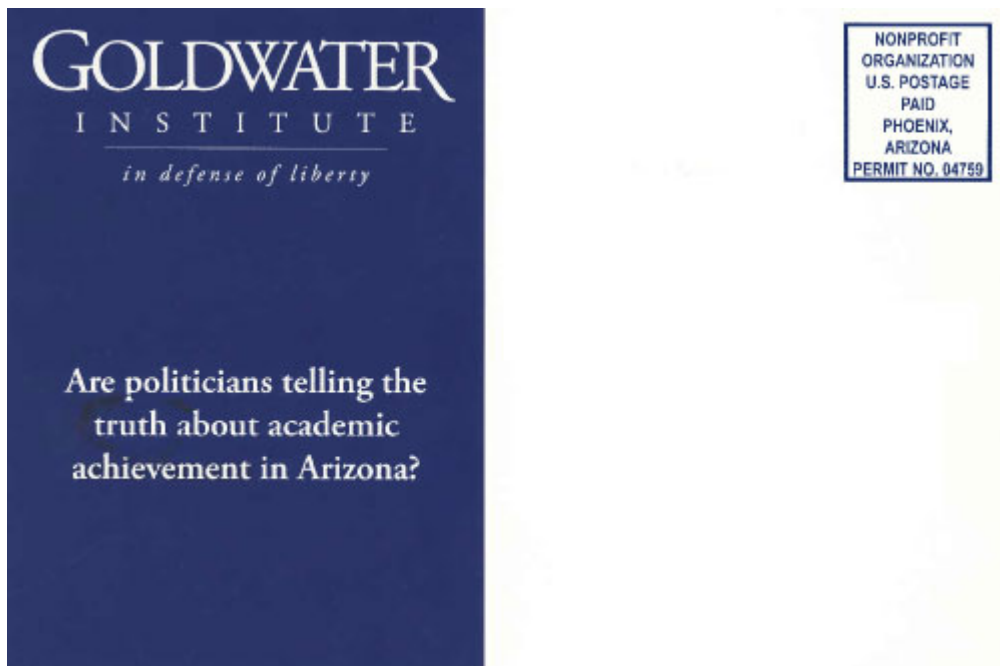
You know from page 1 above what criterion-referenced means. If a question is a measurement of a standard, it is a criterion-referenced question. Let's say that the standards have a performance objective that calls for the student to know about the military career of George Washington, and there is no standard dealing with Benjamin Franklin. If a question asks about the military career of George Washington, then it is

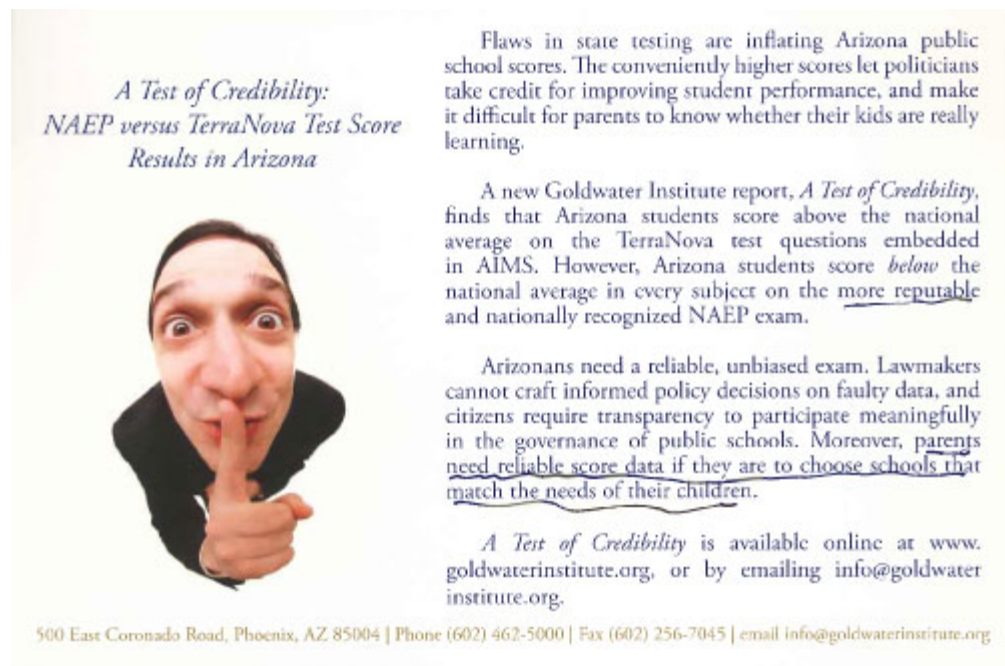
a criterion-referenced question; if it asks about Benjamin Franklin, it is not a criterion-referenced question. It does not matter what are the “scores on the examinations” or “the use of the resultant measures.” If the item measures a standard, it is criterion-referenced.

Like much of the Goldwater Institute report, the above-quoted GI statement that an item is not criterion-referenced without knowing, is nonsense.

6. Demagogic Postcard.

In addition to the above-referenced report, the Goldwater Institute chose to send out the demagogic postcard that is reprinted below:





The postcard states that “parents need reliable score data if they are to choose schools that match the needs of their children.” But the NAEP does not report results by school. (See p. 2 above.) The Goldwater Institute’s position is total nonsense.

7. Goldwater Institute Hides From Debate:

When this report came out in June, and was full of faults and misleading information, I challenged them to a debate on their own turf. Their initial response was “we’d be delighted to host an event and will get back to you after considering the few logistical questions.” There was a mini-debate on Channel 8’s Horizon, which was much too short to point out all the fallacies, but was telling. (It can be viewed at www.azpbs.org/horizon/index.asp.) They then apparently lost their taste for debate. Their last email stated “I’m not willing to use GI [Goldwater Institute] resources to have another one-on-one...”

What’s in a name? Barry Goldwater would never have attacked someone, and then refuse to debate him.

8. Help Correct the Record.

It is sometimes hard for the truth to catch up with false statements. The Goldwater Institute regularly sends out misleading emails. If you know anyone on its mailing list, please send me his or her email address so that I can get the facts to them. You can reach me at tom.horne@azed.gov.