OBSERVER

Mute the Messenger

When Dr. Walter Stroup showed that Texas' standardized testing regime is flawed, the testing company struck back.

by Jason Stanford Published on Wednesday, September 3, 2014, at 8:00 CST

Rebellions sometimes begin slowly, and Walter Stroup had to wait almost seven hours to start his. The setting was a legislative hearing at the Texas Capitol in the summer of 2012 at which the growing opposition to high-stakes standardized testing in Texas public schools was about to come to a head. Stroup, a University of Texas professor, was there to testify, but there was a long line of witnesses ahead of him. For hours he waited patiently, listening to everyone else struggle to explain why 15 years of standardized testing hadn't improved schools. Stroup believed he had the answer.

Using standardized testing as the yardstick to measure our children's educational growth wasn't new in Texas. But in the summer of 2012 people had discovered a brand-new reason to be pissed off about it. "Rigor" was the new watchword in education policy. Testing advocates believed that more rigorous curricula and tests would boost student achievement—the "rising tide lifts all boats" theory. But that's not how it worked out. In fact, more than a few sank. More than one-third of the statewide high school class of 2015 has already failed at least one of the newly implemented STAAR tests, disqualifying them from graduation without a successful re-test. As often happens, moms got mad. As happens less often, they got organized, and they got results.

Texas Education Commissioner Robert Scott, long an advocate of using tests to hold schools accountable, broke from orthodoxy when he called the STAAR test a "perversion of its original intent." Almost every school board in Texas passed

resolutions against over-testing, prompting Bill Hammond, a business lobbyist and leading testing advocate, to accuse school officials of “scaring” mothers. State legislators could barely step outside without hearing demands for testing relief. So in June 2012, the Texas House Public Education Committee did what elected officials do when they don’t know what to say. They held a hearing. To his credit, Committee Chair Rob Eissler began the hearing by posing a question that someone should have asked a generation ago: What exactly are we getting from these tests? And for six hours and 45 minutes, his committee couldn’t get a straight answer. Witness after witness attacked the latest standardized-testing regime that the Legislature had imposed. Everyone knew the system was broken, but no one knew exactly why.

Except for one person. Stroup, a bookishly handsome associate professor in the University of Texas College of Education, sat patiently until it was his turn to testify. Then Stroup sat down at the witness table and offered the scientific basis behind the widely held suspicion that what the tests measured was not what students have learned but how well students take tests. Every other witness got three minutes; it is a rough measure of the size of the rock that Stroup dropped into this pond that he was allowed to talk and answer lawmakers’ questions for 20 minutes.

A tenured professor at UT with a doctorate in education from Harvard University, Stroup isn’t frequently let out of the lab to address politicians in front of cameras. He talks with no evident concern that he might upset the powerful, and he speaks so quickly that his sentences have to hurry to keep up as he darts down tangents without warning. He taught in classrooms for almost a decade—it must have been a nightmare for his students.

But his testimony to the committee broke through the usual assumption that equated standardized testing with high standards. He reframed the debate over accountability by questioning whether the tests were the right tool for the job. The question wasn’t whether to test or not to test, but whether the tests measured what we thought they did.

Everyone knew the system was broken, but no one knew exactly why. Except for one person.

Stroup argued that the tests were working exactly as designed, but that the politicians who mandated that schools use them didn’t understand this. In effect, Stroup had caught the government using a bathroom scale to measure a student’s height. The scale wasn’t broken or badly made. The scale was working exactly as designed. It was just the wrong tool for the job. The tests, Stroup said, simply couldn’t measure how much students learned in school.

Stroup testified that for $468 million the Legislature had bought a pile of stress and wasted time from Pearson Education, the biggest player in the standardized-testing industry. Lest anyone miss that Stroup’s message threatened Pearson’s hegemony in the accountability industry, Rep. Jimmie Don Aycock (R-Killeen) brought Stroup’s testimony to a close with a joke that made it perfectly clear. “I’d like to have you and someone from Pearson have a little debate,” Aycock said. “Would you be willing to come back?”

“Sure,” Stroup said. “I’ll come back and mud wrestle.”

But that never happened. Stroup had picked a fight with a special interest in front of politicians. The winner wouldn’t be determined by reason and science but by politics and power. Pearson’s real counterattack took place largely out of public view, where the company attempted to discredit Stroup’s research. Instead of a public debate, Pearson used its money and influence to engage in the time-honored academic tradition of trashing its rival’s work and career behind his back.

Standardized testing has been a part of American life since the U.S. Army used bubble tests to separate the officer material from the infantry during World War I. And accountability has been a part of public education since Lyndon Johnson’s Elementary and Secondary Education Act offered federal money to states with strings attached.

