Private Data – The Real Story: A Huge Problem with Education Research

R. James Milgram

***Abstract*—A very influential paper on improving math outcomes was published in 2008. The authors refused to divulge their data claiming that agreements with the schools and Family Educational Rights and Privacy Act rules (FERPA) prevented it.**

- **It turns out that this is not true.**
- **The claimed legal foundations do not say what these authors said they do, this this is a widespread misconception among education researchers.**

When we found the identities of the schools by other means, serious problems with the conclusions of the article were quickly revealed.

- **The 2008 paper was far from unique in this respect.**
- **There are many papers that have had enormous influences on K-12 mathematics curricula, and could not be independently verified because the authors refused to reveal their data.**

In this article we describe how we were able to find the real data for the 2008 paper, and point out the legal constraints that should make it very difficult for authors of such papers to withhold their data in the future.

STANFORD Professor of Education, Jo Boaler, and her student, Megan Staples, published a very influential paper, [4], on improving math outcomes for high school students in 2008. The paper had so many policy implications that it was critically important for researchers to be able to check the results. But the authors refused to divulge their data, claiming agreements with the participating schools and FERPA rules prevented it.

- This seems to be a very common occurrence within education circles.
  - For example, the results of a number of papers with enormous effects on curriculum and teaching, such as [5] and [6] have never been independently veified.
  - Yet, [5] was the only independent research that demonstrated significant positve results for the Everyday Math program for a number of years. During this period district curriculum developers relied on [5] to justify choosing the program, and, today, EM is used by almost 20% of our students. Likewise [6] was the only research accepted by What Works Clearinghouse in their initial reports that showed positive effects for the elementary school program "Investigations in Number, Data, and Space," which today is used by almost 10% of our students. Neither one was ever independently verified.
- Between one quarter and 30% of our elementary school students is a huge data set. Consequently, if these programs were capable of significantly improving our K-12 student outcomes, we would surely have seen evidence by now.

So it is vitally important to analyze the legal foundations on which these authors base their refusal to share crucial data with other researchers.

It seems to be settled law that public schools, the lead researchers, and even individual teachers directly involved in conducting the research have no privacy protections and are subject to the Freedom of Information Act (FOIA) requirements.†

It turns out that FERPA only applies to student names and educational records, so it is not relevant here as we fully expected that student names and records would be redacted. However, the part of the Federal Code that does apply to human protections, [7], while providing strong protections for human research subjects, has critical exemptions for exactly the types of research done in [4], [5], and [6]. As a result, the claimed right to privacy does not hold for the data in these three papers, nor for papers like them, and their data is subject to FOIA.

Additionally, as regards [4], since one of the authors is a Stanford faculty member, she is subject to the requirements of the Stanford Openness in Research regulations, [12], that require Stanford faculty to make their data available to qualified researchers on request.

However, 7 years ago, the authors of [1] were unaware of the exemptions in [7] and the details of [12], so it was necessary for us to try to find the names of the schools studied in [4] by other means. We were lucky enough that a close examination of the data recorded in [4] allowed us to do just this.

Serious - perhaps even fatal - problems with the conclusions in [4] were quickly revealed once we had the school names.

I. Introduction.

Jo Boaler recently wrote a very pointed criticism of Prof. Wayne Bishop and me, [3]. It referred to a paper jointly authored by Wayne Bishop, Paul Clopton, and me, [1], that

R. James Milgram is a Professor of Mathematics Emeritus, Stanford University, USA. e-mail: milgram@math.stanford.edu. Special Thanks to Veronica Norris, J.D., RN for her legal analysis.

she seemed to be unaware had been accepted for publication in the peer reviewed education journal, Education Next, on 3/22/2006, and so she claimed it "has never been peer reviewed." What actually happened was that, for various reasons, it was held back and simply made available via the Internet for archival reasons.

Our paper, [1] studies a published article of J. Boaler and M. Staples, [4], a study focused on three California high schools, of which the most important for the study was called "Railside." If [4] is correct, the paper is extremely important as it indicates that most math instruction in U.S. high schools is ineffective, and indicates that the data in [4] appears to show another method is significantly more effective. *Since [4] is potentially so important and has so many implications about the best ways to teach our students, it needs to be independently verified.* **Indeed, a high ranking official from the U.S. Department of Education asked me to evaluate the claims in [4] in early 2005, because she was concerned that if those claims were correct U.S. ED should begin to reconsider much if not all of what they were doing in mathematics education.** This was the original reason we initiated our study, not some need to persecute Jo Boaler as she claims ([3], paragraph 3).

In any case, the conclusion of [1] was that no change in funding policies was required, as we had identified three critical areas where much more information was absolutely required before the results of [4] could be justified. The three areas will be described and discussed later in this note, and we have serious doubts about the possibility of fixing the issues that we've identified.

One of the reasons I held [1] back was that some of our math educators felt that when Boaler left Stanford, there was no real need for this paper to appear. This was not, of course, my focus, but it gave me concerns that if we published [1], it would be impossible for me to work with the community of math educators in this country, and there were still things that I felt a research mathematician could do to help improve the current mess in our K-12 mathematics outcomes. Indeed, at that time

- I was working on the Common Ground project,
- I was a member of the National Board for Education Science,
- the NASA Advisory Council (NAC),
- the NAC Human Resources committee,
- and was involved in a number of other projects directly related to math education.

Though the article was placed on my ftp site, very few people knew about it, so I was able to continue to work on the problems with our K-12 math outcomes.

II. The laws and regulations involved in [4], [1]

Boaler claimed [1] "contravenes federal law that protects human subjects" (see [3], Bullet 10), so it is worth noting first that [1] had been submitted to the Stanford IRB and was approved for publication after minor changes, again something that Boaler appears not to have known.

1) The law she quotes, Family Educational Rights and Prvacy Act, (FERPA), applies only to students names and educational records.
2) The much more relevant rule is *Subpart A – Basic HHS Policy for Protection of Human Research Subjects*, [7], with the relevant paragraph reproduced in the appendix to this note for convenience.
3) It turns out that [7] specifically exempts certain types of research from its protections. That are
   a) "research on regular and special education instructional strategies,"
   b) specifically "on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods."
   c) Additionally, there is an exclusion for "Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement)".[1]

Since these exclusions are the topics in [4], there is no expectation of privacy, except for the identities and school records of the individual students. [Subpart A – Basic HHS Policy for Protection of Human Research Subjects 45 C.F.R §46.101 (1991), Paragraph (b) www.hhs.gov]. *In particular, the article [1] contravenes no federal laws that protect human subjects.*

Also, Stanford's stated expectations for openness in research by Stanford faculty members are unambiguous. (See §2-6 of the Stanford Research Policy Handbook, [12].) The summary of the requirements in [12] starts as follows:

> [RPH 2-6] ... Expresses Stanford's commitment to openness in research; defines and prohibits secrecy, including limitations on publishability of results; specifies certain circumstances which are acceptable under this policy....

and the exact resolution is

> *That the principle of openness in research - the principle of freedom of access by all interested persons to the underlying data, to the processes, and to the final results of research - is one of overriding importance. Accordingly, it is the decision of the Senate that that principle be implemented to the fullest extent practicable, and that no program of research that requires secrecy (as hereafter defined) be conducted at Stanford University, subject to the exceptions set forth in Paragraph 4 of this Resolution.*

The exceptions are not applicable to either [4] or our paper, [1]. *Accordingly, our data – including the names of the three schools – is entirely available to researchers on request.*

It is possible that Boaler was trying to circumvent the requirements of §2-6 using the FERPA constraints. But as we

[1] The original purpose of these exemptions seems to have been to enable researchers to work on these topics with students as subjects without notifying and informing their parents. But an exemption is an exemption.

have seen, they do not apply to [4], and in any case, they are the wrong rules. She should have referenced [7], but that would not have helped either. Moreover, if her research grant contained restrictions on access to the data, then Stanford would not have allowed it, [13], [14]. It seems evident that [4] is subject to Stanford's openness of research requirements, and she was obliged to release the data, just redacting student names and identifiers.

III. Boaler's promised rebuttal to [1]

There was an article written about Boaler's critique, [3], followed by a large number of comments.[2] Perhaps motivated by this article, a number of people read [1] and noted that most of it was indisputable, (where we dissect the actual mathematics involved in [BS]), but it was necessary to have the real names of the schools in [4] to verify the remaining details in [1]. However, the actual names were not included in either [4] or [1].