Texas first linked standardized testing with accountability in education in the 1980s. H. Ross Perot suggested running public schools like private businesses when he chaired the Legislature’s Select Committee on Public Education, and since Perot made money in the punch-card business it made sense that standardized, fill-in-the-little-oval-with-a-No.-2-pencil bubble tests offered the best way to measure classroom learning.

Dr. Walter Stroup

That led four years later to the North Dallas business community putting a young Democratic lawyer named Barnett A. “Sandy” Kress in charge of the Committee on Educational Excellence. Kress’ committee devised an elaborate system that offered schools incentives and punishments based on their standardized test scores. Kress and other businessmen lobbied then-Gov. Ann Richards to expand the program statewide, but it only really took off when George W. Bush succeeded her. Bush took Kress with him to the White House in 2001 as a senior adviser to sell the “Texas Miracle” to Congress as the No Child Left Behind Act.

A dozen years later, standardized tests have become the pre-eminent yardstick of classroom learning in America, and Pearson is selling the most yardsticks. Besides selling tests to Texas, Pearson has the contract for the National Assessment of Educational Progress, also known as the “Nation’s Report Card.” Pearson writes the tests for the Program for International Student Assessment, the tests that always show the United States lagging behind Singapore and China. Pearson also handles the writing portion of the Scholastic Aptitude Test (SAT).

That’s all good news for Kress, now a lobbyist for Pearson. But to paraphrase George W. Bush, no one ever checked whether these tests answered the question, “Is our children learning?”

That was precisely the question Stroup started asking after he thought he found a way to use cloud computing to expose poor, minority children to basic math concepts using calculus. Stroup’s work with a program called the Algebra Project—the reason UT recruited him in the first place—earned him a National Science Foundation grant a decade ago to design a cloud-computing simulation to teach children math. Texas Instruments heard about the program and asked Stroup to use it’s TI Navigator calculator to work with younger students who had failed the state math test. In 2006, he implemented the curriculum at a Dallas-area middle school with impressive results. The same kids branded as failures by the state tests embraced the project, using the cloud technology collaboratively to learn basic math concepts. This was the breakthrough that everybody—Kress, Perot and lawmakers in Austin—had been looking for.

Stroup needed only to measure the improvements to show how successful his methods had been. And for that he had the state math test. As Stroup later explained in a plenary address to the 2009 convention of the North American Chapter of the International Group for the Psychology of Mathematics Education, the teachers and a former official with the Texas Education Agency (TEA) “were very sure, based on some practice tests and also based on what they observed in class, that the students were ready to ‘rock’ the tests.” However, the students’ scores rose only 10 percent, a statistically valid variance but hardly the change that he had observed in the classroom.

Using UT’s computing power, Stroup investigated. He entered the state test scores for every child in Texas, and out came the same minor variances he had gotten in Dallas. What he noticed was that most students’ test scores remained the same no matter what grade the students were in, or what subject was being tested. According to Stroup’s initial calculations, that constancy accounted for about 72 percent of everyone’s test score. Regardless of a teacher’s experience or training, class size, or any other classroom-based factor Stroup could identify, student test scores changed within a relatively narrow window of about 10 to 15 percent.

Stroup knew from his experience teaching impoverished students in inner-city Boston, Mexico City and North Texas that students could improve their mastery of a subject by more than 15 percent in a school year, but the tests couldn’t measure that change. Stroup came to believe that the biggest portion of the test scores that hardly changed—that 72 percent—simply measured test-taking ability. For almost $100 million a year, Texas taxpayers were sold these tests as a gauge of whether schools are doing a good job. Lawmakers were using the wrong tool.

The tests, Stroup said, simply couldn’t measure how much students had learned in school.

The paradox of Texas’ grand experiment with standardized testing is that the tests are working exactly as designed from a psychometric (the term for the science of testing) perspective, but their results don’t show what policymakers think they show. Stroup concluded that the tests were 72 percent “insensitive to instruction,” a graduate- school way of saying that the tests don’t measure what students learn in the classroom.

This claim earned Stroup a rebuke from the TEA, which stated that his findings betrayed “fundamental misunderstandings” about the way tests were constructed. The idea that most of a student’s test score carries over almost automatically, with little variance, year to year, was new, but it shouldn’t have been. After three years, STAAR scores have not budged much at all, and the TEA’s own recent report on the STAAR test results largely agrees with Stroup’s finding: The state agency declared that about 58 percent of middle school test scores showed little change from year to year.

Pearson wasn’t going to let Stroup’s findings go unchallenged. The company’s pushback against Stroup glossed over his most compelling findings and focused instead on what the company perceived as a mislabeled column in one of Stroup’s spreadsheets. In a public statement posted on the Pearson website, Dr. Walter “Denny” Way, senior vice president for measurement services at Pearson, said the 72 percent number was “not supported through valid research and will not stand up to a rigorous review by qualified experts.” After correcting what Pearson interpreted as the mislabeled column, Way wrote, the tests were “only 50 percent” insensitive to instruction. This alone was a startling admission. Even if you accepted Pearson’s argument that Stroup had erred, here was the company selling Texas millions of dollars’ worth of tests admitting that its product couldn’t measure half of what happens in a classroom.