In any case, Prof. Boaler promised a direct rebuttal to our paper, [1], prompted by these comments. She wrote

> *"I see in some of the comments people criticizing me for not addressing the detailed criticisms from Milgram/Bishop. I am more than happy to this. [...] I will write my detailed response today and post it to my site."*

(See footnote(2) and the comments at the end of the article it references.)

As I write this, nearly two months have passed since Boaler's rebuttal was promised, but it has not appeared. Nor is it likely to. This is because there is every reason to believe [1] is not only accurate but, in fact, understates the situation at Railside from 2000 - 2005.

Indeed, a high official in the district where Railside is located called and updated me on the situation there in May, 2010. One of that person's remarks is especially relevant. It was stated that as bad as [1] indicated the situation was at Railside, the school district's internal data actually showed it was even worse. Consequently, they had to step in and change the math curriculum at Railside to a more traditional approach.

Changing the curriculum seems to have had some effect. This year (2012) there was a very large (27 point) increase in Railside's API score and an even larger (28 point) increase for socioeconomically disadvantaged students, where the target had been 7 points in each case.

All this reminds me very much of a promise some years ago when I wrote an article on a widely heralded experiment at Andover High School in Michigan, [8]. I showed that the end result was a disastrous experience for a very large group of students in their first year college mathematics courses the year after graduating from the high school.

> *Then, too, a rebuttal was promised but never appeared, and as time went on, further material was published that made it clear that our analysis was both correct and justified,* [9].

[2] *insidehighered.com/news/2012/10/15/stanford-professor-goes-public-attacks-over-her-math-education-research*

IV. The details of the Boaler/Staples Article

In the remainder of this note we focus on the content of [4], and try to show what our difficulties with it were. We then explain what we did in [1] and why we needed to do it.

Our commentary in [1] refers to [4], published in the journal Teachers College Record in 2008, but the version we worked from was the preprint that appeared on the Boaler web-site for about 1 year as a PDF file dated 3/2/2005. The title in both cases was

> *Transforming Students Lives through an Equitable Mathematics Approach: The Case of Railside School,* and the authors were listed as Jo Boaler, then at Stanford University, and Megan Staples, then at Purdue University

The article [4] studies the mathematics outcomes for the ninth grade students who entered three California high schools in 2000.

The cohorts that [4] studied at the first two were selected from the ninth graders who took the standard Algebra I course in 2000-2001, while the cohort at the third was selected from the ninth graders (almost the entire class) who started with the ninth grade math course that year, as the program at the third school was non-standard.

The students were followed till they left high school, and detailed records were kept for them as they progressed through the mathematical programs at their respective schools for the three years from 2000-2001 through 2002-2003.

The schools are identified using the pseudonyms Greendale, Hilltop, and Railside in [4], and were described as follows on pages 5 and 6 of the preprint:

> "Both Greendale and Hilltop schools offered students (and parents) a choice between a traditional sequence of courses, taught using conventional methods of demonstration and practice, and an integrated sequence of courses in which students worked on a more open, applied curriculum called the Interactive Mathematics Program (Fendel, Fraser, Alper, & Resek, 2003), or 'IMP.' Students in IMP classes worked in groups and spent much more time discussing mathematics problems than those in the traditional classes. Railside school did not offer a choice and the approach they used was 'reform oriented. The teachers worked collaboratively and they had designed the curriculum themselves, drawing from different 'reform curriculum such as the College Preparatory Mathematics Curriculum (Sallee, Kysh, Kasimatis, & Hoey, 2000), or 'CPM' and 'IMP.' "

With a number of changes, *this pre-print is the paper that later appeared in print with the same title (See [4]).* Since there were changes between the two versions, we will send the original pre-print on request, if possible.

It is also worth noting that Prof. Wayne Bishop had requested the identities of the three schools from Prof. Boaler shortly after the 3/2/2005 preprint had appeared, but she refused, saying that it was against the law, the requirements of her NSF grant, and her agreements with the three schools.[3]

It is worth noting again that her refusal is contrary to federal FOIA requirements (see Appendix for the specific section of the federal code that makes studies like [4] subject to FOIA), and to Stanford's *openness of research* requirements.

The point of the Boaler-Staples paper seemed to be that the standard measures of student achievement – STAR exams, SAT, AP exams etc. – were not valid measures of what the students understood and could do. So the authors, together with the involved teachers at the three high schools, created 4 tests, a ninth grade pre-test and ninth, tenth and eleventh grade post-tests, with the ninth grade post-test given as a pre-test at the start of tenth grade. They were adminstered to the treatment groups as the students advanced from 9th through 11th grade.

The paper posits that the pre- and post-tests were a more valid indication of what the students actually knew and understood. Assuming this, the article goes on to say that the students at Railside started out at a much lower level than those at the other two schools, but as they advanced, this difference quickly evened out on the four tests, and by the end of the study the Railside students significantly outperformed the others.

The authors claimed, but did not conclusively demonstrate, that the three cohorts were roughly equivalent. They included a table, (Table 5 on page 12 of the preprint) [this table, with the numbers rounded appears as Table 6 in the published version] that showed students at Railside outperforming the students at the other two schools in Algebra in 2003, the final year of the study. Also, they asserted that on many, if not most, standard measures, the Railside students did not do well when compared to the students at the other two schools.

So, to validate the claims in the Boaler/Staples article, one has to do three things.

> 1. First one has to have accurate accounts of how the students in the treatment groups at the three schools did on the standard end of high school measures, including percent needing remediation in mathematics at the college level. (This last is clearly crucial, since the real measure of success in K - 12 mathematics instruction is success in college or the workplace.)
> 2. Second, one needs to verify that the treatment groups at the three schools were sufficiently similar that they could be meaningfully compared.
> 3. Third, one needs to evaluate the four tests to see if, in fact, they give a valid measure of the mathematics the students need to know and how well they know it.

Not one of these three items is addressed in any detail in the Boaler/Staples article. There are some general assertions that each of items 1 and 2 was done, but no details are included in [4]. Moreover, there are no indications of a detailed evaluation of the tests in [4] at all.

We believe that in a paper having the potential importance of this one – *implying the need for major changes in instruction and even curricula at the high school level* – the authors must give details for all three.

V. The status of the evidence for the three items above

In [1], we address each of these three items.

> 1. We show that there is no external evidence for improved student outcomes at Railside after the Boaler/Staples treatment.
>
> - The most telling data we find is that the mathematics remediation rate for the cohort of Railside students that Boaler was following who entered the California State University system in 2004 was 61%.
> - This was much higher than the state average of 37%,
> - Greendale's remediation rate that year was 35%
> - and Hilltop's was 29%.
>
> (And this is the case in spite of the information from [4] that over 40% of the Railside cohort had taken a pre-calculus or calculus course. (preprint - [4] top of page 2, claims the students had taken calculus), (published version, - [4], claims the students had taken pre-calculus or calculus, line 10, p. 612).) For more details about this see [2].
>
> - Also, the success rates on the mathematics/statistics AP exams at Greendale and Hilltop were at least at state averages, while there were *no students at Railside who took any mathematics/statistics AP exams during the period of the study*, (data obtained via FOIA requests to the three schools).
>
> These apparent contradictions to the Boaler/Staples claims need to be explained.
> 2. For the second, we show that there is strong evidence suggesting that the treatment groups at the three schools were significantly different. It appears, from state data, that the cohort at Railside was comprised of students in the top half of the class in mathematics. For Greendale, it appears that the students were grouped between the 35th and 70th percentiles, and that the students at Hilltop were grouped between the 40th and 80th percentiles.
> The way we determined this was to note the percentages of ninth grade students in the academic

[3] You may see the dramatic language she used in her refusal, which is quoted in a recent article by Prof. Bishop, [2]. The quote starts as follows: "I have documented the different lies and insults you have written about me and I have decided not to engage you in discussions of my study..."

year 2000-2001 who had taken more advanced mathematics STAR exams, Integrated 2, Algebra II, or Geometry at the end of the academic year at the three high schools. We focus on the ninth graders and their 2001 STAR exams because these students were in the entering classes that year, and this is the group that [4] follows. We can be reasonably sure that virtually all the students taking the more advanced tests in 2001 would have been in a more advanced mathematics class in 2000-2001, and not the ninth grade Algebra I classes or the first year Railside math course, *from which the respective school treatment groups were obtained, (preprint, [4], pages 7, 9), (published version, p. 615).* Details are given in §IX of this article.