Determining whether the number was 50 percent or 72 percent is one thing, but the real question is what that percentage meant. Stroup thought it quantified the portion of the test that measured test-taking ability. Another theory, from James Popham, emeritus professor in the Graduate School of Education and Information Studies at the University of California-Los Angeles, was that these types of tests measured innate intelligence, a morally dubious deduction when the results neatly correlate with race and ethnicity.

Way hypothesized that the 50 percent correlation “most likely reflects the fact that students are retaining what they’ve learned in previous years’ instruction and are building on that knowledge in the expected way.” But if that were true, then some students would do better in math than in reading, for example.

But that's not what the research showed. A student in the third grade did as well on a math test as that same student did in the eighth grade on a language arts test as the same student did in the 10th grade on a different test. Regardless of changes in school, subject and teacher, a student could count on a test result remaining 50 to 72 percent unchanged no matter what. Stroup hypothesized that the tests were so insensitive to instruction that a test could switch out a science question for a math question without having any effect on how that student would score.

Recently, the American Statistical Association condemned the use of student test scores to rate teacher performance. In a statement last April, the association cautioned that most studies find that "teachers account for about 1% to 14% of the variability in test scores," largely confirming Stroup's apparently controversial conclusion.

If it's true that the test measured primarily students' ability to take a test, then, Stroup reasoned to the House Public Education Committee in June 2012, "it is rational game theory strategy to target the 72 percent." That means more Pearson worksheets and fewer field trips, more multiple-choice literary analysis and fewer book reports, and weeks devoted to practice tests and less classroom time devoted to learning new things. In other words, logic explained exactly what was going on in Texas' public schools.

When business lobbyists and legislators desired tests that measure whether a student was "college and career ready," they didn't dramatically reform the curriculum. They needed harder questions based on the same curriculum, a trick Pearson managed by incorporating logic puzzles into questions about knowledge.

"My son came home from the third grade, and he said, 'You know daddy, someone is out there trying to trick me, and all I have to do is figure out how they're tricking me,'" Stroup told the legislators. "I'm not sure if it translates all that well to society if we teach kids gaming. All right, we end up with adults and professionals spending most of their time gaming the system."

Regardless of the substantial agreement that the TEA, Pearson and the American Statistical Association had with Stroup, the whisper campaign worked. Pearson kept pointing out the error on his spreadsheet, a bit of minutiae that caused even anti-testing academics to begin privately talking about "Stroup's mistake" and reporters to promptly lose interest.

Stroup disagreed with Pearson's analysis—vainly pointing out the agreement in his findings overshadowed the relatively insignificant error—but his 15 minutes of fame were up.

Rep. Eissler never called another hearing to have the debate between Stroup and a Pearson representative as Rep. Aycock had suggested. Eissler retired from the Legislature and now lobbies for Pearson.

Expecting part-time legislators to understand the implications of misusing standardized tests to hold public schools accountable is a tall order. In fact, Stroup had largely faded from public view by the time the Legislature came back into session in January 2013, but that's when his real problems started.

In retrospect, Stroup might have anticipated that the UT College of Education wouldn't celebrate his scholarship on standardized tests. In 2009, the Pearson Foundation, the test publisher's philanthropic arm, created a $1 million endowment at the College of Education, which in turn engendered the Pearson Center for Applied Psychometric Research, an endowed professorship, and an endowed faculty fellowship.

Tax law allows corporations to establish charitable foundations. What tax law doesn't allow is endowing a nonprofit to supplement the parent corporation's profit-driven mission. Last December, Pearson paid a $7.7 million fine in New York state to settle charges that the Pearson Foundation "had helped develop products for its corporate parent, including course materials and software," reported *The New York Times.* There is some evidence that the same thing is going on at UT, mainly because Pearson said so in a press release posted on the College of Education's website:

"Pearson Foundation's donation underscores the company's dedication to designing and delivering assessments that advance measurement best practice, help ensure greater educational equity and improve instruction and learning in today's global world," wrote Steve Dowling, Pearson executive vice president. "Through our endowment with The University of Texas at Austin, we are investing in technology-driven assessment research that will promote and personalize education for all."