This difference is enough to largely explain the Boaler/Staples results, since stronger students would naturally be expected to have higher scores. So the issues raised above need to be explained and the equivalence of the three cohorts carefully justified if this is even possible.

3. Finally, for the third item, we directly analyze the Boaler/Staples exams in [1]. Our analysis shows that they contain numerous mathematical errors, even more serious imprecisions, and also that the two most important post-tests were at least 3 years below their expected grade levels. We give some details in the last section, **The four Boaler/Staples tests and what they measure**. Full details are given in [1].

We do not claim the first two items were not addressed in [4], only that they had not been addressed nearly adequately. As regards the tests – undoubtedly the most important part of the Boaler/Staples study – we are very sure that the study fails. The four tests cannot measure what they must, unless mathematical imprecisions, errors, and low level mathematical content knowledge are what is required for success in college and the workforce. The effect of low level content knowledge is especially severe. Students who come to college in this situation must start with a remedial math course, and their chances of being able to major in any high tech area become extremely poor.

The details of our analysis of the three items above is the focus of [1].

VI. Details on how one could obtain the school identities from the [4] preprint

Now we explain how the Boaler/Staples preprint revealed the real identities of the three schools. Here is a key table that appears on page 12 of the preprint and with the numbers rounded in the published version as Table 6:

Table 5: California Standards Test, Algebra, 2003. Percent of students attaining given levels of proficiency.

	Greendale	Hilltop	Railside
n	125	224	188
Advanced	0	0	1
Proficient	6	13	15
Basic	27	28	33
Below basic	55	43	36
Far below basic	12	15	15

This table turns out to uniquely identify the schools. There are two things to note.

1) As California only has about 1300 high schools, the data in each of the columns in Table 5 should almost certainly identify a unique school.
2) The full STAR data set for each year is publicly available, and can be downloaded from the California Department of Education web site, [10].

We now give two examples from the 2003 STAR data-set. They show the form of the report that appears on the net for each individual school. Surprisingly, in the cases below the exact columns that appear in the Boaler/Staples Table 5 are seen as representing the performance of the 2003 NINTH GRADERS at the two respective schools[4], and the same is true of the STAR results for the third school.

	Grades										
	2	3	4	5	6	7	8	9	10	11	EOC
Reported Enrollment								489	336	341	
English Language Arts											
Students Tested								440	280	273	
% of Enrollment								90 %	83 %	80 %	
Mean Scaled Score								314.7	310.5	300.6	
% Advanced								5 %	7 %	5 %	
% Proficient								17 %	19 %	20 %	
% Basic								40 %	25 %	25 %	
% Below Basic								24 %	28 %	20 %	
% Far Below Basic								15 %	21 %	31 %	
General Mathematics (Grades 6 & 7 Standards)											
Students Tested								123			123
% of Enrollment								25 %			
Mean Scaled Score								289.6			289.6
% Advanced								0 %			0 %
% Proficient								11 %			11 %
% Basic								30 %			30 %
% Below Basic								35 %			35 %
% Far Below Basic								24 %			24 %
Algebra I											
Students Tested								188	44	19	251
% of Enrollment								38 %	13 %	6 %	
Mean Scaled Score								301.5	272.7	272.6	294.3
% Advanced								1 %	0 %	0 %	0 %
% Proficient								15 %	3 %	0 %	12 %
% Basic								33 %	20 %	22 %	30 %
% Below Basic								36 %	45 %	44 %	38 %
% Far Below Basic								15 %	33 %	33 %	20 %
Mathematics											
Students Tested								14			
% of Enrollment								5 %			
Mean Scaled Score								0.0			
% Advanced								0 %			
% Proficient								0 %			
% Basic								0 %			
% Below Basic								0 %			
% Far Below Basic								100 %			
General Mathematics (Grades 6 & 7 Standards)											
Students Tested								27			27
% of Enrollment								10 %			
Mean Scaled Score								290.7			290.7
% Advanced								0 %			0 %
% Proficient								7 %			7 %
% Basic								30 %			30 %
% Below Basic								48 %			48 %
% Far Below Basic								15 %			15 %
Algebra I											
Students Tested								125	44	11	180
% of Enrollment								46 %	16 %	4 %	
Mean Scaled Score								290.0	289.0	298.6	290.2
% Advanced								0 %	0 %	0 %	0 %
% Proficient								6 %	5 %	9 %	6 %
% Basic								27 %	30 %	36 %	28 %
% Below Basic								55 %	50 %	36 %	53 %
% Far Below Basic								12 %	16 %	18 %	13 %

[4]In an e-mail from Boaler dated 10/23/2006, she asserts that "There is no way of knowing the grade levels of the students [taking the STAR Algebra I exam] (that I know of) and your depiction of the data as that of '9th grade students' is incorrect."

In detail, here is the method we used. We took the data above from Table 5, and one of us (P. Clopton, Director of the Veterans Medical Research Foundation VetStats Core,) checked the entire publicly available 2003 California STAR data-base, looking for schools for which any column was identical to one of the columns in Table 5. In each case we found that there was one and only one school that had that data. But the students in the cohorts Boaler was studying should have been in 11th grade, not ninth, in 2003![4] So Table 5 is not data for the population studied in [4].

Looking more closely at the data for these schools from 2000-2004 we saw that what is remarkable about the supplied data in their Table 5 is that this 2003 ninth grade algebra data is THE ONLY TIME WHERE THE RAILSIDE STUDENTS CLEARLY OUTPERFORMED THE STUDENTS AT THE OTHER TWO SCHOOLS DURING THIS PERIOD. For example here is the data for ninth graders in 2004:

	Greendale	Hilltop	Railside
n	108	250	188
Advanced	1	0	0
Proficient	22	14	10
Basic	34	38	33
Below basic	39	42	48
Far below basic	5	6	8

So we can say that in reality, Boaler and Staples had already publicly identified the schools and then misidentified the data in their Table 5. Moreover, there is a possibility that they picked the unique data that might strengthen their assertions, rather than make use of the data relevant to their treatment groups.

VII. DOUBLE CHECKING THE SCHOOL IDENTITIES

We double checked the identifications in various ways, for example checking whether Boaler had ever worked with faculty members from any of the three schools. We found that this was the case for the schools we identified as Greendale and Railside.

Additionally, we were told by parents at the school we believed to be Greendale that Boaler had been studying the students there, and also at the school we believed was Hilltop.

Finally, we found an article of hers (entitled *Stanford University Mathematics Teaching and Learning Study: Initial report – A comparison of IMP 1 and Algebra 1 at Greendale School*) prominently posted on the web-site of the school we had identified as Greendale, [11].

> (A number of the parents at Greendale were not happy about that article. About 1/5 of the families with students at Greendale had made a huge effort to get a traditional mathematics track reinstated there, but students had to actively select it. The parent's perspective seemed to be that Boaler had interfered with this selection process. Not only had an article extolling the virtues of the IMP track been posted on the schools web-site, but I was told that Boaler had attended private meetings between the individual parents and the lead math teacher at Greendale.)

VIII. THE STUDENT OUTCOMES AT THE THREE SCHOOLS

Once we had the identities of the three schools we could fill in the missing data on student outcomes. This is fully reported in [1] so we don't repeat the details here. It suffices to say that every standard measure that had a strong correlation with the likelihood of student success after high school was much stronger for the Greendale and Hilltop students.

In particular, as indicated in §V, Railside's 61% mathematics remediation rate in the California State Universities was much higher than the 37% state average, while the rates for Greendale and Hilltop (35% and 29% respectively) were at or significantly below the state average.

Boaler and Staples were obligated to discuss this and explain why it does not contradict their results. But there is no indication of this in [4].