Here's how that works in practice: Sharon Vaughn, Ph.D., is the H.E. Hartfelder/Southland Corp. Regents Chair and executive director of The Meadows Center for Preventing Educational Risk at the University of Texas College of Education. For the last four years, she has been simultaneously consulting for Pearson Learning, for an undisclosed sum, "to serve as an author to promote a K-6 reading program," according to disclosure forms she filed with the university. Having written 35 books, Vaughn is unquestionably qualified to "serve as an author." One of them, *Teaching Students Who Are Exceptional, Diverse, and At Risk in the General Education Classroom*, is an e-textbook published by Pearson and is now in its sixth printing.

Last August, she was the presenter on a webinar aimed at teachers to tout Pearson's iLit, "a comprehensive literacy solution designed to produce two or more years of reading growth in a single year." The website for Pearson iLit lists Vaughn as a member of the "authorship team," though it's unclear whether the product originated at UT or Vaughn helped create it off campus. And though she told UT that she would be serving "as an author," the webinar promo touted her academic job at UT. And lest anyone miss the point that this was the pedagogical equivalent of a timeshare pitch, John Guild, the senior

product and marketing manager for Pearson iLit, moderated the webinar. In short, it's the kind of arrangement that got Pearson in hot water in New York.

Most of what UT's Pearson Center does seems more promotional than productive. The Pearson Center invites pro-assessment scientists to give lectures at UT. The center's website has long lists of research presentations it has funded, indicating that the center is less a think-tank for groundbreaking research and more a mouthpiece in the marketplace of ideas.

In January 2013—six months after his testimony and less than a week after a story featuring Stroup aired on the Austin ABC affiliate—he received the results of his post-tenure review. It was bad news. The committee gave Stroup an unsatisfactory rating. Under state law, a public university in Texas can remove a tenured professor if he or she gets two successive unsatisfactory annual reviews. A Post-Tenure Review Report dated Jan. 10, 2013, dinged Stroup for "scholarly activity and productivity." In sum, the committee found he was publishing too little and presenting at conferences too seldom. Of his infrequent conference presentations, the committee members wrote, "Further, and equally concerning, is the paucity of presentations at research conferences. Dr. Stroup lists no presentations (competitively reviewed) at research conferences over the past six years (none since 2005)."

That was a curious conclusion, because Stroup had been presenting. When a professor undergoes post-tenure review, he or she completes annual reports using a form published by the Office of the Executive Vice President. Stroup's annual report lists four conference presentations, including two plenary addresses.

Regarding the overall charge of lack of productivity, the review committee failed to note Stroup's work with cloud computing, which led him to a group approach to education technology called "cloud-in-a-bottle" that is being implemented now at Lamar Middle School in Austin. In addition, his work with Texas Instruments on the math intervention program had led the company to create its Navigator system. All of this was in his annual report.

Stroup protested, citing these omissions, misstatements and errors. On Jan. 16, Committee Chair Dr. Randy Bomer submitted the post-tenure review report to the dean of the College of Education after changing Stroup's rating from "unsatisfactory" to "does not meet expectations." Bomer did not, however, correct any of the errors in the committee's report.

This does not mean Stroup's job is safe, however. The department has put Stroup on an aggressive publishing schedule, and forced him to move offices three times.

In a cover letter to the dean, Bomer explained the change from "unsatisfactory" to "does not meet expectations" as the result of a misunderstanding about the rating definitions. No mention was made of the first draft's supposed mistakes. In fact, Bomer wrote that the review committee did not even consider the omissions Stroup pointed out because "these things were not on his [curriculum] vita," even though they were on the forms provided by the university.

Pearson's real counterattack took place largely out of public view.

Bomer's cover letter indicates that the subject of Pearson came up with the review committee when Stroup requested a meeting after receiving the original unsatisfactory rating. "At that meeting, Dr. Stroup asked whether any member of the review committee or I had any relationship to Pearson publishing," Bomer wrote. "None of us has any such relationship."

Maybe Stroup's "emperor has no clothes" rebellion against UT's generous benefactor has nothing to do with his post-tenure review. For its part, Pearson Education said through a spokesperson that the company had no contact with UT about Stroup.

Maybe Stroup and his cloud computing and networked calculators don't fit neatly into an academic world, so his colleagues think he's slacking off. Maybe there's another explanation for why the UT College of Education is seemingly trying to get rid of a tenured professor.

But if Pearson were trying to strike back against a researcher who told legislators that they were paying $100 million a year for tests that mostly measure test-taking ability, it would look an awful lot like what is happening to Walter Stroup.

Correction: An earlier version of this story incorrectly reported that Dr. Randy Bomer's book was published by a Pearson subsidiary. The book was published by Heinemann. Pearson is the parent company of Heinemann's divisions and related companies in the United Kingdom, Canada, New Zealand and Australia, but not the U.S. division that published Bomer's book. The *Observer* deeply regrets the error.

Tags: Jimmie Don Aycock, No Child Left Behind, Pearson, STAAR tests, standardized testing, University of Texas, Walter Stroup