IX. COMPARISON OF THE THREE SCHOOL COHORTS

As to the possibility that the treatment groups at the three schools were equivalent, we only had the state data to work with, so our analysis had to be somewhat indirect and, as a result, a bit subtle. What we noticed was that

- With the change to eighth grade algebra in 1999, a larger number of middle school students began to take Algebra I before high school. But in 1999, this number was quite small. However, the best mathematics students, for years, had been, allowed to take Algebra I in eighth grade or even seventh grade if they did sufficiently well on a state algebra-readiness exam.
- It was still the case that such students would be expect to be the strongest mathematics students in their grades at their middle schools in 1998 and 1999 as a result.
- Students at Greendale and Hilltop could take any math course they qualified for ([4], published version, line -10, p. 614), so the strongest students would have taken a more advanced math course than Algebra 1 at these two schools.
- This was not the case at Railside since Railside required that every ninth grader take the Railside first year mathematics course ([4], published version, p. 614, line -8).
- Consequently, there was a straightforward way to identify the strongest mathematics students in the in the academic year 2000-2001 ninth grade classes at Greendale and Hilltop. *They were the ninth graders that did not take the Algebra I exam or the General Math exam at the end of academic year 2000-2001, but took the Integrated II, Algebra II, or Geometry exam instead.* (Recall that the cohorts that [4] studies were subsets of the ninth-graders at each school in the academic year 2000-2001.)

This difference in approach meant that in 2000-2001 we would only expect to see ninth graders taking the advanced tests

at the end of academic year 2000-2001, if they were among the strongest math students in their classes, and they were at Greendale or Hilltop.

In fact, this was exactly the case. Only 4 ninth grade students took one of the advanced tests at Railside at the end of academic year 2000-2001, while the number was 31% at Greendale, and 18% at Hilltop. *As a result, we can be reasonably sure that the top 30% of the ninth graders at Greendale and the top 20% of the them at Hilltop were not taking Algebra I in 2000-2001.* But, according to [4] and the information on the Railside web-site, all ninth graders at Railside were required to take the first year math course.

- **This is important because the treatment groups in [4] were taken from the students taking ninth grade Algebra I at Greendale and Hilltop in 2000-2001, while the group for Railside was selected from the students taking the first year math course that year.**

[4] tells us that about half of each cohort continued with the standard sequence at each of the high schools, and this half is the group that [4] evaluates to draw their conclusions.

So what do we have? We can assume that the surviving half of each treatment group was roughly made up of the strongest students in each original group. This leads to the following conclusions.

> i. the students in the percentiles from 35% to 70% were the group at Greendale that [4] focused on,
> ii. the students from the 40th to the 80th percentile were the group at Hillsdale,
> iii. the group at Railside consisted of the top half of the class.

These are hardly comparable populations if we make the standard assumption that abilities are uniformly distributed among populations and ethnic groups.

Boaler/Staples needs to come to grips with this issue. If the final treatment group at Railside contained a significant number of stronger students than the groups at the other two schools, that alone could be enough to explain why the Railside students did better on the Boaler/Staples exams than the groups at the other two schools.

X. The four Boaler/Staples tests and what they measure

Finally, we come to the third issue – how closely the tests used by [4] measure the mathematics students need to know. For this we had to do a close analysis of the four tests (actually we only looked closely at two of the three post-tests), and fully 2/3 of [1] is devoted to this.

There were many serious mathematical errors on the tests, and even more uses of language so imprecise that students who were never privy to the shorthand being used had little chance of finding the "correct" answers.

Here are just two of the problems we found with the tests. One of the errors was especially amusing, [1], p. 17. The "correct" response to the question the first time it appeared on one of the exams was wrong, so it was changed when the identical question appeared on a later test. However, the new answer while not technically wrong, was so peculiar as to be unbelievable.

Another question, this time on the final post-test, is

> 4. A triangle has an area of 62 sq units. If one side is 10 units, and one angle measures 40 degrees find possible measurements for the other sides and angles. Draw the triangle and label sides and angles.

The question implies there is only one answer up to congruence. This is not true. There would have only been one if the area had been larger than 68.687 square units, but for 62, there are two. One is very easy to analyze since it depends on a general argument. The other is quite tricky since it depends on a very special argument as the second triangle only exists for certain areas.

All of this is discussed in full detail in [1]. where we also show that the first and second post-tests were at least three years below California's expectations.

Taking all this into consideration, together with the very low level of the tests, we had no hesitation in asserting that the Boaler tests could not have been an accurate measure of the mathematics the students knew. Similarly, there was no evidence that they measured the mathematics students *needed to know* for the workplace or for success in college.

It remains to describe our professional qualifactions. Two of us are practicing Ph.D. mathematicians and the third, P. Clopton, is a well respected statistician. All three of us were involved in the creation of the 1998 California Mathematics Standards and Framework, and I have held a number of national positions overseeing research in mathematics education, as well as overseeing the creation of the recent Common Core Mathematics Standards.

XI. Conclusion

Taking all this into consideration, it should be clear that we have to apply the same high standards to research in education as we do for papers in medicine and other areas where we study human subjects. Failure to do so has extremely high costs.

XII. Appendix: The relevant part of 45 C.F.R §46.101 (1991): especially paragraph (b).

Authority: 45 U.S.C. 289(a).
Subpart ABasic HHS Policy for Protection of Human Research Subjects 45 C.F.R §46.101 (1991).
§**46.101** To what does this policy apply?
(a) *Except as provided in paragraph (b) [www.hhs.gov] of this section*, this policy applies to all research involving human subjects conducted, supported or otherwise subject to regulation by any federal department or agency which takes appropriate administrative action to make the policy applicable to such research. This includes research conducted by federal civilian employees or military personnel, except that

each department or agency head may adopt such procedural modifications as may be appropriate from an administrative standpoint. It also includes research conducted, supported, or otherwise subject to regulation by the federal government outside the United States.

> (1) Research that is conducted or supported by a federal department or agency, whether or not it is regulated as defined in 46.102 [www.hhs.gov] must comply with all sections of this policy.
> (2) Research that is neither conducted nor supported by a federal department or agency but is subject to regulation as defined in §46.102(e) [www/hhs.gov] must be reviewed and approved, in compliance with §46.101 [www.hhs.gov], §46.102 [www.hhs.gov], and §46.107 [www.hhs.gov] through §46.117 [www.hhs.gov] of this policy, by an institutional review board (IRB) that operates in accordance with the pertinent requirements of this policy.

(b) Unless otherwise required by department or agency heads, research activities in which the only involvement of human subjects will be in one or more of the following categories are exempt from this policy:

> **(1) Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as**
>
> > **(i) research on regular and special education instructional strategies, or**
> > **(ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.**
>
> **(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:**
>
> > **(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and**
> > **(ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.**
>
> **(3) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (b)(2) of this section, if:**
>
> > **(i) the human subjects are elected or appointed public officials or candidates for public office; or**
> > **(ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.**
>
> **(4) Research, involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.**

REFERENCES

[1] Wayne Bishop, Paul Clopton, R. J. Milgram, *A close examination of Jo Boaler's Railside Report*, ftp://math.stanford.edu/pub/papers/milgram/combined-evaluations-version3.pdf
[2] Wayne Bishop, "A response to some points of: when academic disagreement becomes harassment and persecution," (to appear).
[3] Jo Boaler, "When Academic Disagreement Becomes Harassment and Persecution," www.stanford.edu/˜joboaler
[4] Jo Boaler, Megan Staples, "Transforming Students Lives through an Equitable Mathematics Approach: The Case of Railside School," preprint 3/2/2005, and Teachers College Record, 110(3) (2008), 608-645.
[5] D. Briars, L. Resnick, *Standards, assessments – and wht else? The essential elements of Standards-based school improvement.* CRESST Technical Report 528, (2000)
[6] J. Riordan, P. Noyce, "The impact of two standards-based mathematics curricula on student achievement in Massachusetts," J. for Research in Mathematics Education 32 (2001), 368-398.
[7] Subpart A – Basic HHS Policy for Protection of Human Research Subjects 45 C.F.R. §46.101 (1991), paragraph (b). http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html
[8] ftp://math.stanford.edu/pub/papers/milgram/andover-report.html
[9] www.math.msu.edu/˜parker/monthly905-921.pdf
[10] star.cde.ca.gov
[11] www.gphillymath.org/StudentAchievement/Reports/Initial_report_Greendale.pdf
[12] rph.stanford.edu/2-6.html.
[13] dot.stanford.edu/C-Res/ITARlist.html.
[14] When we first heard Boaler's claim about restrictions in her NSF grant making it impossible for her to release data for [BS], Stanford's IRB checked with NSF and was assured that this was not the case.